AUDIOVISUAL SPEECH PERCEPTION IN TONAL LANGUAGE SPEAKERS: EVIDENCE FROM THE MCGURK PARADIGM

WENG YI

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University

Department of Chinese and Bilingual Studies

Audiovisual Speech Perception in Tonal Language Speakers: Evidence from the McGurk Paradigm

WENG Yi

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

November 2024

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ _____ (Signed)

_____WENG YI_____ (Name of student)

Abstract

Daily face-to-face communication, in its most natural form, involves information from multiple channels, especially audition and vision. In addition to audition, the primary modality for speech perception, the compensatory role of vision in speech perception among typical perceivers has become increasingly evident. However, studies on Indo-European language speakers have demonstrated that the relative weighting of auditory and visual information in speech perception is influenced by a variety of factors, leading to considerable diversity in audiovisual processing strategies across populations. Among these factors, language background, developmental stage, and neurotypicality have been highlighted as key moderators of this diversity in previous research. Furthermore, auditory noise and talking face processing patterns have been proposed to exert broader influences on audiovisual speech processing. Nonetheless, these factors have rarely been examined among speakers of tonal languages, despite tonal languages constituting the majority of the world's languages. To address this gap, the present study employs the classic McGurk paradigm to investigate audiovisual speech perception in tonal language speakers with both behavioural and eye-tracking data, placing a particular focus on various groups of native Cantonese speakers.

To clarify the role of tonal language background, the performances in the classic McGurk paradigm of Cantonese- and Mandarin-speaking young adults were compared under both quiet and noisy conditions. In terms of behavioural responses, Cantonese- and Mandarin-speaking participants exhibited comparable accuracy in identifying congruent stimuli and exhibited a high degree of audiovisual integration when perceiving incongruent stimuli in the quiet condition. However, native speakers of Cantonese, a language characterized by greater complexity in both segmental and suprasegmental aspects, made significantly more audiovisual-integrated responses along with fewer visual-dominant ones compared to their

Mandarin-speaking counterparts under both noisy conditions. As revealed by eye-tracking data, greater phonological complexity might also lead Cantonese speakers to rely more on fine-grained visual linguistic cues offered by the mouth area of speakers. Taken together, the impact of language background could be seen by the preference for audiovisual-integrated strategy in noisy conditions among Cantonese-speaking participants, which was potentially attributable to the inherent complexity of Cantonese phonology and the specific patterns of visual attention allocation.

Drawing upon the clarification of the role of language background, we further portrayed the developmental trajectory of audiovisual speech perception within a Cantonese-speaking context using the McGurk paradigm with various levels of auditory noise. A cross-sectional study covering Cantonese children aged 4 to 11 years was conducted. For behavioural responses, Cantonese-speaking children aged 4 to 9 years could not achieve comparable accuracy as adults in identifying congruent stimuli. Meanwhile, consistent with previous findings, the current study supported that tonal language background would not eliminate the developmental shift regarding sensory dominance in audiovisual speech perception as children aged 4 to 9 years made significantly more audio-dominant responses while fewer audiovisual-integrated responses to the incongruent stimuli compared to adults in the quiet condition. No significant differences were detected between the 10–11-year-old group and adults. Eye-tracking data revealed a synchronized developmental course in visual allocation to talking faces, as younger children aged 4 to 9 years fixated less on the mouth area of the speaker during the first half of the stimulus window than adults. In contrast, the 10–11-year-old group showed no significant differences from adults. The findings indicate that the tonal nature of languages does not exempt young speakers from undergoing the developmental shift in audiovisual speech perception, while the timing of the shift may be modulated by language background. Moreover, the link between speech perception and talking face processing is supported by the current

findings, which highlight the synchronous developmental courses shared by these two interacting processes.

In light of a more comprehensive understanding of the developmental course in typically-developing (TD) children, learning about the profile of audiovisual speech perception of Cantonese-speaking children with ASD compared with their TD counterparts was also pursued. Autistic individuals aged 8 to 11 years showed atypical patterns in processing audiovisual speech stimuli. When identifying congruent stimuli, autistic individuals achieved poorer accuracy compared with their TD counterparts matched on chronological age (CA-matched TD) or language ability (LA-matched TD). When perceiving incongruent stimuli, the autistic group made significantly more audio-dominant responses relative to their CA-matched TD counterparts. When compared with a group of TD children matched on language ability, they still exhibited differences in terms of within-group comparisons. The preference for the audio-dominant strategy in autistic individuals might result from their avoidance of social stimuli, as suggested by eye-tracking results, which might lead them to reduce visual intake at source in face-to-face contexts. However, differences in mouth-looking time between the autistic group and their TD counterparts were limited. This finding suggests that the core barrier for autistic children in audiovisual speech perception lies in their lack of general social interest rather than an inability to utilize informative visual linguistic cues from mouth movements.

Findings from the current thesis enhance our understanding of the mechanisms underlying audiovisual speech perception among tonal language speakers, with a particular emphasis on the understudied Cantonese-speaking population. Furthermore, the integration of behavioural and eye-tracking data provides new insights into the interconnected roles of speech decoding and talking face processing.

Acknowledgments

First and foremost, I would like to extend my most sincere gratitude to my supervisor, Prof. Peng Gang. Whether in academic pursuits or in personal life, Prof. Peng has always been a role model of mine. During one of the most uncertain times of my life, it was Prof. Peng who gave me the opportunity to pursue my doctoral studies at PolyU. With his encouragement, I embarked on my investigation into phonetics, which eventually led me to a deeper exploration of speech processing. During moments of self-doubt, he encouraged me not to dwell too much on the past but to look forward, reminding me that it is not the past that shapes the future but the choices we make in the present. This shift in mindset brought new fulfillment to my life. Over the three years of doctoral studies, I have learned far more than I ever imagined I could.

I am also deeply grateful to Prof. William Wang Shi-Yuan, who provided valuable suggestions for this series of studies, both in the classroom and during my confirmation process. A special thanks go to my host supervisor at the University of Minnesota, Twin Cities, Prof. Zhang Yang, whose invaluable insights have greatly contributed to both my current research and future career.

I would also like to express my heartfelt thanks to Dr. Rong Yicheng, my fellow companion in this challenging yet rewarding journey. I am also thankful to other members and alumni of Prof. Peng's research group for their camaraderie and support, and they are Dr. Tao Ran, Dr. Zhang Kaile, Prof. Chen Fei, Dr. Feng Yan, Ms. Ye Yanyuan, Ms. Qi Jing, Ms. Shu Yuqin, Ms. Li Jiaxin, Ms. Ji Jinxin, Ms. Hu Yiying, Ms. Yang Yifan and Mr. Zhang Gaode. I also owe a great deal of gratitude to my friends from different stages of my life: Ms. Li Xueying, Ms. Huang Qianxin, Ms. Huang Ziqing, Mr. Chen Guoji, Mr. Kong Weikun, Dr. Gan Zhanhui, Ms. Wang Jingyi, Dr. Wang Lu, Dr. Lu Yao and Mr. Chang Rui. My heartfelt thanks also go to my doctoral classmates: Ms. Yu Jiayu, Ms. Lin Jueyao, and Ms. Peng Yingying.

I would also like to thank the Speech Therapy Unit of our department, Zhaoqing No. 2 Kindergarten, and Xiamen Hand-in-Hand Kindergarten for their tremendous support in participant recruitment and experiment administration, as well as all the participants and their caregivers for their participation and support.

I reserve my deepest gratitude to my family, Mr. Weng, Ms. Liang and Mr. Chen, for their unfailing love, unconditional support, and unwavering trust. It was their understanding and encouragement have made all these miracles possible.

Table of Contents

List of Tables

List of Figures

Figure 5.7 (A) The time course of the face-looking ratio (empirical-logit transformed) in 50-ms time bins by three groups of participants, and (B) the estimated temporal courses of fixation (empirical logit-transformed) towards the face areas of the speaker for three groups of participants derived from GAMM.

List of Abbreviations

ADHD	attention deficit hyperactivity disorder
AIC	Akaike information criterion
AOI	area-on-interest
AQ	Autism Spectrum Quotient
AQ-Child	Autism Spectrum Quotient—Children's Version
ASD	autism spectrum disorder
BOLD	blood oxygenation level dependent
CA-matched	chronological-age-matched
CV	consonant-vowel
DSM-5	The Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
f_0	fundamental frequency
fMRI	functional magnetic resonance imaging
GAMMs	generalized additive mixed model
GLMM	generalized linear mixed model
HK SAR	Hong Kong Special Administrative Region
IQ	intelligence quotient
LA-matched	language-ability-matched
LMM	linear mixed model
LRT	likelihood ratio test
pSTS	posterior superior temporal sulcus
SD	standard deviation
SE	standard error
TD	typically-developing

V	vowel

Chapter 1. Introduction

1.1 Research Background

Humans gather information about the physical world through the interaction and cooperation of senses, which form a coherent perception of the surrounding environment and bolsters higher-order cognitive activities (Burr & Gori, 2012). In its most natural form, a simple face-to-face dialogue can be signalled by a variety of senses—voice, facial movements, gesture, tactile contact, or even olfaction—which are integrated by our brain, and create a holistic atmosphere about the conversation (Lalonde & Werner, 2021). While speech perception has been comprehensively studied from the auditory modality, it has received less attention from a multimodal perspective (Denes & Pinson, 1993). In fact, the pivotal role of visual speech has been witnessed with both behavioural and neural evidence (Ménard et al., 2014; Sato et al., 2013; van Wassenhove et al., 2005). Specifically, visual information consistent with auditory information facilitates speech perception, while conflicting visual cues might alter the outcome of auditory perception (Ma et al., 2009; McGurk & Macdonald, 1976). For instance, the classic McGurk effect (i.e., McGurk fusion) refers to a phenomenon where participants tend to perceive a /da/ when presented with a combination of auditory /ba/ and visual /ga/ (Macdonald & McGurk, 1978; McGurk & Macdonald, 1976). The illusory /da/, lacking both auditory and visual substance, emerges as a more precise estimate based on conflicting audiovisual information, given that perceivers may expect an initial lip openness corresponding to the visual /ga/ or a fronter place of articulation in accordance with the auditory /ba/ (Ernst & Banks, 2002; Weng et al., 2024). Consequently, the strength of the McGurk illusion has been frequently adopted to measure the magnitude of audiovisual speech integration (e.g., Hirst et al., 2018; Sekiyama, 1997; Stevenson et al., 2014a, 2014b; Zhang et al., 2018). With the McGurk paradigm, considerably differing perceptual strategies, as informed by the degree of the McGurk illusion, have been revealed among perceivers, indicating the intensity of

audiovisual speech integration can be susceptible to various factors. Following this line, the current study also adopts the classic McGurk paradigm to investigate audiovisual speech processing.

Previous studies using the McGurk paradigm centered on Indo-European-language speakers have suggested three factors that considerably influence audiovisual speech processing across populations, to which the current studies would pay particular attention. Firstly, previous studies have suggested that the strategy for audiovisual speech processing is affected by the phonological properties of specific language backgrounds, giving rise to differing behavioural responses to the same audiovisual stimuli by perceivers varying in native languages (e.g., Sekiyama, 1997; Sekiyama & Tohkura, 1993; Zhang et al., 2018). Secondly, the development of neurotypical children can profoundly influence audiovisual speech processing as the adult-like audiovisual processing manner takes a long journey to develop (e.g., Hirst et al., 2018; Tremblay et al., 2007; Weng et al., 2024; Weng & Peng, 2023). This results in a gap between adults and children and results in a developmental shift in sensory dominance when responding to the McGurk paradigm (Burr & Gori, 2012; Ernst, 2008). Thirdly, sensory disturbance encountered by children with neurodevelopmental disorders, such as autism spectrum disorders (ASD), induces difficulties in processing audiovisual speech stimuli and causes atypical behavioural manifestations (Feng et al., 2021a, 2021b, 2022; Stevenson et al., 2014a, 2014b, 2016). To date, research in this field has been primarily conducted among individuals from the West, with much fewer studies focusing on the East and even scarce for speakers of tonal languages that boast a vast number of speakers. Taking Cantonese as an example, this typical tone language has over 85.5 million speakers (Eberhard et al., 2022). However, it is rarely known how language background, developmental stage and neurotypicality relate to the audiovisual speech processing in tonal-language-speaking populations.

Besides, two other critical factors closely associated with audiovisual speech perception, namely, auditory noise and face-processing manner, may contribute to a more nuanced understanding of this topic. According to recent theories of multisensory processing, our brain generates perceptual outcomes in a statistically optimal manner resembling a maximum-likelihood integrator (Ernst & Banks, 2002). When calculating the optimal percept, our brain will allocate less weight to the specific sensory modality with greater variance due to noise, giving rise to shifts in behavioural responses (Ernst, 2006). Since our surrounding environment is noisy in nature, the magnitude of the shifting effect on the perceptual strategy in processing audiovisual speech stimuli across populations deserves additional attention. As for the face-processing manner, it can be explored through tracking simultaneous eye movements recorded with eye-tracking techniques when processing audiovisual speech stimuli (e.g., (Feng, Lu, Wang, et al., 2021; Feng et al., 2022; J. Irwin et al., 2017; Yamamoto et al., 2019). The McGurk design distinguishes from other non-social paradigms by tapping on two critical elements of social communication: speech decoding and talking face processing. These two elements, previously studied as separate domains, have been proposed to be linked with each other by recent studies (Pascalis et al., 2014). Therefore, another issue being explored is whether attention allocation on talking faces can provide valuable insights into unravelling the behavioural differences across populations.

Through the behavioural responses and eye movements recorded in audiovisual speech perception tasks under various auditory conditions from tonal language speakers varying in language background, developmental stage and neurotypicality, a more important question awaits clarification: whether stable, universal features exist across different groups of tonal language speakers and whether specific characteristics vary according to population differences? Answers to this question may be expected to may help reconcile the inconsistencies found in previous studies on audiovisual speech perception (Basu Mallick et al., 2015).

1.2 Research Purposes

In recent decades, the role of visual cues in speech perception gradually has become clear, confirming that our experience of speech communication is often constructed by information from multiple sensory channels (Sumby & Pollack, 1954). However, speech perception has been predominantly studied in the auditory modality, leading to a less developed understanding of multisensory speech processing. Thus, with the McGurk paradigm, the current study focused on speech perception under the audiovisual bimodal context, with the main purpose of observing the responses from participants and exploring the corresponding mechanism underlying behavioural manifestation. To achieve this primary objective, stimuli were presented in both channels, and participants would be required to report their perceptual outcome in an audiovisually holistic way. In response to the bimodal design, online eye movement data, in addition to behavioural responses, were also recorded using eye-tracking devices in order to capture a more comprehensive picture of participants' reactions to bimodal stimulation.

Moreover, existing studies on audiovisual speech perception using the McGurk design have been primarily conducted on Indo-European-language-speaking individuals, leaving the situation of perceivers from the tonal language background understudied. While the tonal aspect of languages has been suggested to implicitly influence the strategy of processing audiovisual stimuli in an implicit approach (Sekiyama, 1997; Zhang et al., 2018), relatively few studies have focused on tonal-language-speaking populations. In fact, lexical tones are estimated to be employed in 60–70% of the world's languages to contrast meaning (Yip, 2002), with over half of the world population speaking in tonal languages (Fromkin, 1978). Accordingly, the current study may place specific emphasis on tonal language speakers, aiming at documenting the behavioural and eye movement characteristics to clarify the moderating role of tonal properties and to supplement findings from non-tonal language speakers. Specifically, the current study is concerned with five main issues in audiovisual speech

perception with behavioural and eye-tracking evidence from tonal-language-speaking individuals.

The first aim of this thesis is to clarify the role of language background in the preferred strategy adopted in processing the McGurk stimuli. Given that the moderating effect of language background can be illustrated in several aspects, here, the complexity of a phonological system is particularly emphasized, as it has been suggested to profoundly impact the experience of the McGurk illusion by cross-linguistic studies (Sekiyama, 1997; Sekiyama & Burnham, 2008; Zhang et al., 2018). On the one hand, a more complex segmental phonology has been proposed to intensify visual reliance because producing segments offers rich identifiable visual cues to assist speech processing (Zhang et al., 2018). A richer suprasegmental phonology, on the other hand, has been argued to cause a low level of audiovisual speech integration due to greater attention consumption by auditory modality (Sekiyama, 1997; Sekiyama & Burnham, 2008). Taking lexical tone as an example, its changes are mainly realized by the vibration of vocal folds (Hayes, 2009; Kong, 2007; Ladefoged & Johnson, 2015), which corresponds well to the response pattern of the auditory system while involving limited distinguishable visual cues (National Research Council (US) Committee on Disability Determination for Individuals with Hearing Impairments, 2004; Sekiyama, 1997). In previous studies, the tonal property was attributed to the cause of a lower degree of audiovisual integration as indexed by the McGurk illusion, though this argument has been challenged by recent studies (Hazan et al., 2010; Magnotti et al., 2015). Study 1 of the current thesis aimed to access the issue of phonology complexity, both segmental and suprasegmental, through the comparison between speakers of two typical tonal languages, namely, Cantonese and Mandarin. The obtained data allow us to figure out whether tonal nature necessarily leads to a weaker audiovisual integration. Additionally, the commonality and differences shared by the two languages give rise to higher cross-linguistic comparability, and therefore, contribute

to a clearer clarification of the influence of phonology complexity of language background in audiovisual speech perception.

Building on the clarification of the role of language background, the second objective of this thesis is to portray the developmental trajectory of audiovisual speech perception within a Cantonese-speaking context. The adult-like audiovisual-integrated perceptual strategy has been likened to a "late bloomer" in the lifespan, given that such an optimal perceptual strategy does not appear to be inherent (Ernst, 2008; Hirst et al., 2018; Tremblay et al., 2007; Weng et al., 2024). Instead, studies on children from Indo-European-speaking backgrounds have revealed a common bias for unimodal auditory information during earlier stages of lifespan (Dupont et al., 2005; Hirst et al., 2018; Robinson & Sloutsky, 2010; Tremblay et al., 2007). The gap between favoured perceptual strategies between children and adults makes room for the emergence of a developmental shift in the strategy opted for audiovisual speech perception. Yet this developmental shift has been proposed to be rather language-specific, which is subject to the intensity of audiovisual integration attained by adults (Li et al., 2008; Sekiyama & Burnham, 2008). For instance, the comparatively lower level of audiovisual integration exhibited by Japanese-speaking adults may narrow the gap between children, resulting in a compressed space for the occurrence of the developmental shift (Sekiyama & Burnham, 2008). If tonal properties necessarily lead to hyper-reliance on auditory modality, it can be predicted that the developmental shift may be exempted from children from tonal backgrounds. However, the recent cross-sectional study by Weng et al. (2024) presented a clear and earlier developmental shift in Mandarin-speaking children. Thus, Study 2 in the current thesis sought to revisit the developmental trajectory of experiencing the McGurk illusion in another group of tonal-language-speaking children to illustrate whether the developmental shift is exemptible. If not, the time point of its occurrence will be of particular interest.

Based on understanding the developmental trajectory in typical children, learning about the profile of audiovisual speech perception of Cantonese-speaking children with ASD compared with their typically-developing (TD) counterparts remains the third purpose of this thesis. ASD is a neurodevelopmental disorder characterized by deficits in social interaction and communication, repetitive behaviours and restricted interest (American Psychiatric Association, 2013). By comparing the responses to speech (i.e., the McGurk stimuli) and non-speech (e.g., flash-beep stimuli) audiovisual stimuli, previous studies pointed out that the deficit in audiovisual processing tended to be speech-specific (Feldman et al., 2022; Stevenson et al., 2016). It is noteworthy that the social- and speech-related nature of the McGurk stimuli is very likely to touch upon the core barrier of ASD. Recent studies have highlighted the correlation between the degree of the McGurk illusion and the verbal/communicative ability of autistic individuals (Feldman et al., 2018, 2022), leading us to pay extra attention to the variable of language ability of autistic perceivers. In order to tackle the problem of language ability, a TD group matched on language ability (LA-matched TD group), in addition to another TD group matched on chronological age (CA-matched TD group), would be included in Study 3 to clarify if language ability may account for their atypical audiovisual speech processing in autistic individuals.

The fourth objective of the current thesis is to explore how the strategy adopted for processing audiovisual stimuli is influenced by auditory noise. Recent theories suggest that producing percepts from multisensory processing is highly contingent upon the reliability of the multiple sensory cues (Ernst & Banks, 2002). Auditory noise, as a degradation of audio information, has been predicted to impair the reliability of audition, and therefore, lead to shifts in perceptual strategies (Hirst et al., 2018; Weng et al., 2024). Across the three studies covered by the current thesis, the role of auditory noise was illustrated by the interaction between the three group-level variables. Particularly, the magnitude of the noise-induced shifting effect would be investigated by comparisons with the results under the quiet condition. Also, whether

the influence of language background, developmental stage and neurotypicality is altered with the presence and absence of auditory noise also remains a focus of the current thesis.

The eye movements simultaneously recorded during the audiovisual speech perception task seek to achieve the fifth goal of the current thesis, that is, to clarify the impact of face-processing pattern in audiovisual speech perception. The link between speech perception and face processing, which have been studied in separate domains, has become more and more clear (Hisanaga et al., 2016; J. Irwin et al., 2017; Pascalis et al., 2014; Yamamoto et al., 2019). More specifically, existing studies have suggested a correlation between the intensity of audiovisual speech integration indexed by the McGurk illusion and the frequency of looking at the mouth area of the speakers, which may potentially account for individual discrepancies in audiovisual speech perception (Gurler et al., 2015). Whereas, whether this association can be applied to illustrate the differing behavioural patterns across populations remained sparsely studied. For instance, Japanese speakers were found to rely more on overall features while their Caucasian-American-speaking counterparts tended to opt for a holistic approach, implying the linguistic and cultural background might foster differing face-processing manners (Miyamoto et al., 2011; Wang et al., 2015). Similarly, children who need to experience a developmental trajectory in sensory dominance have also been revealed to undergo a developmental process in face processing. A handful of studies tracking the development of face-viewing patterns during audiovisual speech perception tasks among children show an increasing trend in looking at the mouth region of the speakers (J. Irwin et al., 2017; Yamamoto et al., 2019). Considering the link between mouth-looking time and the strength of audiovisual integration, an attention shift towards the mouth might underly the developmental shift manifested in behaviour. In the case of autistic individuals, as a social-related stimulus, the processing of talking faces appears even more challenging (Klin et al., 2002). Since the first description of ASD, atypical eye gaze has been identified, and reduced eye contact remains a vital diagnostic indicator in today's clinical practice (Kanner, 1943; Lord et al., 2012). Such an eye-avoidance fashion seems to

cause avoidance of looking at the entire human face, which potentially prevents visual intake at its end during audiovisual speech perception and gives rise to unimodal auditory processing (Feng, Lu, Wang, et al., 2021). Taken together, the online eye movements would be recorded in every study and compared across populations, aiming to clarify to what extent visual allocation on talking faces during audiovisual speech perception tasks can account for the behavioural pattern within and between different groups of tonal language speakers.

1.3 Research Questions

With the current thesis, three original studies were conducted and reported.

Study 1 (Chapter 3: Exploring the Role of Language Background in Audiovisual Speech Perception)

This study focuses on the role of language background in audiovisual speech perception. Specifically, given that Cantonese is phonologically more complex than Mandarin, the current study investigates four main questions with the classic McGurk design: 1) whether the tonal property of tone languages necessarily leads to a low level of audiovisual integration in the McGurk paradigm, 2) whether there are differences between perceptual strategies adopted by Cantonese and Mandarin speakers under conditions with and without auditory noise, 3) whether these two groups of speakers adopt different perceptual strategies for audiovisual speech perception under noisy conditions, and 4) how their behavioural responses are linked to their face-processing patterns, particularly the time spent looking at the speaker's mouth area?

Study 2 (Chapter 4: The Development of Audiovisual Speech Perception in Cantonese-speaking Children: Effects of Language Background and Face-processing Pattern)

Study 2 seeks to explore the developmental course of audiovisual speech perception among children native to Cantonese. With the McGurk paradigm, the following research

questions are of particular interest: 1) whether the capability of identifying audiovisual congruent stimuli under clear and noisy conditions in Cantonese-speaking children improves with development, 2) for the perception of incongruent stimuli in the quiet condition, whether Cantonese-speaking undergo a developmental shift in sensory dominance, 3) whether introducing auditory noise to the audiovisual stimuli affects the development of audiovisual speech perception, and 4) whether the face-processing patterns induced during the speech perception task also exhibit a developmental effect?

Study 3 (Chapter 5: Deficient Attention Allocation Towards Human Faces Hampers Audiovisual Speech Perception in Children with Autism Spectrum Disorder)

Through the comparison between autistic children and their CA-matched and LA-matched TD counterparts, we aim to explore the following research questions: 1) whether autistic children behave differently in making responses to the McGurk paradigm from their chronological-age-matched TD groups, and 2) to what degree the abnormal behavioural responses can be addressed by the language ability of autistic children with the comparison between their language-ability-matched TD group, 3) whether the autistic group shows atypical visual attention allocation the speaker's face and the speaker's mouth area compared to the other two TD groups, and 4) to what extent the atypicalities manifested in audiovisual speech perception can be accounted for by the abnormal visual attention allocation in autistic children?

1.4 Structure of the Thesis

Chapter 2 is a literature review on audiovisual speech perception and its influencing factors. Firstly, an overview of audiovisual processing is provided, with the particularity of the McGurk paradigm being introduced. Next, current literature on the audiovisual speech processing among participants varying in language background, developmental stage and neurotypicality is reviewed. Also, existing findings on the influence of auditory noise and face-processing manner are summarized.

Chapter 3 investigates the role of language background in processing audiovisual speech stimuli with the comparison between Cantonese- and Mandarin-speaking adults. Behavioural response and eye-tracking data from forty college students, with half native to Cantonese and the other half native to Mandarin, are compared. Additionally, the predictability of the proportion-looking time towards the speaker's mouth to the accuracy of identifying audiovisual congruent stimuli and the strategy adopted for the perception of audiovisual incongruent stimuli is analyzed. Furthermore, the time courses of proportion-looking time dwelled on different facial features will be compared between the two groups of participants.

Chapter 4 portrays the developmental trajectory of experiencing the McGurk illusion as well as the face-processing pattern of Cantonese-speaking children. Specifically, Cantonese-speaking 4–5-year-olds (n = 20), 6–7-year-olds (n = 25), 8–9-year-olds (n = 20), 10–11-year-olds (n = 17) as well as adults (n = 21) are compared on their identification of congruent stimuli and perception of incongruent stimuli under quiet and noisy conditions. The direction of development will be explored by the predictability of age to the perceptual strategy for audiovisual speech processing. The proportion-looking time directed to different facial components, as well as its time course, will be compared across age groups to detect developmental differences regarding face-processing manner.

Chapter 5 compares both behavioural responses and face-processing patterns among autistic children (n = 20) as well as their language-ability-matched TD (n = 20) and chronological-age-matched TD (n = 20) peers under the McGurk design. The identification of audiovisual congruent stimuli and the perception of audiovisual incongruent stimuli are compared among groups. Meanwhile, the predictability of the visual attention allocated to the entire face area and the mouth region of the speaker to the behavioural responses of three groups of children is pursued.

Chapter 6 of the thesis provides a general discussion based on the three main studies based on findings from Chapters 3 to 5. How behavioural response to audiovisual speech stimuli is modulated by language background, developmental stage and neurotypicality is discussed respectively. Moreover, the observed alternation of strategies for audiovisual speech processing given rise by the introduction of auditory noise is also discussed. Additionally, the profound impact of talking face processing during audiovisual speech perception across studies will be summarized and highlighted.

Chapter 7 summarizes the findings and their significance, reviews the limitations and proposes future directions.

Chapter 2. Literature Review

2.1 Audiovisual Speech Perception and the McGurk Paradigm

Social communication, in its most essential form, is a multisensory process that incorporates multiple perceptual cues. In this process, listeners receive speech information not only through dynamic, transient auditory signals but also through the explicit movements and gestures of speakers in a face-to-face context. The role of visual cues in audiovisual speech perception has been witnessed under two circumstances. When the information from the visual modality is consistent with that from the auditory modality, visual cues considerably facilitate the efficiency of speech perception, especially in auditorily adverse circumstances (Sumby & Pollack, 1954; van Wassenhove et al., 2005). Visual cues that are incongruent with auditory ones, on the other hand, can alter the perceptual outcome to a certain extent, giving rise to perceptual illusions (Macdonald & McGurk, 1978; Shams et al., 2000). Both circumstances demonstrate the functioning of an audiovisual integrative mechanism underlying speech perception. The incorporation of vision, as well as other modalities, into speech perception, is believed to be motivated by the demand from our brain for a coherent and robust sensory estimate (Ernst, 2006).

2.1.1 The Perception of Audiovisual Congruent Stimuli

According to the early study by Sumby and Pollack (1954), the facilitation of vision in audiovisual speech perception may be more pronounced in adverse auditory conditions. One approach to address this issue is investigating speech processing among individuals whose hearing is poor or even impaired. The compensatory role of vision is best reflected by their expertise in lip-reading, a critical technique for individuals with hearing impairment to “read” the visible speech and interact with the hearing community under the assistance of surrounding context and language knowledge (Kyle et al., 2013). Enhanced visual perception can be observed among well-practiced lip-reading users in speech recognition tasks at phonemic,

syllabic and sentence levels (e.g., Auer & Bernstein, 2007; Bernstein et al., 2000; Mohammed et al., 2006) relative to normal hearing individuals, particularly in noisy environment (Ma et al., 2009). For instance, Bernstein et al. (2000) instructed 72 adults with profound hearing impairment and 96 controls with normal hearing to identify consonant-vowel (CV) nonsense syllables and words in isolation as well as in sentences. Results revealed that the group with impaired hearing exhibited better sensitivity to visual phonetic information. Consistently, Auer and Bernstein (2007) obtained a significantly higher accuracy from participants with early-onset hearing loss than their hearing counterparts in recognizing words embedded in visually presented sentences, highlighting the enhanced speechreading ability in the individuals with early-onset hearing loss.

Another common approach to assessing visual enhancement is to deprive the audition of hearing individuals by lowering the intelligibility of speech information with auditory noise. The classic study by Sumby and Pollak (1954) compared the performance of two groups of participants in identifying disyllabic words with either audiovisual (i.e., speaker's facial movements were visible) or auditory-only information (i.e., speaker's face was away) provided. Results indicated that visual information significantly boosted word recognition efficiency as participants from the audiovisual group showed enhanced accuracy relative to the auditory-only group. Notably, the comparisons across the speech-to-noise ratios (SNRs) of stimuli revealed that the contribution of vision was even more marked when auditory intelligibility was low, indicating that the extent of visual enhancement increases with more auditory deprivation. Visual facilitation in speech perception has been witnessed in numerous studies (e.g., Chen & Hazan, 2009; Gijbels et al., 2021; Hazan & Li, 2008; Lalonde & Holt, 2015; Sekiyama & Burnham, 2008; Sommers & Phelps, 2016), but visual gain does not appear to always increase monotonically with auditory noise, as recent studies have proposed that moderate noise, instead of heavy noise, induces maximized visual gain in speech recognition tasks, rendering an inverted U-shape relationship (e.g., Ma et al., 2009; Ross et al., 2007).

Visual gain, derived by subtracting the accuracy achieved in the auditory-only condition from that in the audiovisual condition, peaked at around -12dB SNR in both Ross et al. (2007) and Ma et al. (2009). Taken together, the contribution of vision becomes more pronounced with auditory intelligibility dropping from quiet to moderately noisy and reaching its maximum at around -12dB SNR. Even when SNR falls below -12dB, vision remains a contributing factor, albeit to a lesser extent.

2.1.2 Perception of Audiovisual Incongruent Stimuli

Incongruent stimuli refer to audiovisual combinations consisting of conflicting auditory and visual information. Along with audiovisual congruent stimuli, incongruent stimuli have also been frequently employed in audiovisual measurement. As opposed to congruent ones that can be encountered in real-world scenarios (e.g., speech components including phonemes, words, and sentences), usually incongruent stimuli are artificially synthesized for experimental purposes. The inconsistent modal information may evoke sensory illusions by mixing up perception outcomes from respective channels, whose magnitude allows us to observe the strength of modality blending and opens a window for measuring the magnitude of audiovisual integration. Audiovisual illusions can be triggered in both speech and non-speech settings.

2.1.2.1 Non-speech paradigm

A classic paradigm adopting non-speech stimuli was the sound-induced flash illusion (also called the illusory flash paradigm or flash-beep paradigm; Hirst et al., 2020; Shams et al., 2000, 2001, 2005). In Shams et al. (2000), participants who were presented with a single rapid flash accompanied by more than one beep reported that they had experienced more than one flash. The multiple perceived flashes were perceptual illusions as auditory interference successfully altered visual perception. Subsequent studies corroborated the robustness and prevalence of this illusion, making it a promising tool for measuring the degree of audiovisual integration (e.g., Parker & Robinson, 2018; Stevenson, et al., 2014a; Tremblay et al., 2007). Subsequent

studies further break the flash-beep illusions into fusion and fission, where fusion refers to the underestimation of the number of flashes while fission is the overestimation. These two subtypes of illusions have been claimed to be derived from distinct neural networks (Mishra et al., 2007, 2008) and manifested in behavioural differences (McGovern et al., 2014; Parker & Robinson, 2018; Vatakis & Spence, 2006).

2.1.2.2 The McGurk paradigm

The McGurk effect is a stable illusion involving speech stimuli (Basu Mallick et al., 2015; Hirst et al., 2018). In the original report of this effect, McGurk and MacDonald (1976) included 21 preschool children (3–4 years), 28 primary school children (7–8 years) and 54 adults (18–40 years) and recorded their responses to four audiovisual mismatched stimuli created by dubbing auditory "ba-ba" onto visual "ga-ga" ("AbVg"), "pa-pa" onto "ka-ka" ("ApVk") and vice versa ("AgVb" and "AkVp"). Interestingly, adult participants were strongly prone to report fused responses, namely, "da-da" and "ta-ta", to "AbVg" and "ApVk", respectively. As for "AgVb" and "AkVp" conditions, however, adult participants made a considerable proportion of combination responses, that is, composites comprising relatively less modified elements from each modality (e.g., "gabga", "bagba", "baga" and "gaba" for "AgVb"). Though combination responses signal audiovisual integration to some degree, several studies have reported that participants rarely make combination responses if consonant clusters were illegal in their native language (Li et al., 2008), making combined percepts less stable across linguistic backgrounds. In contrast, McGurk fusion is considered the most representative of the McGurk effect as it has been proven robust across language backgrounds (Hazan et al., 2010; Magnotti et al., 2015), regardless of the varying strengths. Since the fused responses lack both auditory and visual substance, they are purely perceptual illusions, opening a window for us to measure the magnitude of audiovisual integration.

The McGurk paradigm distinguishes itself from paradigms using non-speech stimuli in several aspects. Firstly, the McGurk paradigm utilized syllables as stimuli, which are physically more complex than flashes and beeps (Stevenson et al., 2014a). Secondly, the visual component of the McGurk stimuli is the human face, a special type of stimulus with communicative and social attributes (Stevenson et al., 2014a; Feldman et al., 2022). The speech and social characteristics of the McGurk paradigm give rise to divergent results from those using non-speech stimuli in populations with higher variability in communicative capabilities. For instance, children of different ages can be considered a group of perceivers who are sensitive to these two elements, as their developmental stage may pose constraints on their processing of the McGurk stimuli. For instance, Tremblay et al. (2007) compared the degree of audiovisual integration using both non-speech (illusory flash fission and fusion) and speech (the McGurk fusion) illusions among French-speaking 5–9-year-olds, 10–14-year-olds and 15–19-year-olds. Age-related differences were only found between 5–9-year-olds and either of the elder groups when processing the McGurk stimuli, which was absent in terms of non-speech illusions, indicating that the maturation processes for developing these two types of illusions were independent and asynchronous. On the other hand, individuals with ASD can be regarded as another group of populations who are sensitive to verbal and social skills, as verbal ability varies greatly within this population (Feldman et al., 2018, 2022). Stevenson et al. (2014b) compared the temporal window binding between autistic children and their TD counterparts using both non-speech and speech stimuli. Autistic children only showed an atypically wider temporal binding window in processing the McGurk stimuli but not in the flash-beep or tool-handle stimuli, concluding that the binding impairment tended to be speech-specific. The correlation between verbal and communicative skills and the strength of experiencing the McGurk illusion has been further substantiated by recent research (Feldman et al., 2022). Collectively, both non-speech and speech stimuli can successfully trigger robust audiovisual illusions among typical perceivers. However, there may be distinct mechanisms underlying

these two types of illusions, accounting for the asynchronous developmental rate among typical children varying in age and the unbalanced atypicalities among autistic individuals.

Wrapping up, the functioning of audiovisual integration can be captured in studies using audiovisual congruent or incongruent stimuli. When processing congruent stimuli, visual contribution will be magnified when the primary modality of speech communication, namely, audition, is impaired or deprived. When processing incongruent stimuli, the measurement of audiovisual integration can be as straightforward as the strength of audiovisual illusions. The mechanism underlying illusions evoked by non-speech and speech stimuli can be different. The social and speech-related nature remains an important rationale for adopting the McGurk paradigm to investigate audiovisual speech perception.

2.2 The Variability of Experiencing the McGurk Effect and Three Key Moderators

The robustness of the McGurk effect has been examined under various circumstances, even when the bimodal signals were simultaneously degraded (Hirst et al., 2018), temporally mismatched (Soto-Faraco & Alsius, 2009), or gender conflicting (Green et al., 1991). On top of that, previous studies have also concluded that there is variability regarding the strength of audiovisual integration exhibited in the McGurk paradigm, as some of the perceivers show a relatively weaker integration or even never experience the illusion while the rest of them showed a high magnitude of integration, despite the illusion exhibited high stability over time (Basu Mallick et al., 2015; Gurler et al., 2015; Keil et al., 2012). The variability in experiencing the McGurk illusion suggests that the strategies adopted for processing audiovisual speech stimuli appear to vary across individuals. Then, what drives such individual differences? One possible explanation is the existence of moderators and their interactions which contribute to making the perceptual pattern of audiovisual speech stimuli more complex. The three group-level factors specifically studied in the current thesis have exhibited significant moderating

effects within the population in previous studies using the McGurk paradigm, which potentially affects the strength of audiovisual integration.

2.2.1 Impact of Language Background

The moderating effect of language background on the strength of experiencing the McGurk illusion can be illustrated by several aspects, including exposure to a foreign-language-speaking environment (Sekiyama, 1997), bilingual experience (Marian et al., 2018), proficiency in a second language (Marian et al., 2018), and so forth. In the current thesis, how audiovisual speech perception is conditioned by phonological complexity is of particular interest. The impact of phonological complexity has been addressed by the comparison in the strength of visual influence between Japanese- and English-speaking individuals as significantly weaker audiovisual integration was found in Japanese-speaking participants (Sekiyama, 1994; Sekiyama & Tohkura, 1993). The phonological property of Japanese was proposed as a possible attribute by the authors (Sekiyama, 1994, 1997; Sekiyama & Burnham, 2008). From the suprasegmental aspect, pitch-accent utilized for contrasting lexical meanings in Japanese is mainly realized by the fundamental frequency of vocal folds vibration, whose production takes place at the phonatory stage and involves limited observable facial movements for visual identification (Hasegawa & Hata, 1992; Sekiyama, 1997; Tsujimura, 1999). Hence, the pitch-accent property of Japanese was argued to lead to over-reliance on auditory modality among Japanese speakers (Sekiyama, 1997). In terms of the segmental aspect, since around 70% of Japanese syllables were simple-structured syllables (e.g., isolated vowel (V) and consonant-vowel (CV), Kubozono, 1995; Otake, 2015), the simpler segmental phonology might meanwhile loosen the demand for compensatory visual assistance for audiovisual speech decoding (Sekiyama & Burnham, 2008). Taken together, the suppressed audiovisual integration, signalled by the weaker McGurk effect, in Japanese has been suggested to be jointly affected by its segmental and suprasegmental properties (Sekiyama, 1994, 1997). This account was attempted to be generalized to typical tone language speakers, as Sekiyama

(1997) also recorded a reduced McGurk effect from Mandarin-speaking participants residing in Japan. However, these early results were challenged by recent studies with a larger sample size where tone language speakers did not necessarily exhibit a weaker audiovisual integration as indexed by the McGurk effect compared with Indo-European speakers (e.g., Magnotti et al., 2015). Such mixed results might stem from the low comparability between languages that markedly differed in many aspects. Besides, there is no evidence indicating which is more phonologically complex between tone and stress languages (Hazan et al., 2010; Weng et al., 2024). From the suprasegmental aspect, Cantonese features a dense tone inventory with six unchecked tones as well as three checked tones (i.e., entering tones attached to syllables with final stops), which is more complex compared to the four-tone system in Mandarin (Bauer & Benedict, 1997; Department of Chinese Language and Literature of Peking University, 2004). In terms of segmental phonologies, again, Cantonese contains a greater number of vowels and syllable codas (i.e., -m, -p, -t, -k), resulting in a larger inventory of base syllables (Bauer & Benedict, 1997a; T. Lee et al., 2002). This theoretical foundation allows for more cautious discussions on how language background impacts audiovisual speech perception. With behavioural evidence from both typical tone languages, it is possible to investigate whether the tonal property of a language will enhance auditory reliance, lead to reduced visual influence, and eventually cause weakened audiovisual speech integration. By comparing the performance in the McGurk paradigm of Cantonese- and Mandarin-speaking participants, how phonological complexity tunes the sensory modality weighting during audiovisual speech processing is also sought to be clarified.

2.2.2 Developmental Effect in Audiovisual Speech Perception

As an optimal perceptual strategy, cross-modal integration has been suggested to develop late in humans as the maturity of the integrative mechanism entails incorporating senses from sensory organs varying in development rate (Burr & Gori, 2012; Ernst, 2008). Children at early stages of development may present differing perceptual outcomes from adults when

confronting identical multimodal stimuli. Being a form of multimodal processing, audiovisual speech perception has been revealed to take a long journey to mature in Indo-European-language-speaking children (Dupont et al., 2005; Hirst et al., 2020; Massaro, 1984; Massaro et al., 1986; Taitelbaum-Swead & Fostick, 2016; Tremblay et al., 2007). Specifically, this developmental process initiates with a preference for auditory unimodal information and gradually evolves into an adult-like manner where audiovisual bimodal information is both taken into account in percept generation. This developmental shift in sensory dominance has been corroborated by both behavioural and neural evidence (Heikkilä et al., 2018), with ten years of age as the critical milestone of maturation (Tremblay et al., 2007; Hirst et al., 2018). However, previous studies claimed that such a developmental shift was rather language-specific, given that children from the East appeared to be exempted from experiencing this process (e.g., Sekiyama & Burnham, 2008; Li et al., 2008). For instance, Japanese-speaking children exhibited little progress from 6 to 8 years in visual utilization as opposed to their English-speaking counterparts (Sekiyama & Burnham, 2008). Again, linguistic property and social custom were considered as consequences. Recalling the higher auditory reliance and reduced audiovisual integration in Japanese-speaking adults (Sekiyama & Tohkura, 1993; Sekiyama, 1994), it can be concluded that the prerequisite for the developmental shift lies in the gap in the magnitude of audiovisual integration between children and adults. Taking the case of Japanese speakers as an example, the weak audiovisual speech integration shared by children and adults yields limited developmental differences, constraining the potential for a developmental shift in sensory dominance. This conclusion was seemingly extended to Mandarin speakers, as no differences regarding the strength of audiovisual integration reflected by the McGurk effect were found between children and adults by Li et al. (2008). Agreeing with Sekiyama et al. (2003), Li et al. (2008) tried to attribute the absent developmental shift to the shared social custom as well as the tonal nature of Mandarin and Japanese. Yet this result contradicts recent findings that Mandarin-speaking participants are comparably strong, or even

stronger, in audiovisual integration with Indo-European-language speakers, which inevitably leads to a gap against the universal auditory bias among children (Chen & Hazan, 2009; Hazan et al., 2010; Magnotti et al., 2015). Such inconsistency has been tentatively intermediated by a recent study using the McGurk paradigm (Weng et al., 2024), where Mandarin-speaking children were found to experience this developmental shift at the age of around five, which is earlier than expected. This study highlights two implications for the development of audiovisual speech perception. First, the tonal property of a language will not eliminate the developmental shift among tone-language-speaking children. Second, the time point at which the shift occurs may be subject to language background. Guided by these insights, Study 2 of the current thesis intends to track the developmental trajectory of audiovisual speech perception in another group of tone-language-speaking children to examine whether the developmental shift is necessary. If this holds, the time point of its occurrence will be of particular interest.

2.2.3 Audiovisual Speech Perception in Individuals with ASD

ASD is a neurodevelopmental disorder characterized by deficits in social interaction and communication, repetitive behaviours and restricted interest (American Psychiatric Association, 2013). Current theory has suggested a continuum view of ASD diagnosis, which assumes that there is an autism spectrum ranging from normality to autism according to autistic traits (Baron-Cohen et al., 2001; Wing et al., 1988). In this sense, the autism spectrum is extended to non-autistic individuals who fall on the other end of the spectrum. Audiovisual speech perception has been reported to be conditioned by autistic traits among neurotypical population (Ujiie et al., 2015), as Autism Spectrum Quotient (AQ) positively predicted the frequency of making audio-dominant responses to the McGurk stimuli while negatively predicted audiovisual-integrated responses, indicating autistic symptoms themselves may moderate the pattern of audiovisual integration at a population level. When examining this issue by classifying participants based on the categorical diagnosis of ASD, despite the results were not always consistent, several review papers agree on atypical behavioural responses in

autistic individuals, who are prone to generate percepts depending on unimodal auditory information, resulting in weaker integration of bimodal cues (Feldman et al., 2018; Zhang et al., 2019). Such atypical audiovisual processing aligns with the symptom of sensory disturbance commonly associated with individuals with ASD (Stevenson et al., 2014a, 2014b. It is noteworthy that such an atypical integrative mechanism was mainly reflected in social stimuli processing (e.g., the McGurk stimuli), suggesting a close association with the social deficits in autistic individuals. Firstly, the strength of the McGurk illusion has been revealed to positively predict language/communication abilities while negatively correlating autism severity (Feldman et al., 2022). Secondly, the atypical temporal binding mechanism in audiovisual integration was only found in the McGurk paradigm instead of non-social stimuli, and therefore, this deficit has been proposed as speech-specific (Stevenson et al., 2014a, 2016). Two vital elements touched by the McGurk paradigm, namely, speech decoding and social stimuli processing, have led us to two factors that are potentially associated with the atypical audiovisual speech perception in autistic children: language ability and visual attention allocation on talking faces. While the correlation between language skills and audiovisual speech perception has been clearly presented by Feldman et al. (2022), how audiovisual speech perception, especially audiovisual integration, is impacted by visual attention allocation on talking faces is less understood. Specifically, whether the observed weakened audiovisual integration was constrained by the reduced general interest in social stimuli (e.g., attention allocated to the human face) and/or by the reduced linguistic knowledge in extracting finer visual cues (attention directed to the mouth area) requires further investigation.

Altogether, the three moderators identified in the literature may contribute to the variability of perceptual strategies used in audiovisual speech processing. Therefore, a dedicated investigation of these moderators is warranted to clarify their individual and collective roles. By isolating these factors and examining their impacts in a controlled manner,

we aim to better understand the mechanisms driving variability in the observed effect across different groups of tonal language speakers.

2.3 Auditory Noise and Audiovisual Speech Perception

As a common element in everyday environments, auditory noise introduces significant challenges to understanding spoken language in daily contexts by affecting listeners' ability to accurately decode and interpret speech signals, often leading to increased cognitive load (Zekveld et al., 2011). While we perceive the environment through the signals from various senses, noise inevitably disrupts the transmission of useful information. SNR is a straightforward indicator that measures the ratio of unwanted noise relative to intelligible speech. Given the critical role of auditory noise in shaping perception outcomes in audiovisual speech perception (Burr & Gori, 2012; Ernst & Banks, 2002), it offers an opportunity for us to observe how different groups of tonal language speakers react to audiovisual speech stimuli under differing auditory conditions.

2.3.1 Statistically Optimal Hypothesis in Multisensory Processing

That humans integrate multisensory information in a statistically optimal fashion has been elucidated in visual-haptic processing. Ernst and Banks (2002) investigated how visual and haptic information are integrated in height estimation using a two-interval forced-choice paradigm. Participants were required to compare the height of a bar in haptic-only, visual-only, or visual-haptic settings, with varying levels of visual noise introduced to the latter two conditions. Results showed that when the reliability of vision is lowered by noise, perceivers tended to weigh more on the haptic modality in generating perceptual judgments. Accordingly, a principle of predicting the perceptual outcome in multisensory processing was proposed: to minimize the variance in generating the final estimate, which corresponded well with a maximum likelihood indicator. This framework has been tested out in other multisensory practices, including visual-auditory localization and visual-haptic integration in sensing the

shape of objects (Helbig & Ernst, 2007). Following this line, in audiovisual speech processing, the introduction of auditory noise may lower the reliability of auditory, and perceivers will, therefore, weigh more on visual modality, since it is less likely to be affected by auditory noise. Findings from recent studies on audiovisual speech perception using the McGurk paradigm support this statistically optimal hypothesis (e.g., Hirst et al., 2018; Weng et al., 2024), showing that noise alone can alter the perceptual strategy opted for audiovisual speech processing.

2.3.2 Identifying Audiovisual Congruent Stimuli in Auditory Noise

When identifying audiovisual congruent stimuli, audition and vision are in a cooperative relationship, with a common aim to decode the speech information that is consistent across modalities (Weng et al., 2024). The introduction of auditory noise interferes with the intelligibility of auditory information, depriving a vital information source supporting successful identification. Meanwhile, vision that is less affected by auditory noise will take over the dominance in percept generation in accordance with the statistically optimal hypothesis (Ernst & Banks, 2002). Since visual speech is less phonetically detailed compared to auditory speech (Kuhl & Meltzoff, 1988), a drop in identification accuracy can be generally observed in previous studies on identifying audiovisual congruent stimuli (e.g., Chen & Hazan, 2009; Ma et al., 2009; Ross et al., 2007; Sekiyama & Burnham, 2008).

2.3.3 Perceiving Audiovisual Incongruent Stimuli in Auditory Noise

In the case of perceiving audiovisual incongruent stimuli, auditory and visual modalities are in a competitive relationship as they convey conflicting information (Robinson & Sloutsky, 2010). Taking the McGurk paradigm as an example, a strategy that incorporates both auditory and visual information in producing percepts can be considered the optimal strategy, as implied by the widespread reports of the McGurk fusion in typical perceivers. As auditory noise increases, the role of the visual modality is elevated in the competition between modalities accordingly. For instance, Stacey et al. (2020) measured the strength of the McGurk effect in various levels

of auditory noise, obtaining significantly more non-auditory responses in noisy conditions relative to the quiet one. The significant decrease in audio-dominant responses was attributed to the higher reliance placed on the vision when auditory information was disrupted by noise. However, it is noteworthy that the term “McGurk response” in this study was defined as any non-auditory response, which blurred the line between audiovisual-integrated (i.e., “Da”) and visual-dominant (i.e., “Ga”) responses. Splitting these two distinct responses allows for portraying the noise-shifting effect more clearly. Marian et al. (2018) compared the response patterns to the audiovisual incongruent stimuli in quiet and noisy conditions among early bilinguals, late bilinguals and monolinguals. Monolinguals, who showed a low level of audiovisual integration (14%) under the quiet condition, were observed a significant rise in making audiovisual-integrated responses in the noisy condition (43%), indicating auditory noise could lift audiovisual integration from a lower level by boosting visual utilization. In the cross-sectional studies by Weng et al. (2024), adult participants already exhibited a large proportion of audiovisual-integrated responses in the quiet condition. When auditory noise was introduced, the number of visual-dominant responses that showed an increasing trend alongside decreasing SNR eventually exceeded that of audio-dominant and audiovisual-integrated responses, demonstrating that the statistically optimal strategy was shifted from the audiovisual-integrated to the visual-dominant one with auditory noise loud enough.

Taken together, these studies suggest that the noise-induced shifting effect reflected from the perception of audiovisual incongruent stimuli follows the statistically optimal hypothesis (Ernst, 2006; Ernst & Banks, 2002). Along with increasing auditory noise, our brain tends to allocate less reliability to the auditory modality while weighing more on vision that is less likely influenced by auditory noise (Hirst et al., 2018; Weng et al., 2024). The resulting enhanced visual utilization under noisy conditions may boost audiovisual integration to a certain extent. When auditory intelligibility keeps falling and the reliability of audition continues to decline, the audiovisual-integrated strategy will no longer be considered the

statistically optimal manner in perception generation. Instead, with the significantly raised status, visual modality eventually takes over the dominance in producing perceptual outcomes, and the visual-dominant strategy is taken as the statistically optimal option.

2.4 Face-processing and Audiovisual Speech Perception

When it comes to the multisensory nature of speech communication, the primary source of visual intake should be the human faces. The human faces are special in many ways, and we start to show preferences for faces during infancy (Birulés et al., 2023). Interestingly, such a bias seems to be, at least partially, driven by communication, as the attention allocated to faces by infants will not be enhanced without being involved in communicative events (Farroni et al., 2002; Vecera & Johnson, 1995). Among all the facial components, the mouth area remains especially crucial for speech perception mainly because the final step of speech production is carried out by the finely tuned configuration and cooperation among articulators at the upper surface of the vocal tract, which is densely distributed in the oral region (Ladefoged & Johnson, 2015). The observable mouth movements have been proposed to elicit the motor plan of the viewer via the putative mirror neuron system (Skipper et al., 2005, 2007), as a motor network was found to be activated when perceiving speech in an audiovisual setting, including the cerebellum and cortical motor areas relating to both planning and executing the production of speech (Skipper et al., 2005). Yet, the audio-only processing of speech seems not to activate these motor areas (e.g., Zatorre & Belin, 2001). To conclude, looking at human faces engages with a general interest in social communication (Grelotti et al., 2002), while the visual attention allocated to the mouth area relates to delicate visual speech decoding (Skipper et al., 2007), with both of which profoundly impacting the perceptual outcome of audiovisual speech perception.

Traditional studies of face processing are primarily concerned with distinguishing faces and recognizing identities (e.g., Maurer et al., 2002; Pascalis et al., 2011), which is rarely

associated with language processing. Whereas, in multisensory language processing, face processing is a key factor that cannot be ignored. To our surprise, how talking faces are processed has received very little attention. A handful of existing studies have already recognized a trend of enhanced processing of the mouth area during audiovisual speech perception tasks, indicating the criticality of mouth movements in speech production can draw more visual attention allocation and, therefore, adjust the manner of face processing (Feng et al., 2021b; Hisanaga et al., 2016; Irwin et al., 2017; Yamamoto et al., 2019). That is to say, there is very likely to be a link between these two fundamental human abilities, face processing and speech processing, which are conventionally studied separately (Rennig et al., 2020). In fact, this association has already been advocated and obtained preliminary findings.

This link has been supported by recent behavioural and neural findings. Previous studies using functional magnetic resonance imaging (fMRI) showed that there is an overlap between the responses from the posterior superior temporal sulcus (pSTS) to both voice and faces (e.g.,Belin et al., 2011; Deen et al., 2015; Rennig & Beauchamp, 2018). Deen et al. (2015) recorded the response from different subregions of STS among a group of adults to a battery of social and linguistic stimuli, including human faces and human vocal sounds (e.g., coughing, laughing, humming, sighing and speech sounds). One of the striking findings was that the face-sensitive region of pSTS also strongly responded to voice contrast bilaterally, indicating the observed “face region” was rather an area for audiovisual processing. In a more recent study, Renning and Beauchamp (2018) simultaneously recorded the eye movements and brain activities with blood oxygenation level dependent (BOLD) fMRI of participants when presented with the McGurk paradigm (Experiment 1) and talking faces (Experiment 2). Results revealed distinctions in different subregions of the pSTS, as the anterior pSTS preferred trials where participants fixated more on the mouth region of the speaker, while the posterior pSTS favoured fixations on the eye region. The anterior pSTS was further found to more strongly respond to auditory and audiovisual speech relative to the posterior subregion, indicating a

shared neural mechanism for preferring both mouth movements and audiovisual speech perception. Taken together, these findings support resembling and interconnected neural circuits for face and speech processing (Pascalis et al., 2014)

The link has also been proposed from the perspective of language development. Studies using eye-tracking techniques showed that infants experienced developmental shifts in processing talking faces. For instance, Lewkowicz and Hansen-Tift (2012) found that infants, who initially viewed more at the speaker's eyes, gradually directed more attention to the speaker's mouth by eight mouths when they started babbling. During 8 to 12 months of age, however, the speaker's mouth would no longer consume more visual attention of infants when processing a talking face in their native language. The intensified visual attention directed to the speaker's mouth by eight months has been interpreted as being driven by perceptual narrowing, a critical process of language acquisition where the ability to discriminate native speech sound at the expense of losing the sensitivity to non-native speech contrasts (Pascalis et al., 2014). In another study centring on children, Irwin et al. (2017) observed developmental progression in terms of visual attention allocation on talking faces with the comparison among adults and children aged 5–6, 7–8, and 9–10 years. Specifically, an increase in fixation dwelling on the mouth area of the speaker alongside aging was identified among the three child groups. With similar findings by Yamamoto et al. (2019), it could be concluded that the developmental adjustments of talking face processing are continuously evolving during early to middle childhood, likely driven by the demand for speech processing.

To summarize, although the processing of human faces and audiovisual speech perception have long been considered independent domains, talking face processing bridges the two. fMRI evidence suggests that face processing and audiovisual speech processing share at least some neural mechanisms, while eye-tracking studies show a developmental trend of mutual reinforcement in these areas. Given the potential close relationship between the two,

and the limited research on their direct connection, it remains unclear whether the manners of talking face processing varies across different populations and whether such variations influence audiovisual speech perception.

Chapter 3. Exploring the Role of Language Background in Audiovisual Speech Perception

3.1 Introduction

With the McGurk paradigm, a number of studies have suggested that perceivers from different linguistic backgrounds do not always adopt the same perceptual strategy for processing an identical audiovisual stimulus (de Gelder et al., 1995; Li et al., 2008; Sekiyama, 1994, 1997; Sekiyama & Burnham, 2008, 2008; Zhang et al., 2018), proposing the probability that the perceptual strategy employed for audiovisual speech perception can be cultivated by linguistic background (de Gelder et al., 1995; Sekiyama, 1997; Zhang et al., 2018). To explain this phenomenon, the complexity of the phonological aspect of a language was considered a contributing factor since the phonetic and phonological layer is at the surface of speech communication, organizing the actual sounds and speech signals (Goldstein & Fowler, 2003). Phonological complexity can be further evaluated from segmental and suprasegmental perspectives, which have been proposed to impact the cue weighing of information from auditory and visual modalities in different ways (Zhang et al., 2018).

3.1.1 Role of Phonological Complexity in Audiovisual Speech Perception

Segments refer to the discrete units in the speech stream that are identifiable either auditorily or physically (Ladefoged & Johnson, 2015, p.10), whose articulation entails the fine-tuned coordination of the vocal tract to modify the airflow into various sounds (Crystal, 2008). As a result, the articulation of segments triggers observable movements in the larynx and facial muscles, which are believed to offer rich, distinguishable visual cues for identification and intensify the visual reliance of its native speakers in audiovisual speech processing (Sekiyama & Burnham, 2008, Zhang et al., 2018). For instance, Zhang et al. (2018) observed a weaker McGurk effect among participants native to Mandarin relative to Cantonese, which is characterized by a more complex segmental phonology compared to Mandarin. Such enhanced

visual utilization was credited to the higher segmental density in Cantonese, which increases aural ambiguity and, hence, calls for compensation from visible physical movements. In the case of Japanese, quite the opposite, Sekiyama and Tohkura (1993) revealed a reduced visual bias among Japanese speakers in comparison with American English speakers in terms of the strength of audiovisual integration when perceiving syllables with mismatched audiovisual information. The observed lower dependence on the visual modality in Japanese was attributed to the simpler structure of the phonological system, namely the relatively smaller vowel and consonant inventory and the large proportion of simple-structured syllables (i.e., 70% are Vowel (V) and Consonant-Vowel (CV) syllables, Otake, 2015; Kubozono, 1995).

On the other hand, utilizing pitch to contrast lexical meanings, especially the existence of pitch accent and lexical tone, has been put forward to increase auditory reliance when deriving audiovisual perceptual outcomes. Given that pitch is primarily determined by the fundamental frequency (f_0; Hayes, 2009, p.291; Kong, 2007; Ladefoged & Johnson, 2015, p.264; Wang, 1967), changes in pitch are mainly realized by the speed of vocal fold vibration, and therefore, seemingly induce limited distinguishable visual cues (Sekiyama, 1997). Apart from the segmental attribute, the pitch accent characteristics of Japanese were also suggested to result in increased reliance on auditory information among Japanese speakers (Sekiyama, 1997). Furthermore, this tonal account has also been extended to interpret the performance of Mandarin speakers. Sekiyama (1997) measured the magnitude of the McGurk effect among 14 Mandarin speakers residing in Japan, finding it to be weaker compared to English or even Japanese speakers. The suppressed audiovisual integration in Mandarin speakers was attributed to the utilization of lexical tones, which heightened aural ambiguity, consumed more attentional resources, but failed to provide reliable visual cues for identification.

Taken together, the linguistic background of perceivers seems to subtly influence audiovisual speech perception by adjusting the weighing of auditory and visual cues. Specifically, more complex segmental phonology seems to enhance visual utilization, as the

production of segments involves the configuration and cooperation of various articulators. Conversely, using tones to contrast lexical meanings appears to consume greater auditory attention while offering limited visual support.

3.1.2 Impact of Language Background on Audiovisual Speech Processing in Noise

In terms of the shifting effect induced by noise, there is controversy over whether speakers across various language backgrounds manifest equal visual enhancement with the addition of the same amount of noise. Several studies showed that the degree of such a noise-shifting effect appeared consistent among speakers from various language backgrounds, despite their differing strength of audiovisual integration in the quiet condition. For instance, the cross-linguistic study between Mandarin and English speakers by Chen and Hazan (2009) detected a similar visual enhancement in the noisy condition compared to the quiet one, but no significant effect in terms of language background was revealed, implying a comparable increase in visual utilization was captured in an auditory-degraded condition relative to the noise-free one. Similar results were observed between Japanese and English speakers whose absolute magnitude of visual influence differed, as a consistent increase in visual influence was found (Sekiyama & Burnham, 2008). However, in another comprehensive comparison by Hazan et al. (2010), a strengthened audiovisual integration was found in Mandarin speakers in the quiet condition than their English-speaking counterparts, but their differences in the noisy condition (-10 dB SNR) did not reach significance, inferring a weaker noise-shifting effect in Mandarin speakers. To conclude, it is unclear whether speakers varying in linguistic background will experience an equivalent visual gain with a certain amount of auditory noise.

3.1.3 Language Background and Face-viewing Pattern

Facial movement remains the primary source of visual cues in our daily face-to-face communication, and therefore, face processing is strongly associated with audiovisual speech perception at both behavioural and neural levels (e.g., Skipper et al., 2011, Rennig &

Beauchamp, 2018). Existing studies suggested that the way of scanning faces, namely, face-viewing patterns, among participants are subject to language and cultural background (Miyamoto et al., 2011; Wang et al., 2015). With a face identification task, Miyamoto et al. (2011) found that Japanese-speaking participants who were prone to rely on overall resemblance relative to matching features achieved higher accuracy relative to Caucasian Americans in the subsequence speedy identity-match task, addressing the cultural influence in face-processing strategy. When it comes to conversations, the mouth and eye areas of the speakers were believed to convey crucial information regarding speech content and the social cues of the speaker (de Boer et al., 2020; Yamamoto et al., 2019). More importantly, results from eye-tracking studies further illustrated that visual attention allocated to different facial components while processing a talking face might vary due to language and cultural background. For instance, Hisanaga et al. (2016) compared the gaze bias between Japanese and English speakers when watching speech stimuli presented audiovisually, finding a significantly longer looking times directed to the speaker's mouth compared to the eyes and nose areas in English speakers only. In stark contrast, Japanese speakers tended to direct more eye gaze towards the eyes and nose relative to the mouth area from the movie onset to the audio onset. Even when the speaker was uttering, their mouth-looking time did not significantly exceed eyes- and nose-looking time. According to the authors, such a lowered visual attention to the mouth areas in Japanese speakers was linked to the face processing strategy fostered by specific language and/or cultural atmosphere, which possibly accounted for their looser demand for visual aids in audiovisual speech perception.

The differing face-viewing patterns fostered by linguistic and cultural backgrounds possibly account for the various degrees of audiovisual integration, as indexed by the McGurk effect, experienced by speakers native to different languages if the hypothesis regarding the correlation between participants' behavioural response in the McGurk design and their visual attention given to talkers' mouth area holds. An earlier study by Paré et al. (2003) did not

observe a correlation between eye fixation directed to the speaker's mouth and the likelihood of the McGurk effect among nine participants. On the contrary, more and more studies have captured positive predictability of mouth-looking time of the magnitude of the McGurk effect (e.g., Feng et al., 2021b, 2022; Gurler et al., 2015; Stacey et al., 2020). For instance, Gurler et al. (2015) included 40 participants to investigate their face-viewing patterns in a McGurk task, finding that participants with prolonged mouth-looking time exhibited a significantly stronger McGurk effect relative to those whose mouth-looking time was shorter. This positive relationship was also observed in clinical populations, especially among children with autism spectrum disorder (Feng et al., 2021b, 2022).

Accordingly, discrepancies in audiovisual speech perception might be traced back to variations in the manner of viewing faces, which is potentially varied by linguistic background. Since facial features convey distinctive amounts of speech information, the face-viewing pattern may directly affect visual intake, thereby profoundly influencing the encoding of sensory input from visual modality while generating perceptual outcomes.

3.1.4 The Current Study

The current study seeks to revisit the role of linguistic background in audiovisual speech perception by comparing the behavioural responses between Cantonese and Mandarin speakers with the McGurk paradigm. Previous cross-linguistic studies directly comparing two significantly different languages showed striking inconsistency. While Sekiyama (1997) obtained a weakened visual influence indexed by the McGurk illusion in Mandarin speakers relative to English speakers, Chen and Hazan (2009) found a comparable degree of integration between the two, and Hazan et al. (2010) and Magnotti et al. (2015) found an even stronger visual utilization in Mandarin speakers as opposed to their English-speaking counterparts. Two concerns arise from these mixed results. Firstly, it remains unclear if the tonal property of languages leads to higher auditory reliance in audiovisual speech perception. Secondly, there

are a number of uncontrollable variables associated with the factor “linguistic background”, as there is no substantial evidence to suggest that the phonological complexity of Mandarin is lower than that of English (Hazan et al., 2010; Weng et al., 2024).

The comparability between Cantonese and Mandarin is higher because they can be viewed as two variants of Chinese, sharing high commonality in phonology, vocabulary and grammar. As another widely spoken tone language, Cantonese is the *lingua franca* of Hong Kong Special Administrative Region (HK SAR, HK hereafter). Despite many similarities between the two, Cantonese is believed to be more complex than Mandarin from both segmental and suprasegmental perspectives as a consequence of a less dramatic sound change (Bauer & Benedict, 1997). From the segmental aspect, Cantonese is richer in rime inventories and allows for more base syllables (Bauer, 2016; Lee et al., 2002; Department of Chinese Language and Literature of Peking University, 2004). From a suprasegmental angle, Cantonese keeps the three checked tones (“entering tone” attached to closed syllables ending in plosives) in addition to the six unchecked tones (Bauer & Benedict, 1997), giving rise to a more complex tone system relative to the four-unchecked-tone system in Mandarin (Department of Chinese Language and Literature of Peking University, 2004).

The high comparability between Cantonese and Mandarin allows us to step forward to explore how language background, especially phonological complexity, impacts audiovisual speech perception. In the current study, the following questions were particularly pursued: 1) whether the tonal property of tone languages necessarily leads to a low level of audiovisual integration in the McGurk paradigm, 2) whether there are differences between perceptual strategies adopted by Cantonese and Mandarin speakers under conditions with and without auditory noise, 3) whether these two groups of speakers adopt different perceptual strategies for audiovisual speech perception under noisy conditions, and 4) how their behavioural

responses are linked to their face-processing patterns, particularly the time spent looking at the speaker's mouth area?

3.2 Methods

3.2.1 Participants

Twenty young adults native in HK Cantonese (ten females and ten males, mean age = 22.14, *SD* = 2.54) and twenty native in Mandarin (ten females and ten males, mean age = 22.67, *SD* = 1.26) were recruited from the university campus to participate in the current study. All Cantonese participants were born in HK with Cantonese as their first language and reported no experience of leaving HK for more than one year. For Mandarin participants, all of them were born and brought up in northern China with Mandarin as their first language. At the time of testing, none of the Mandarin-speaking participants had stayed in Hong Kong for more than four months. All participants had normal or corrected-to-normal vision and none of them reported any auditory disorders. Each participant signed a written consent and received monetary compensation for their participation. The detailed gender and age information of both groups of participants is displayed in Table 1.

Table 3.1 The gender and age information of Cantonese- and Mandarin-speaking participants.

Group	**N (Female/Male)**	**Ages (Range, in year)**	
		Mean	SD
Cantonese	20 (10/10)	22.14 (18.27–26.77)	2.54
Mandarin	20 (10/10)	22.67 (18.60–24.70)	1.26

3.2.2 Stimuli

Four speakers, with two native speakers of HK Cantonese (one female) and two native speakers of Mandarin (one female), were invited to videotape their articulation process while uttering /ba/ [pa˥], /da/ [ta˥], /ga/ [ka˥] in their native language with a frame rate of 30 f/s and a resolution of 1080 pixels. The recordings were conducted in a sound-attenuated room. Speakers' faces and necks were set against a background in solid colour. For the congruent stimuli, the original videos were adopted. While for the incongruent stimulus, the auditory component of /ba/ was dubbed on the muted visual component of /ga/ (AbVg). Pink noise at 10 dB and -10 dB SNRs was added to the auditory component of both congruent and incongruent stimuli using a MATLAB script. The intensity of the auditory component in each stimulus was scaled to 70 dB based on root mean square.

3.2.3 Procedure

The study was conducted in a sound-attenuated room at the university. Participants were required to sit 60 cm away from the monitor of a 23-inch Tobii Pro TX300 eye tracker, whose resolution was 1920 × 1200 pixels. Stimuli were presented using E-prime 3.0, with sound displayed through professional headphones.

Using E-prime 3.0, the experiment consisted of four sessions, each presenting videos of a single speaker. Each session included a practice phase preceding the formal experiment during which three congruent trials free of auditory noise were presented twice at random. In the formal experimental phase, each block contained 12 trials, consisting of four stimuli (three congruent stimuli and one incongruent stimulus) presented in three auditory conditions (quiet, 10 dB SNR, and -10 dB SNR). Each block was repeated five times. Participants were allowed a 10-second break after each block and a five-minute rest after each session. For each experimental trial, participants were presented with a fixation cross (800 ms), followed by a black screen (1000 ms), a stimulus (2000 ms) and a response screen (infinitive). After watching

the video, participants were required to make a behavioural response from three options: /ba/, /da/ and /ga/ by pressing the first, third and fifth buttons from the left on Chronos, respectively. Participants took approximately 60 minutes to complete 240 trials. A sample trial is shown in Figure 3.1.

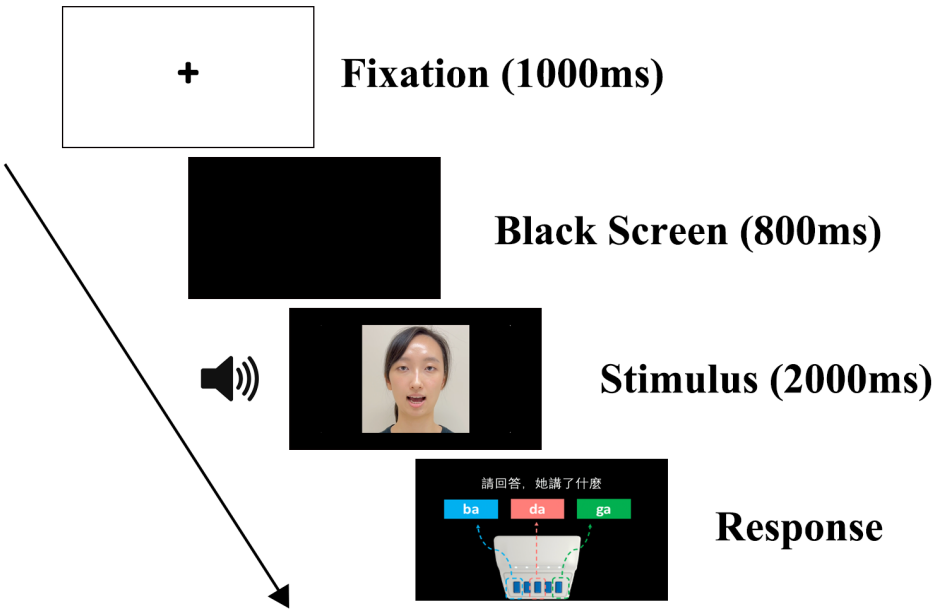


Figure 3.1 A sample trial of the experiment.

Given that the eye movements of participants were simultaneously recorded with the Tobii Pro TX300 eye tracker at a sampling rate of 300 Hz, a nine-point calibration method was adopted prior to the formal experimental session. A calibration result was accepted only if a participant could accurately fixate on all nine points or recalibration would be reperformed. When the experiment commenced, the eye gaze data, including information on which predefined area-on-interest (AOIs) was being fixated by participants, were automatically documented and stored using the TET packages provided by the E-prime Extension for Tobii (Psychology Software Tools, Inc). Figure 3.2 displays a sample of the definition of AOIs adopted for eye movement data analysis.

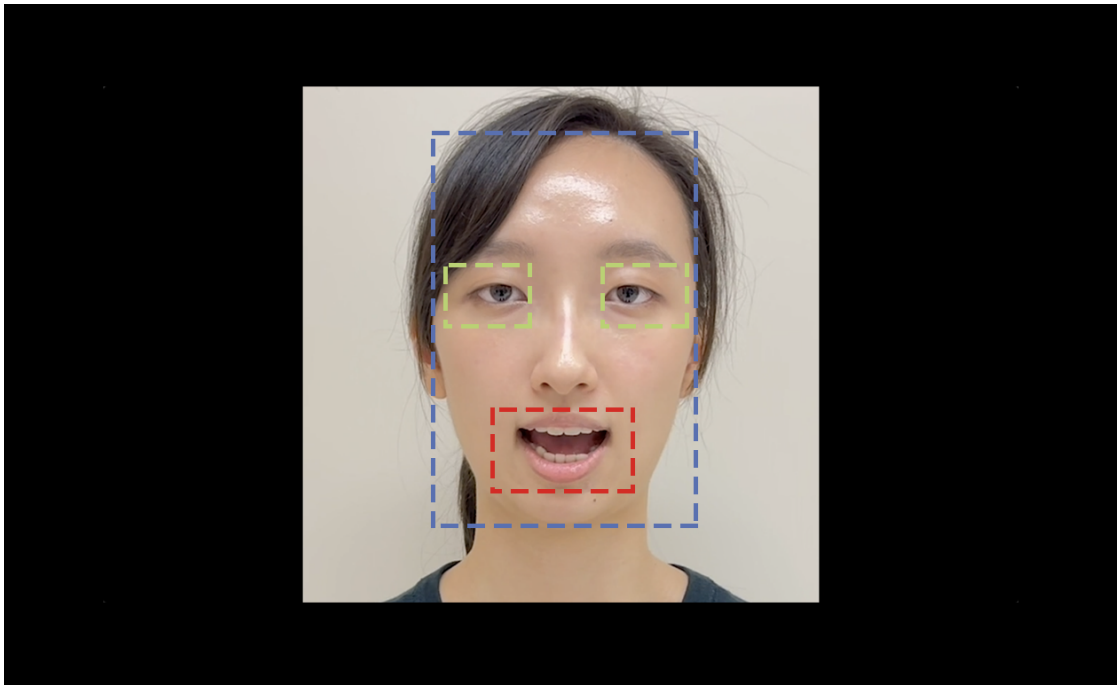


Figure 3.2 A sample of the definition of AOIs adopted for eye movement data analysis. The three AOIs were the mouth, eyes (both left eye and right eye) and the other area of the speaker's face.

3.2.4 Data analysis

All the statistical analyses were carried out in R (R Core Team, 2024).

3.2.4.1 Behavioural data

Mixed-effects models were fitted for behavioural data analysis using the *lme4* package (Bates et al., 2015), which allowed for the inclusion of random intercepts for Participant and Speaker, respectively. A generalized linear mixed model (GLMM) with a logit link function was constructed for responses to congruent stimuli. The correctness of a response served as the dependent variable (correct response = 1, incorrect response = 0). In this model, Language (Cantonese vs. Mandarin), Noise Level (quiet, 10 dB SNR and -10 dB SNR), Stimulus Type ("Ba," "Da," and "Ga"), Mouth-looking Time, and their interactions acted as fixed effects. Model construction followed a backward stepwise selection approach and started with the full model containing all the fixed effects together with by-participant and by-speaker random intercepts. In each following step, the full model was compared with a simplified model with

a specific fixed factor excluded using likelihood ratio tests (LRTs). If the inclusion of a term did not significantly improve the model's goodness-of-fit, the factor was excluded from the final model (Barr et al., 2013).

For the response to incongruent stimuli, three GLMMs with logit link function were respectively constructed for three types of responses: audio-dominant, audiovisual-integrated and visual-dominant responses, indexed by "Ba," "Da," and "Ga", respectively. Within each model, the dependent variable was whether participants made the corresponding type of response, coded as 1 if they did and 0 otherwise. The full model, including by-participant and by-speaker intercepts together with the fixed effects of Language, Noise Level, Mouth-looking Time and their interactions, was refined through backward stepwise comparison guided by the outcome of LRTs.

Post-hoc pairwise comparisons with Bonferroni correction were conducted on the best-fitting models, balancing predictive accuracy and overfitting, using the "*emmeans*" package.

3.2.4.2 Eye-tracking data

The eye gaze data obtained from the 2000-ms stimulus window were extracted and checked for validity by participants and trials. None of the participants or trials were excluded from analysis according to a 75% threshold (Grandon et al., 2023). Subsequently, the number of fixations directed to three fixed pre-defined areas-of-interest (AOIs) was computed: mouth, eyes and other facial areas (i.e., areas other than mouth and eyes). Proportion-looking time was calculated by dividing the duration of looking at a specific AOI divided by the total duration of looking at the speaker's face area according to Feng et al. (2021).

In order to explore whether there existed differences between two groups of speakers in terms of the temporal evolution of eye fixation dwelling on the different AOIs, three generalized additive mixed models (GAMMs) were fitted in R. For each GAMM, the empirical logit of fixation towards the specific AOI was treated as the dependent variable (Barr, 2008).

Following Grandon et al. (2023), model construction followed a progressive stepwise method, starting from the simplest model that only contained the smooth of Time. Subsequently, random smooths for Participant and Item within each language group were then added to the model, followed by the fixed-effect factor of Group. Model comparisons were carried out whenever a new factor was introduced to the model on Akaike information criterion (AIC). If model goodness was not significantly improved, as informed by a lower AIC, the simpler model would be retained, or the more complex one would be kept over. Finally, the time window where the effect occurred was estimated for each GAMM.

3.3 Results

3.3.1 Identification of Congruent Stimuli

Figure 3.3 presents the identification accuracy achieved by both groups of participants under varying auditory conditions.

GLMM on the correctness of the identification of the congruent stimuli revealed the main effects of Noise Level (χ^2 (2) = 160.97, $p < .01$), Stimulus Type (χ^2 (3) = 71.86, $p < .01$), Mouth-looking Time (χ^2 (1) = 4.62, $p = .03$), together with the Language × Mouth-looking Time (χ^2 (3) = 6.21, $p = .01$), Noise Level × Stimulus Type (χ^2 (3) = 38.49, $p < .01$) and Stimulus Type × Mouth-looking Time (χ^2 (3) = 6.15, $p = .046$) two-way interactions significantly improved model fit. For the main effect of Noise Level, post-hoc pairwise comparisons indicate a significantly lower probability of making a correct response in any auditory condition when compared to another with higher intelligibility (all $p < .01$). The main effect of Mouth-looking Time revealed positive predictability, indicating a benefit for making accurate response (β = 1.43, SE = .49, z = 2.91, $p < .01$) with a considerable effect size (OR = 4.20). For the Noise Level × Stimulus Type interaction, post-hoc pairwise comparisons showed that participants could identify the three congruent stimuli comparably in both quiet and 10 dB SNR conditions.

While in -10 dB SNR condition, however, the identification accuracy for “Ba” was the highest, followed by “Ga”, with the lowest for “Da” (all *p*s < .05).

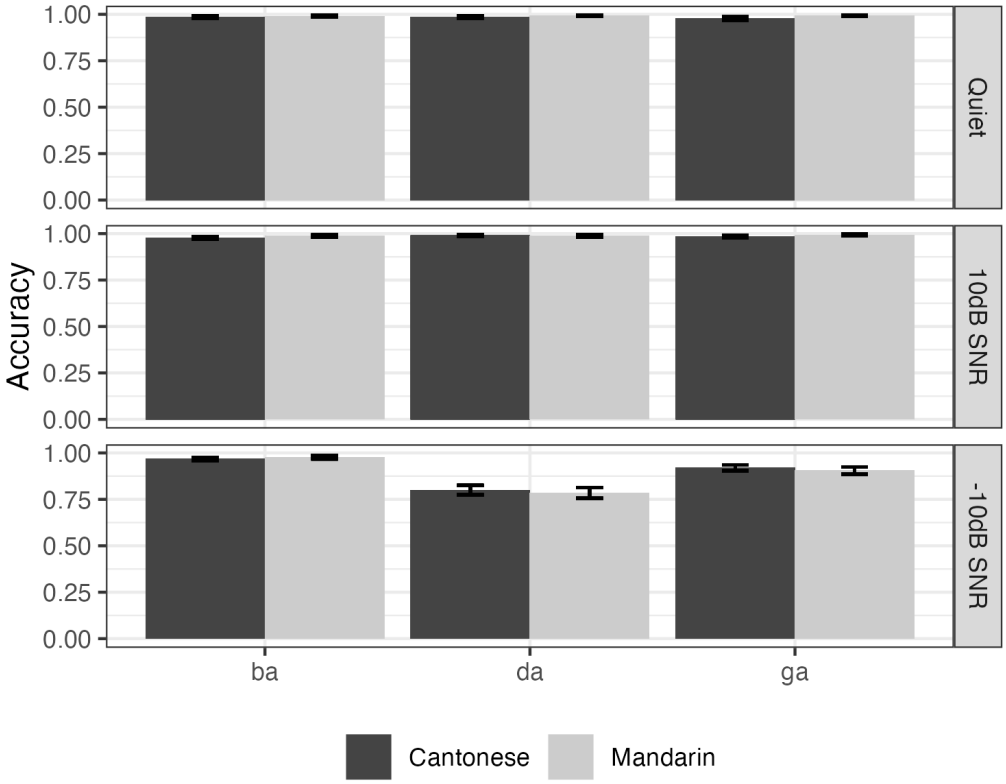


Figure 3.3 The identification accuracy achieved by Cantonese- and Mandarin-speaking participants across stimulus types under three auditory conditions.

To better understand the Language × Mouth-looking Time interaction, LRT was implemented to examine the role of Mouth-looking Time in Cantonese-speaking and Mandarin-speaking groups separately. When analyzed in the Cantonese-speaking group, the fixed effect of Mouth-looking Time was significant (χ^2 (1) = 11.42, $p < .01$) with a small effect size (*OR* = 2.42), which was insignificant when analyzed in the Mandarin-speaking group (χ^2 (1) = 1.25, $p = .26$). The prediction from the best-fitting model was visualized in the upper panel of Figure 3.4. Likewise, the two-way interaction between Stimulus Type and Mouth-looking Time is visualized in the bottom panel of Figure 3.4. Prolonged Mouth-looking Time

was revealed to significantly predict a higher probability of successfully identifying "Ba" (χ^2 (1) = 4.67, p = .03) to a moderate degree (OR = 2.79) instead of "Da" or "Ga" (both p > .05).

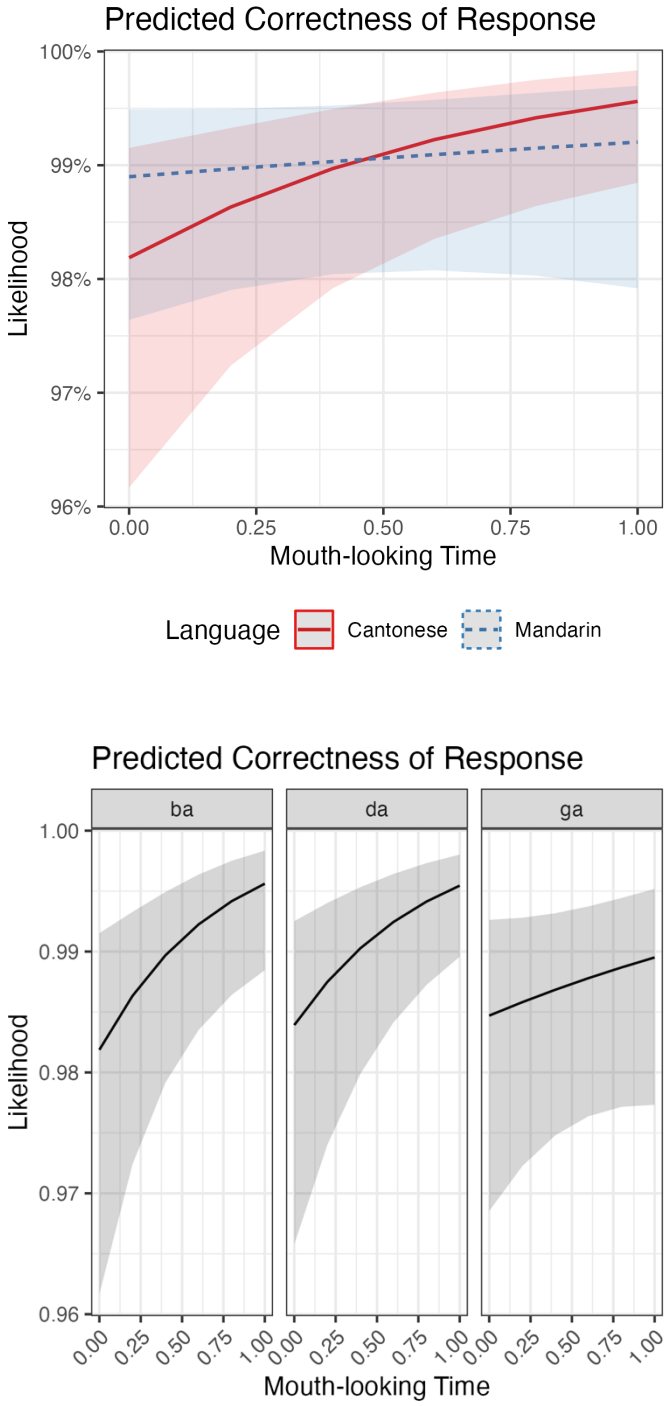


Figure 3.4 The interaction plot of Language and Mouth-looking Time (upper) and the interaction plot of Stimulus Type and Mouth-looking Time (bottom) from the best-fitting logistic GLMM.

3.3.2 Perception of Incongruent Stimuli

Figure 3.5 displays the percentage of three types of responses to the incongruent stimuli from Cantonese- and Mandarin-speaking participants.

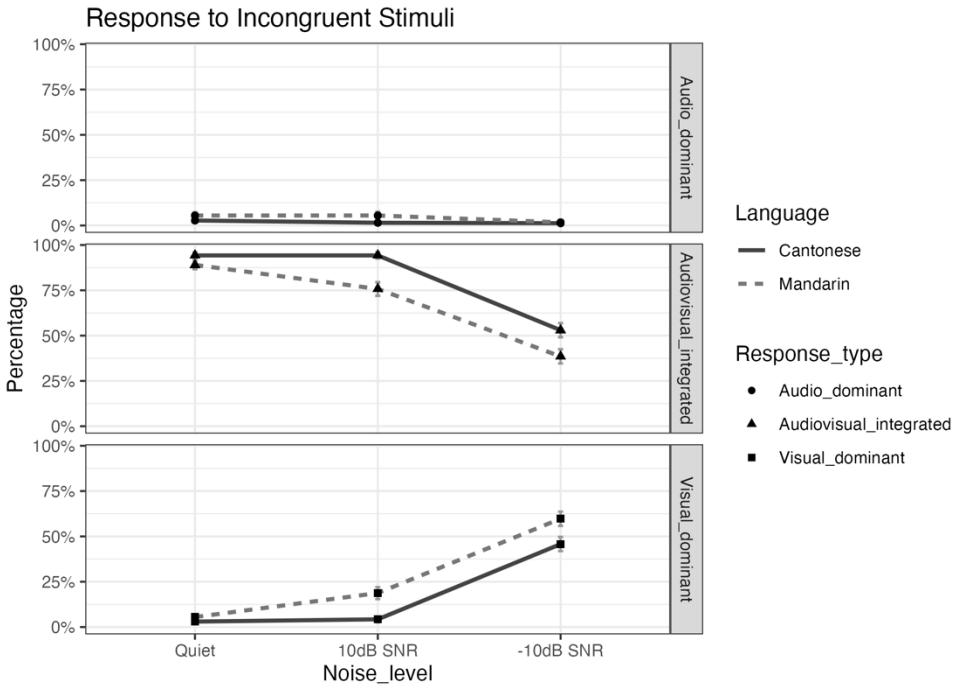


Figure 3.5 The percentage of responses to the incongruent stimuli by Cantonese- and Mandarin-speaking participants.

3.3.2.1 Audio-dominant responses

When fitting the model, the inclusion of the random effect for Speaker resulted in both the variance and the standard deviation being estimated as 0. Additionally, the inclusion of this random effect led to convergence issues, making it difficult to achieve a stable and reliable model. Thus, the random intercept was excluded from the model. GLMM revealed the significant main effects of Noise Level (χ^2 (1) = 13.73, $p < .01$) and Mouth-looking Time (χ^2 (1) = 5.99, $p = .01$). The main effect of Noise Level was yielded by the lowered probability of making an audio-dominant response in the -10 dB SNR condition when compared with the quiet condition ($p < .01$) or 10 dB SNR condition ($p = .02$). On the other hand, Mouth-looking Time was detected to significantly decreased the likelihood of making an audio-dominant response (β = -1.04, SE = 0.43, z = -2.43, p = .02) with a medium effect size (OR = .35).

Figure 3.6 illustrates the estimated effects of Mouth-looking Time derived from the best-fitting model.

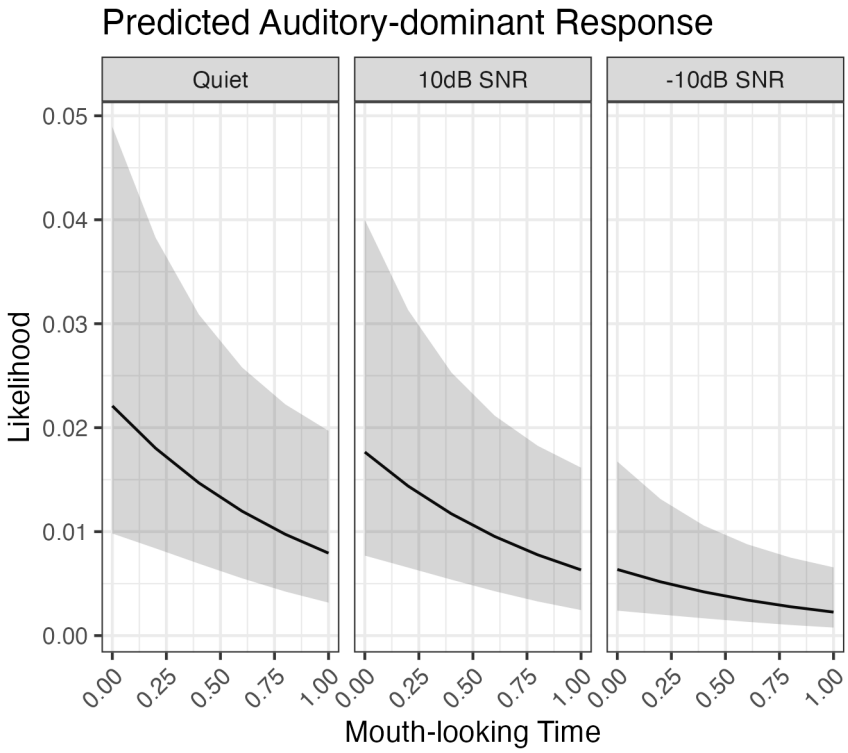


Figure 3.6 The estimated effect of Mouth-looking Time on making audio-dominant responses by the best-fitting GLMM for audio-dominant responses.

3.3.2.2 Audiovisual-integrated Responses

For the GLMM for analyzing audiovisual-integrated responses, the main effects of Language (χ^2 (1) = 8.49, p < .01) and Noise Level (χ^2 (2) = 607.78, p < .01), as well as their interaction (χ^2 (2) = 13.42, p < .01) were identified to significantly improved model's goodness. When the interaction was analyzed on Noise Level, a significantly higher probability of making an audiovisual-integrated response was demonstrated in Cantonese-speaking participants in both noisy conditions (both ps < .05), while such a difference was absent in the quiet condition (p = .15). When the two-way interaction was analyzed under Language, both groups of participants exhibited a lower probability of making audiovisual-integrated responses in the -10 dB SNR condition relative to quiet and 10 dB SNR conditions (all ps < .01). Moreover,

Mandarin group was observed a significantly decreased probability of making audiovisual-integrated response in 10 dB SNR condition compared to the quiet one ($p < .01$), which was absent in Cantonese group ($p = 1.00$). Figure 3.7 visualizes the interaction between Language and Noise Level based on the best-fitting model.

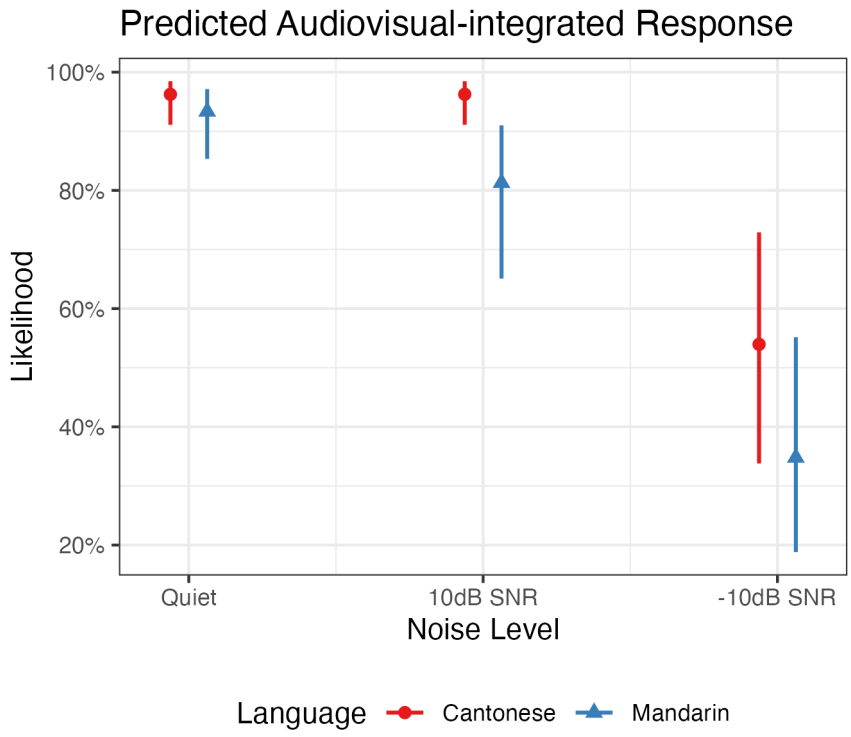


Figure 3.7 Estimated Language and Noise Level interaction based on the best-fitting GLMM for audiovisual-integrated responses.

3.3.2.3 Visual-dominant Response

The significant main effects of Noise Level and Language, as well as the Noise Level × Language interaction, was also revealed by GLMM for analyzing visual-dominant response. When analyzing the two-way interaction on Noise Level, again, no significant difference was obtained between the two groups of participants in the quiet condition. While in both noisy condition, there was a significantly higher probability for Mandarin-speaking participants to make more visual-dominant responses as opposed to their Cantonese-speaking counterparts (both $ps < .05$). When analyzed on Language, the probability of visual-dominant response did

not differ when comparing quiet and 10 dB SNR conditions for Cantonese group (p = .98), but significantly increased in 10 dB SNR condition for Mandarin group (p < .01). Besides, both groups showed an increased probability of making visual-dominant responses in the -10 dB SNR condition compared to quiet and 10 dB SNR condition (all ps < .05). Displayed in Figure 3.8 are the estimated interaction between Language and Noise Level implied by the best model for visual-dominant responses.

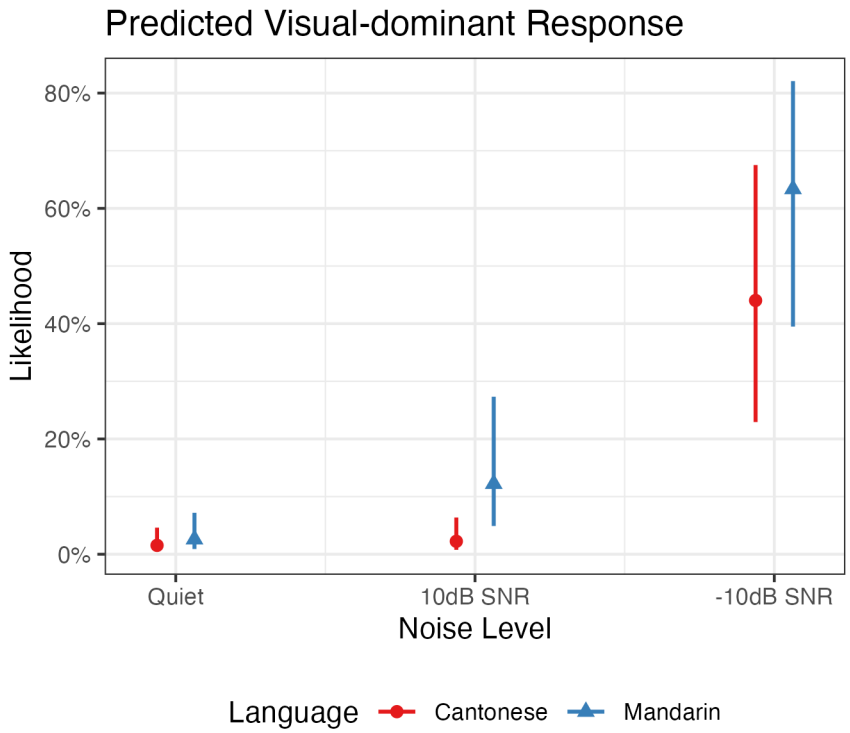


Figure 3.8 The interaction effect between Language and Noise Level estimated by the best-fitting for visual-dominant responses.

3.3.3 Time-course of Fixation Directed to Speakers' Mouth Areas

Figure 3.9 displays the temporal courses of the empirical logit of fixation organized into 50ms time bins for the three AOIs: mouth, eyes and other facial areas. Based on the 50-ms-binned temporal courses, three GAMMs were fitted to explore whether there were differences between the two groups of participants when processing the talking faces of speakers.

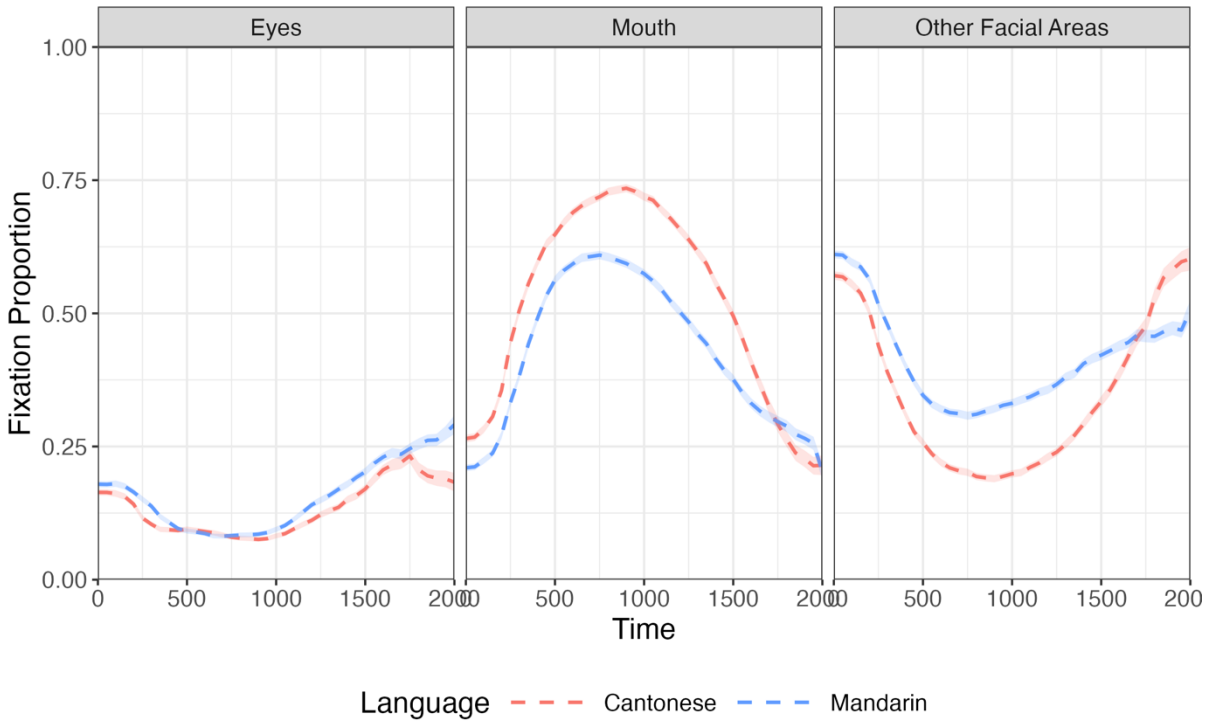


Figure 3.9. Temporal courses of fixation (empirical logit-transformed) allocated into 50 ms time bins for the three AOIs.

The best-fitting GAMM analyzing the gaze fixation directed to speakers' mouth area contained the fixed-effect factor of Language, in addition to the random smooth terms for Participant and Item within each language group. The estimation of the best-fitting model is visualized in Figure 3.10 (A) while the results of the model was summarized in Table 3.2. As informed by the parametric coefficients, the average differences between Cantonese and Mandarin groups reached marginal significance (p = .07). In terms of the fixed effect, the non-linear evolutions of fixation for both groups were both significantly different from zero (both p < .01). When visualizing the difference between Cantonese and Mandarin, a time window situated at the middle-to-late period (848–1495ms) of the 2000-ms window was estimated to be significant (Figure 3.9 (B)), which was mainly driven by the higher probability of fixating at mouth area for the Cantonese group relative to their Mandarin-speaking peers.

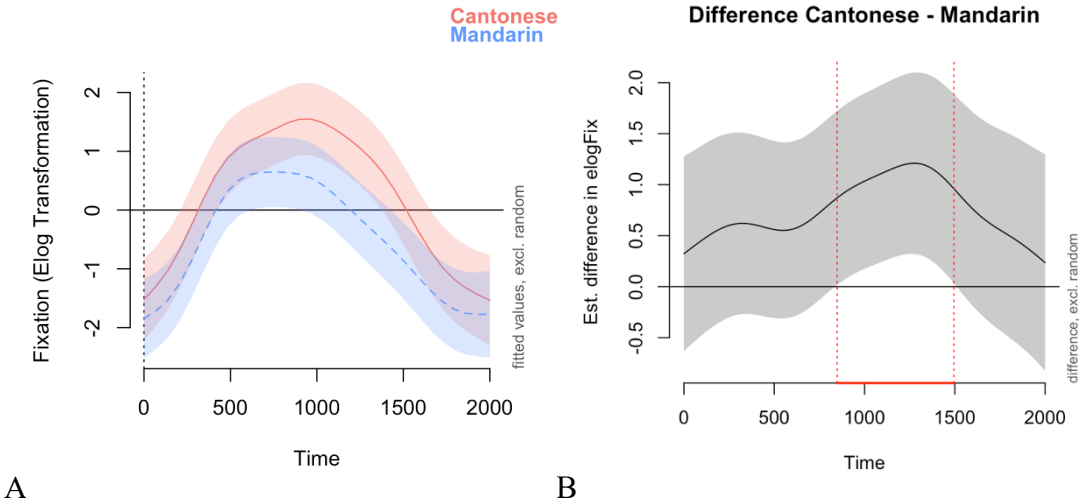
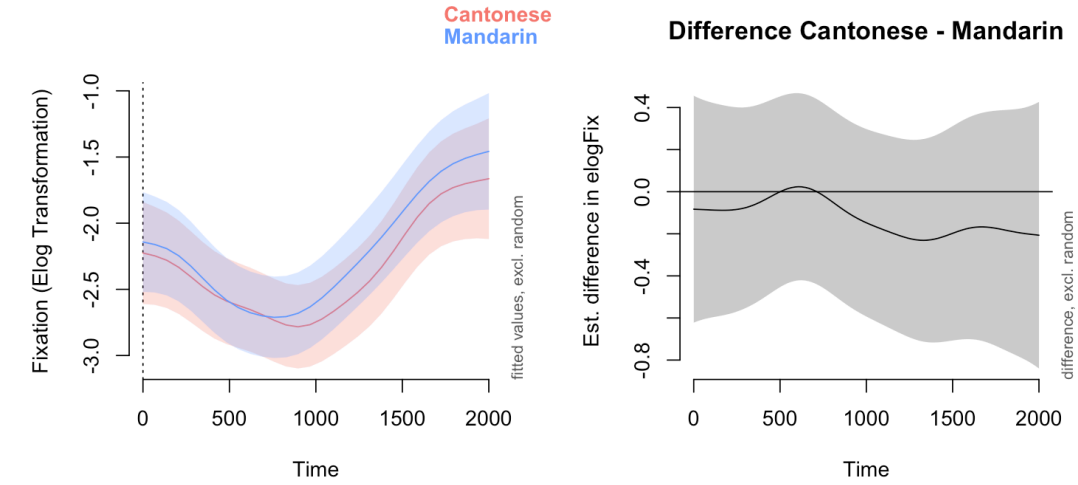


Figure 3.10 (A) Estimated temporal courses of fixation (empirical logit-transformed) towards mouth areas of speakers for Cantonese- and Mandarin-speaking participants derived from GAMM, and (B) difference between fixation of Cantonese and Mandarin groups towards mouth areas, with Cantonese group as the reference.

Table 3.2 Model summary for GAMM on fixation directed to mouth areas.

Parametric coefficients				
	Estimate	*SE*	*t values*	*p values*
(Intercept)	0.39	0.30	1.33	0.19
Language (Mandarin)	-0.76	0.42	-1.82	0.07
Smooth terms				
	Edf	*Ref.df*	*F*	*p values*
s(Time):Language (Cantonese)	8.12	8.23	44.10	<.01
s(Time):Language (Mandarin)	8.24	8.32	34.43	<.01
s(Time, Participant)	669.39	778.00	19.45	<.01
s(Time, Item):Language (Cantonese)	133.47	239.00	1.86	<.01
s(Time, Item):Language (Mandarin)	122.59	239.00	1.47	<.01

Similarly, the best-fitting GAMM analyzing the time course of fixation directed to the eyes of speakers included the fixed effect of Language, and the random smooths for Participant and for Item within Group. No significant difference was obtained in terms of the averaged values between groups, as indicated by the parametric coefficients. The smooth terms modelled for both groups were significantly different from zero, but no time windows of significant difference were detected between groups. The results of GAMM were listed in Table 3.3 and visualized in Figure 11 (A) and 10 (B).



A B

Figure 3.11 (A) Estimated temporal courses of fixation (empirical logit-transformed) towards speakers' eyes for Cantonese- and Mandarin-speaking participants derived from GAMM, and (B) difference between fixation of Cantonese and Mandarin groups towards eye areas, with Cantonese group as the reference.

Table 3.3 Model summary for GAMM regarding fixation directed to speakers' eyes.

Parametric coefficients				
	Estimate	*SE*	*t values*	*p values*
(Intercept)	-2.44	0.15	-16.51	<.01
Language (Mandarin)	0.11	0.21	0.53	0.60
Smooth terms				
	Edf	*Ref.df*	*F*	*p values*
s(Time):Language (Cantonese)	6.54	6.77	8.14	<.01
s(Time):Language (Mandarin)	6.38	6.59	10.21	<.01
s(Time, Participant)	635.77	778.00	11.67	<.01
s(Time, Item):Language (Cantonese)	132.43	239.00	1.69	<.01
s(Time, Item):Language (Mandarin)	138.26	239.00	1.90	<.01

For GAMM modelling the fixation directed to other facial areas of the speakers, the best-fitting model included both random smooths and the fixed-effect factor of Language. Indexed by the parametric coefficient, the curve for the Cantonese group had a marginally lower height relative to their Mandarin-speaking counterparts (p = .07). For the effects of smooth terms, the curves for both language groups were significantly different from zero (p < .05). A time window posited at the middle-to-late period of the stimulus (788–1495ms) was reported to identify group difference. Model results were summarized in Table 3.4 and visualized in Figure 12 (A) and 12 (B).

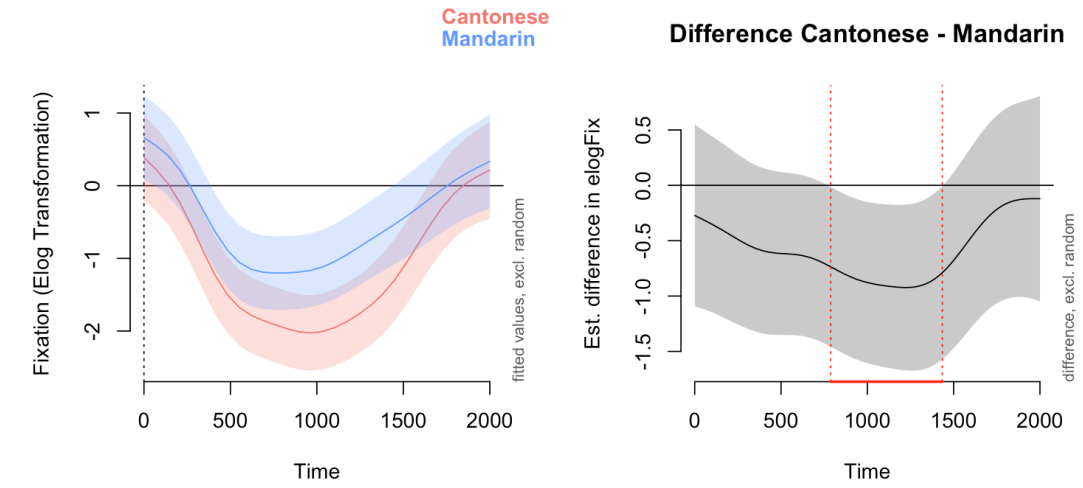


Figure 3.12 (A) Estimated temporal courses of fixation (empirical logit-transformed) towards speakers' other facial for Cantonese- and Mandarin-speaking participants derived from GAMM, and (B) difference between fixation of Cantonese and Mandarin groups towards eye areas, with Cantonese group as the reference.

Table 3.4 Model summary for GAMM regarding fixation toward other facial areas.

Parametric coefficients				
	Estimate	*SE*	*t values*	*p values*
(Intercept)	-1.18	0.25	-4.71	<.01
Language (Mandarin)	0.63	0.35	1.80	0.07
Smooth terms				
	Edf	*Ref.df*	*F*	*p values*
s(Time):Language (Cantonese)	7.86	8.02	33.21	<.01
s(Time):Language (Mandarin)	7.55	7.72	20.28	<.01
s(Time, Subject)	630.69	778.00	10.79	<.01
s(Time, Item):Language (Cantonese)	111.77	239.00	1.17	<.01
s(Time, Item):Language (Mandarin)	130.10	239.00	1.54	<.01

3.4 Discussion

In the current study, the behavioural response and eye movements of Cantonese and Mandarin speakers were compared with the McGurk paradigm in various auditory conditions. For the behavioural response pattern, comparable accuracy in identifying congruent stimuli was achieved by both groups of participants across noise levels. When responding to the incongruent stimuli, albeit no group differences reached significance in the quiet condition across response types, Cantonese group showed a significantly higher likelihood of making audiovisual-integrated responses and a lower likelihood of visual-dominant responses compared to their Mandarin-speaking counterparts in both noisy conditions. Regarding the role of mouth-looking time in audiovisual speech perception in identifying congruent stimuli, it positively predicted the accuracy only among Cantonese-speaking participants, but not among Mandarin-speaking participants. In terms of the incongruent stimuli, mouth-looking time negatively predicts the percentage of audio-dominant responses but did not predict audiovisual-integrated or visual-dominant responses, regardless of the language background. In terms of the face-viewing pattern revealed by the time courses of fixation directed to various AOIs, Cantonese speakers demonstrated a higher probability of fixating the mouth area and a lower probability of fixating on other facial areas compared with Mandarin-speakers.

3.4.1 Behavioural Results

3.4.1.1 Comparable performance in identifying congruent stimuli

The GLMM did not detect a significant group effect on the likelihood of making a correct response to congruent stimuli across noise levels and stimulus types, indicating both groups performed comparably in identifying the congruent stimuli. This result could be primarily attributed to the fact that the three CV syllables used in the current design were real-word syllables in both Cantonese and Mandarin. These syllables shared identical places and manners of articulation in these two languages, despite having different meanings (Lee & Zee, 2003;

Zee, 1991). As real words in both languages, these three syllables were frequently encountered in both language contexts, which made both groups of participants familiar with them. The nearly identical articulation of the three syllables in both languages resulted in highly similar facial and head-neck motion, offering a secure visual basis for identifying the congruent stimuli even in noisy conditions (Lalonde & McCreery, 2020).

3.4.1.2 Comparable performance in perceiving incongruent stimuli in quiet condition

For the perception of incongruent stimuli, the GLMMs did not report significant group differences under the quiet condition across response types, indicating that the incongruent stimulus without auditory noise elicited a similar response pattern in both groups of participants: the likelihood of making audiovisual-integrated response was the highest (exceeding 85%), overwhelmingly surpassing audio-dominant and visual-dominant ones. The comparable strength of the McGurk effect observed in both groups of participants suggested that the incongruent stimulus was fair for both groups of participants, as most participants found it perceptually very resembling /da/ in their respective languages. This audiovisual-integrated response, which did not correspond with inputs from neither modality, was derived from the comparable weighting assigned to audition and vision by perceivers, underscoring the functioning of the audiovisual integration mechanism (Ernst & Banks, 2002). Additionally, results from the quiet condition highlighted the prevalent experience of audiovisual integration, as measured by the McGurk illusion, in both groups of tone language speakers. This finding contradicted the proposal of Sekiyama (1997) that tone language speakers were weaker in visual utilization due to the existence of lexical tone whose production invokes less pronounced visible movement. In fact, the weaker audiovisual integration in tonal language speakers was challenged by recent findings where they showed a comparable, or even higher, magnitude of audiovisual speech perception with their non-tone-language-speaking counterparts (Chen & Hazan, 2009; de Gelder et al., 1995; Hazan et al., 2010; Magnotti et al., 2015). Taken together, Cantonese- and Mandarin-speaking participants responded similarly to our incongruent stimuli

in an auditorily quiet condition, with both of them exhibiting a strong degree of audiovisual integration in the McGurk measurement.

3.4.1.3 Distinct responses to incongruent stimuli under noisy conditions

In contrast to the results in the quiet condition, significant language effects were obtained in both noisy conditions where Cantonese-speaking participants, compared to their Mandarin-speaking counterparts, were more likely to make audiovisual-integrated responses while less likely to make visual-dominant responses. More specifically, Cantonese speakers were prone to generate the perceptual outcome by taking inputs from both modalities into account, while Mandarin-speaking participants tended to weigh relatively more heavily on unimodal visual information. This seemingly echoed the findings from Zhang et al. (2018) where an enhanced audiovisual speech integration was also found in the Cantonese group compared to the Mandarin group. However, note that Mandarin participants in Zhang et al. (2018) made significantly more audio-dominant responses, differing from the current study. Such a discrepancy might be due to the differing criteria for determining whether participants experienced the McGurk effect. Zhang et al. (2018) included more audiovisual mismatched combinations and the key to assessing audiovisual integration was whether the labial feature of phonemes interfered with the perception of non-labials. Yet the current study adopted a narrower definition to ensure fairness of the stimuli for both groups of participants. Another reason may be the noise effect, as Zhang et al. (2018) did not include auditory noise in their design. In contrast to the quiet condition, the reliability of auditory information was lowered as it was deprived to a certain extent under noisy circumstances, leading perceivers to rely more on visual information when generating perceptual outcomes (Hirst et al., 2018; Weng et al., 2024). However, Cantonese speakers maintained a relatively higher attendance to auditory information, potentially because audition was the primary mode of daily communication that was more informative and effective (Robinson & Sloutsky, 2010), even though its intelligibility was reduced. Here, the comparatively greater bias for auditory modality in noisy conditions

might be jointly brought by the more complex segmental and suprasegmental systems in Cantonese. Firstly, as guided by previous studies, suprasegmentals, whose production involved less exaggerated facial motions, might give rise to greater auditory reliance (Sekiyama, 1997; Sekiyama & Burnham, 2008; Zhang et al., 2018). Secondly, apart from its more distinguishable visual hints, it would be the most immediate consequence of the richer segmental system to heighten aural ambiguity in communication, which potentially consumed greater attentional resources in noise. Together, Cantonese demonstrated a greater tendency to adopt an audiovisual-integrated strategy in deriving percepts under noisy conditions. Notably, while we emphasize the relative auditory bias of Cantonese speakers here, we do not consider them to be weaker in visual utilization, since the generation of audiovisual-integrated percepts also entailed visual information (See section 3.4.3 for further discussion).

Meanwhile, a varying degree of noise effect through the comparison between two groups of participants was registered. When the auditory intelligibility dropped from quiet to 10 dB SNR, a significant elevation in visual-dominant response along with decreasing audiovisual-integrated responses was found among Mandarin speakers, while such a noise-shifting effect was absent in Cantonese speakers. Combining the fact that Cantonese maintained more audiovisual-integrated responses under -10 dB SNR condition than Mandarin speakers, it could be concluded that Cantonese showed a stronger resistance to auditory noise. This finding was consistent with Hazan et al. (2010) where Mandarin speakers showing a higher degree of audiovisual integration in the quiet condition showed a reduced visual enhancement in the noisy condition than English speakers. Results from Hazan et al. (2010) and the current study both point to a possibility: populations that more frequently employ an audiovisual-integrated strategy may be less susceptible to the effects of auditory noise.

3.4.2 Eye-tracking Results

3.4.2.1 The predictability of mouth-looking time to the identification accuracy of congruent stimuli

Mouth-looking time was revealed to affect the identification accuracy of Cantonese speakers differently from that of Mandarin speakers, as increased identification accuracy could be predicted by the enlarged proportion-looking time on the speaker's mouth area in Cantonese speakers only. Though the accuracy of Mandarin speakers did not increase along with mouth-looking time, it was comparable with that of Cantonese speakers. Accordingly, it could be observed that the identification accuracy of Mandarin speakers was less likely to be influenced by mouth-looking time, which might be because they also relied on the movements of other facial regions of speakers' faces, compared to their Cantonese-speaking counterparts, to acquire sufficient linguistic cues for identification (See 3.4.1.5 for further discussion). On the other hand, since the eye tracker measured the direct foveations, Mandarin-speaking participants might relatively employ more parafoveal and peripheral vision to complete the tasks (Paré et al., 2003). For Cantonese-speaking participants, their identification accuracy was highly dependent on mouth-looking time, implying that they required finer-grained visual aids from speakers' mouths to correctly screen out the syllable that matched their perception from their more complex syllabic inventory (Lee et al., 2002). Their performance was, in turn, constrained by mouth-looking time, as their accuracy would also drop without direct visual assistance from mouth movements. This constraint potentially accounted for their comparable identification accuracy with Mandarin speakers. Agreeing with Zhang et al. (2019), the heavy dependency on visual cues of oral areas also likely stemmed from the more complex segmental phonology that called for a higher demand for direct visual compensation. Also given that visual speech information was lower in resolution (Kuhl & Meltzoff, 1988), the denser distribution of segmental phonology resulted in greater ambiguity for speech decoding, forcing Cantonese speakers habitually to fixate on the movements of speakers' mouths, the most informative region in speech cues (Skipper et al., 2005).

3.4.2.2 The predictability of mouth-looking time to the perception of incongruent stimuli

However, the role of mouth-looking time did not differ between groups in the perception of incongruent stimuli. GLMMs only revealed the negative predictability of mouth-looking time in audio-dominant responses, which was insignificant in terms of audiovisual-integrated or visual-dominant responses. Mouth-looking time was used as a proxy for attention allocation to the mouth area of the stimuli, reflecting the cognitive load associated with speech processing (Gurler et al., 2015). Thus, the decreasing audio-dominant responses might be due to the intensified visual influence brought by the increased attention allocated to the mouth area. Consistent with to Paré et al (2003), we did not find a positive correlation between mouth-looking time and the strength of the McGurk illusion as indexed by the likelihood of making an audiovisual-integrated response across auditory conditions. However, our results aligned with Gurler et al. (2015) if both audiovisual-integrated and visual-dominant responses were counted as McGurk responses (e.g., Hirst et al., 2018). Moreover, through the time course analysis on fixation, it could be observed that Cantonese speakers who allocated more visual attention to speakers' mouth areas had a higher likelihood of making audiovisual-integrated responses in the noisy conditions relative to their Mandarin-speaking participants, which will be discussed in Section 3.4.2.3.

3.4.2.3 The time courses of mouth-looking time

Comparing the time courses of proportion-looking time directed to the mouth, eyes and other facial areas of speakers' faces revealed differing manners in processing talking faces among participants with varying tonal language backgrounds. Particularly, the Cantonese group showed heightened fixation on the mouth areas while reduced fixation on other facial areas compared to the Mandarin group. Combined with the positive predictability of mouth-looking time in the congruent stimuli identification, a heavier dependence on the mouth area in Cantonese speakers was confirmed, which might be brought about by the more complex segmental phonology of Cantonese. Together with the intensified reliance on auditory modality fostered by the richer suprasegmental phonology, a perceptual preference for taking

information from both modalities into account had been subtly cultivated. This audiovisual-integrated strategy potentially provided robust support for stronger resistance to auditory noise. Compared to Cantonese-speaking participants, Mandarin speakers tended to depend on a relatively more holistic manner of processing talking faces instead of the specific features of mouth movements (Hisanaga et al., 2016; Miyamoto et al., 2011). This face-viewing fashion loosened the demand for focusing on the oral areas and allowed participants to capture the dynamic spatial relationships among various facial features. In this way, Mandarin speakers might be able to track the overall movements and changes in facial expressions more efficiently, achieving a comparable accuracy with their Cantonese-speaking counterparts in identifying congruent stimuli. Also, with multiple facial cues contributing to perception generation, participants might concentrate more on whether a stimulus was visually harmonious in speech production, resulting in higher reliance on visual input when resolving the modal conflict.

Specifically, we observed significantly extended visual bias to the mouth area with a correspondingly reduced likelihood of fixating on other facial areas among Cantonese-speaking participants, occurring during the middle-to-later period of the 2000-ms stimulus. Interestingly, since the stimuli involved were primarily contrasted in terms of the consonant part, the earlier half, instead of the later one, was more information-dense in the current design. As corresponsive to the heavy reliance on mouth movements in Cantonese speakers, the prolonged attention to the mouth areas might be induced by a significantly higher cognitive load in decoding the lip movements (Bernstein, 2012; Gurler et al., 2015). In addition, Cantonese speakers were reported to process syllables as a compact unit while Mandarin speakers managed to skilfully segment syllables into finer-grained components (e.g., onset and rime) with Pinyin mastery (Holm & Dodd, 1996; Lin et al., 2020; McBride-Chang et al., 2004). Therefore, Mandarin speakers might drift their fixation away from the mouth area, with their clear consonant awareness (Holm & Dodd, 1996), once the information density dramatically fell as the speakers' production transitioned from consonant to vowel.

3.5 Conclusion

Audiovisual speech perception of Cantonese- and Mandarin-speaking perceivers was compared using the McGurk paradigm under various auditory conditions. Both groups of typical tone language speakers, who showed no difference in identifying congruent stimuli, experienced a strong McGurk illusion when perceiving the incongruent stimuli under the quiet condition, indicating tonal property did not necessarily lead to heightened auditory reliance. The impact of language background was observed under noisy conditions where Cantonese-speaking participants, whose language is phonologically more complex, were more likely to use an audiovisual-integrated strategy when auditory input was disrupted by noise. Eye-tracking data revealed that Cantonese-speaking participants showed a greater reliance on the fine-grained visual linguistic cues offered by the mouth movements of speakers during audiovisual speech decoding, while Mandarin-speaking participants tended to adopt a more holistic face-scanning approach. The enhanced visual attention towards the mouth areas of speakers among Cantonese-speaking participants might also be attributed to the more complex phonology of their native language, potentially linked to their preference for the audiovisual-integrated strategy in processing bimodal speech stimuli in noise.

Chapter 4. Development of Audiovisual Speech Perception in Cantonese-speaking Children: Effects of Language Background and Face-processing Pattern

4.1 Introduction

Integrating audiovisual information promotes speech comprehension and facilitates communicative efficiency, with which listeners manage to construct a holistic view of the dialogue at hand (e.g., Auer & Bernstein, 2007; Bernstein et al., 2000; Mohammed et al., 2006). However, with a rich body of evidence, audiovisual integration, as a form of multisensory processing, appears not innate in human beings. Instead, it is likened to a "late bloomer" in the course of human life as young children seem to undergo a long journey to develop an adult-like perceptual manner responsive to our complex surrounding environment (Burr & Gori, 2012; Ernst, 2008). Previous studies adopting the McGurk paradigm have suggested that it takes around ten years of age for children from Indo-European-language background to experience a developmental shift in sensory dominance from preferring auditory unimodal information to taking audiovisual bimodal cues into account (Hirst et al., 2018; Tremblay et al., 2007). Yet for children from the East, the situation varies as Mandarin-speaking children have exhibited an earlier development according to a recent study (Weng et al., 2024), while Japanese-speaking children displayed an arrested development during childhood in terms of audiovisual integration (Sekiyama & Burnham, 2008). Together, the underlying mechanism that triggers this development shift remains unclear. As another key element of audiovisual speech perception, face-processing manner has also been uncovered to develop with age. Given that facial movement stands out as an important cue from the visual domain contributing to speech decoding (Meltzoff & Kuhl, 1994), the mastery of skilful audiovisual-integrated perceptual strategy might evolve alongside the maturation of face-processing manner. Although this link has been proposed by existing studies (e.g., Irwin et al., 2017; Pascalis et al.,

2014), no behavioural evidence consistent with eye-tracking data was recorded. In light of these insights, the current study sought to explore the developmental courses respective to audiovisual speech perception and face-processing manner with the combination of behavioural response and eye-tracking data from a group of tonal language speakers aged 4 to 11 years native in Cantonese, whose developmental trajectory in audiovisual speech perception was rarely studied in the literature.

4.1.1 Development of Experiencing the McGurk Illusion

With the McGurk effect, existing studies have revealed a developmental shift in sensory dominance regarding audiovisual integration: young children preferring the audio-dominant strategy in processing audiovisual speech stimuli would gradually grow into an adult-like audiovisual-integrated manner later in life. Such a developmental effect has been noted in the original paper discovering the McGurk effect (McGurk & Macdonald, 1976), as the authors recorded more audiovisual-integrated responses while fewer audio-dominant responses from adults compared to the two younger groups (i.e., 3–5- and 7–8-year-olds). This trend has been corroborated in subsequent developmental studies among English-speaking (Hirst et al., 2018; Massaro, 1984; Massaro et al., 1986), French-speaking (Dupont et al., 2005; Tremblay et al., 2007), and Hebrew-speaking (Taitelbaum-Swead & Fostick, 2016) children. Tremblay et al. (2007) found a significantly weaker McGurk effect in French-speaking 5–9-year-olds relative to their 10–14- and 15–19-year-old counterparts, suggesting that the developmental shift may occur around ten years old. In parallel, English-speaking children aged 3 to 9 years in Hirst et al., (2018) made significantly fewer audiovisual-integrated responses and more audio-dominant responses compared to their adult controls. These findings align with the proposal that optimal multisensory integration emerges late in life (Ernst, 2008; Gori et al., 2008; Nardini et al., 2008).

Nevertheless, when researchers attempted to generalize the developmental shift to children from the East, they encountered inconsistencies. Earlier cross-linguistic studies argued that this

group of children may not necessarily undergo the developmental shift. For instance, Sekiyama and Burnham (2008) investigated the audiovisual speech perception with the McGurk paradigm in Japanese-speaking and English-speaking adults and children aged 6 years, 8 years and 11 years, finding distinctive development patterns subject to language backgrounds. Specifically, a significant increase in visual influence was obtained between English-speaking 6-year-olds and 8-year-olds but not in their Japanese-speaking counterparts. This absence of developmental shift seemed to extend to Mandarin-speaking children, as Li et al. (2008) did not observe significant differences among grade-two pupils, grade-five pupils and university freshmen in terms of the strength of McGurk fusion. Similarly, Chen and Hazan (2009) also failed to detect a significant age effect from the comparison between Mandarin-speaking 8–9-year-olds and adults in the noise-free condition, though their comparison under the noisy condition revealed significant differences. Li et al. (2008) suggested that the cause of the insignificant developmental effect was similar to that observed among Japanese speakers, that is, the face-avoidance custom and the tonal background shared by both language users. However, recent studies on Mandarin speakers have provided evidence for a developmental shift in the Mandarin-speaking population. For example, Weng et al. (2024) have revisited the developmental issue of audiovisual speech perception in Mandarin-speaking children with a larger sample including preschool children and using a more refined age interval between groups. By performing the McGurk paradigm on children aged 3–8 years as well as young adults aged 18–22 years, Weng et al. (2024) observed a clear developmental shift occurring at around age five, given that 3–4-year-olds made significantly more auditory-dominant responses but fewer audiovisual-integrated responses relative to older groups in the noise-free condition. More surprisingly, none of the group differences between 5–6-year-olds and older groups have reached significance, indicating that children as young as 5–6 years have already adopted an adult-like audiovisual-integrated strategy in processing stimuli of the McGurk paradigm. Thus, the authors tended to hold that tonal-language-speaking children may not be

exempted from undergoing this development process, and the absence of the developmental shift in Mandarin-speaking children reported by existing studies might be due to its early occurrence during a period ignored by previous studies.

Collectively, the developmental shift in sensory dominance for audiovisual speech processing has been well-documented in Indo-European-language-speaking children, while it remains inconsistent for children from the East, particularly for tonal language speakers. Therefore, it is necessary to conduct further investigation among another group of tone-language-speaking children to clarify how native language will affect the occurrence and/or the trajectory of this developmental process.

4.1.2 Impact of Auditory Noise on the Development of Audiovisual Speech Perception

Recent developmental studies have highlighted an interplay between development and noise. For instance, Chen and Hazan (2009) only observed a developmental discrepancy in visual effect between 8–9-year-olds and adults in the noisy condition, suggesting that noise might amplify age differences in audiovisual speech perception. However, Hirst et al. (2018) found that younger children needed higher level of auditory noise to attain the threshold at which they made non-audio-dominant responses in 50% of trials, indicating that noise allowed for measuring a comparable strength of the McGurk effect between adults and children. Weng et al. (2024) further examined the role of auditory noise with respect to stimulus congruency, finding that auditory noise might slow down the development of identifying stimuli with congruent audiovisual speech information as children aged 5 to 6 years achieved comparable accuracy with adults in the quiet condition, but not in noisy conditions. However, consistent with Hirst et al. (2018), auditory noise was found to narrow the gap between adults and children in perceiving incongruent stimuli, as the number of audiovisual-integrated responses made by the youngest child group aged 3 to 4 years was comparable in the noisy condition, which was significantly fewer compared to the elder groups in the quiet condition. Furthermore, Weng et

al. (2024) concluded development trends follow the prediction of the statistically optimal fashion proposed by Ernst and Banks (2002), as the number of visual-dominant responses increased with age when the auditory intelligence was as low as -10 dB SNR while the reliability of vision was limitedly impacted, confirming that enhanced visual reliance due to auditory degradation increases with development.

4.1.3 Development of Talking Face Processing

In Tremblay et al. (2007), both speech and non-speech stimuli were adopted to test the developmental change in audiovisual speech perception. Notably auditory preference in 5–9-year-olds was only observed with speech stimuli using the McGurk paradigm, supporting the independent maturational courses underlying speech and non-speech audiovisual illusory effect. Such asynchrony might be rooted in the characteristics of the McGurk paradigm, which taps two key elements of social communication: speech decoding and talking face processing. Though conventionally studied as separate domains, the link between these elements has been witnessed by neural and behavioural evidence. On the one hand, the superior temporal sulcus (STS) has been shown to support multiple cognitive operations, including speech processing, audiovisual integration, and face processing, as reviewed by Hein and Knight (2008). Specifically, the posterior part of STS was identified as a vital region for integrating face and voice information by converging inputs from multiple sensory domains, which exhibited heightened activation to bimodal relative to unimodal stimuli (Belin et al., 2011; Pascalis et al., 2014; Rennig et al., 2020; Rennig & Beauchamp, 2018). Furthermore, there are considerable similarities in the functional organization of cerebral processing for faces and voices, especially evident in the existence of cortical areas selective to face or voice stimuli (Belin, 2017). As suggested by Pascalis et al. (2014), the neural organization for face and voice processing might be similar and/or interacting. From the behavioural perspective, Lewkowicz and Hansen-Tift, (2012) tracked the eye movements of 4–12-month-old infants and adults while listening to native and nonnative speech, revealing a developmental change in face processing in infants

who shifted their focus from the speaker's eyes to the mouth between four to eight months, and then back to the eyes between eight to twelve months alongside with the increasing familiarity to their native language. These findings suggest the timing of shifts in face processing manner corresponded with the development of speech perception and production, which was driven by early linguistics experience. Given these analogies, Pascalis et al. (2014) proposed a framework in which perceptual and cognitive systems, such as face and language processing, are interconnected in social communication despite emerging at different times.

The evolution of talking face processing in children has been studied using audiovisual speech perception tasks, such as the McGurk paradigm, but results have been inconsistent. For instance, Irwin et al. (2017) studied the time course of eye gaze when English-speaking 5–6-year-olds, 7–8-year-olds, 9–10-year-olds and adults completing the audiovisual perception tasks. A significant increase in looking at the speaker's mouth along with age was obtained between the ages of five and ten. Also, adult participants, who had a significantly higher probability of looking at the speaker's face, were less likely to fixate on the speaker's mouth region compared to 7–8- and 9–10-year-olds. However, the association between the development of face processing patterns and that of audiovisual speech perception was not supported by the behavioural responses in Irwin et al. (2017), probably yielded by the ceiling effect. In a more recent study conducted among Japanese-speaking 5–8-year-olds, 9–12-year-olds, and adults, mouth-looking time was observed to increase with age without exception. Again, evidence for a correlation between behavioural responses and eye movements was limited, with the only statistically significant difference being that the youngest group exhibited more audio-dominant responses compared to both elder groups (Yamamoto et al., 2019).

Overall, the similarities and interconnectedness between face and language processing are supported by behavioural and neural evidence. However, studies involving children rarely obtain corresponding behavioural and eye-tracking data simultaneously.

4.1.4 The Current Study

Compared to the substantial evidence for the developmental shift experienced by children who speak Indo-European languages, whether the existence of lexical tone might eliminate this developmental process by enhancing auditory reliance remains unclear, as results from tone-language-speaking children exhibited considerable inconsistency (Li et al., 2008; Liu et al., 2020; Weng et al., 2024). Therefore, more data from tonal language speakers are warranted to verify if the tonal property will exempt them from experiencing the developmental shift. Since audiovisual speech perception closely engages with the processing of talking face (Irwin et al., 2017; Pascalis et al., 2014; Yamamoto et al., 2019), the developmental trajectory of face processing manner during audiovisual speech perception is also of particular interest. Whereas existing studies only obtained limited behavioural evidence corroborating eye-tracking results, despite the similarities and interconnectedness between audiovisual speech perception and face processing suggested by a number of studies (Pascalis et al., 2014; Rennig & Beauchamp, 2018). This may be either related to the ceiling effect of the tasks (Irwin et al., 2017) or the minimal developmental effect in the Japanese-speaking population (Yamamoto et al., 2019). In view of these issues, the current study aims to improve the design by refining the age interval between groups to detect more precise changes in processes for speech and face, introducing auditory noise to increase the difficulties of congruent stimuli identification and to provide room for strategy shifts in perceiving incongruent stimuli. Additionally, time-course analysis was employed for eye-tracking data to specifically capture the window of developmental effect. To summarize, the current study seeks to explore the following research questions: 1) whether the capability of identifying audiovisual congruent stimuli under clear and noisy conditions in Cantonese-speaking children improves with development, 2) for the perception of incongruent stimuli in the quiet condition, whether Cantonese-speaking undergo a developmental shift in sensory dominance, 3) whether introducing auditory noise to the audiovisual stimuli affects the

development of audiovisual speech perception, and 4) whether the face-processing patterns induced during the speech perception task also exhibit a developmental effect?

4.2 Methods

4.2.1 Participants

A total of 83 Cantonese-speaking children from HK aged between 4 years 0 months and 11 years and 11 months participated in the current study. According to chronological age, children were categorized into four groups: 4–5-year-old-group (aged 4 years 0 months to 5 years 11 months), 6–7-year-old-group (6 years 0 months to 7 years 11 months), 8–9-year-old-group (8 years 0 months to 9 years 11 months), and 10–11-year-old group (10 years 0 months to 11 years 11 months). All participants were native Cantonese speakers with normal or corrected-to-normal vision, and their caregivers reported no physical, cognitive, or language impairments. Caregivers of child participants were invited to complete two questionnaires to screen for autism spectrum disorder (ASD) and attention deficit hyperactivity disorder (ADHD): the Autism Spectrum Quotient—Children's Version (AQ-Child, traditional Chinese version, Auyeung et al., 2008) and the Vanderbilt Assessment Scale (Parent Informant, American Academy of Pediatrics and National Initiative for Children's Healthcare Quality, 2002). None of the children scored above the cut-off point on either test, indicating they were not at high risk for ASD or ADHD. Additionally, 20 Cantonese-speaking adults from HK were included as controls. Written consent was obtained from all participants or their caregivers, and monetary compensation was provided for their involvement upon completing the study. Table 4.1 summarizes the characteristics of the child and adult participants.

Table 4.1 The characteristics of the child and adult participants.

Group	*N* (Female/Male)	Chronological Ages	
		Mean	*SD*
		(Range, in year)	
4–5-year-olds	20 (10/10)	5.07 (4.04–5.83)	0.62
6–7-year-olds	25 (12/13)	7.05 (6.09–7.68)	0.47
8–9-year-olds	20 (10/10)	9.19 (8.1–9.98)	0.57
10–11-year-olds	17 (10/7)	10.89 (10.02–11.98)	0.77
Adults	21 (10/11)	22.63 (18.00–27.98)	2.69

4.2.2 Stimuli

The stimuli used in the current study were a subset of those described in Chapter 3. Specifically, we only selected the stimuli produced by a 24-year-old female native Cantonese speaker for two main reasons. First, speech produced at a high pitch by female speakers is acoustically closer to "motherese" (Grieser & Kuhl, 1988), making it more familiar and comforting for children. Second, to accommodate the limited attention span of young children while ensuring sufficient repetition of stimuli, we used stimuli from only one female speaker and increased the number of repetitions to seven. The methods for generating incongruent stimuli and creating different auditory conditions were identical to those described in Section 3.2.2.

4.2.3 Procedure

The experiment was conducted in a soundproof room, with participants seated approximately 60 centimeters from the monitor of Tobii Pro TX300 eye-tracker with a resolution of 1920 × 1080 pixels. Audio tracks were played through a pair of external speakers positioned on either side of the monitor at a medium volume level. Participants were required to complete all tasks independently but under the supervision of an examiner. The experimental procedure was

modelled after Weng et al. (2024), which consisted of training sessions and a formal experimental session.

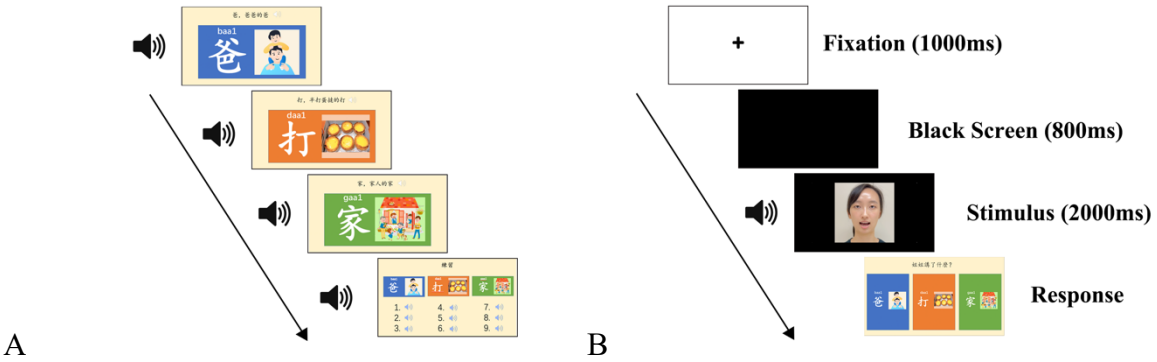


Figure 4.1 (A) Sample of the first training session and (B) a sample trial shared by the second training session and the formal experimental session.

The first training session aimed to ensure participants could distinguish the three CV syllables involved in the current design. As shown in Figure 4.1(A), participants were presented with three slides, each featuring a rectangular pattern with a unique background colour. Each pattern contained a Chinese character, its Romanized pronunciation (*Jyut Ping*, developed by the Linguistics Society of Hong Kong in 1983), and an image semantically related to the character. To reinforce the association between the patterns and the stimuli, the examiner played the syllable recording corresponding to each slide and instructed participants to repeat the syllable. At the end of this training session, participants completed a nine-trial test, where they were required to point to the correct rectangular pattern upon hearing the syllable. Feedback was provided on the correctness of each response. In the second training session, the three congruent stimuli were presented twice in random order, following the procedure of the formal experimental session. Participants who achieved full accuracy in this session were eligible to proceed to the formal experiment.

During the formal experimental session, participants' eye movements were recorded simultaneously with a Tobii Pro TX300 eye tracker at a sampling rate of 300 Hz. The E-prime Extension for Tobii was used for eye gaze collection, time synchronization, and data restoration,

using the TET packages. Before the formal experimental session commenced, the eye tracker was calibrated and validated with a five-point calibration procedure. Calibration was only accepted and proceeded to formal experiment only if all five points were accurately hit. The definitions of AOIs were consistent with those described in Chapter 3 (See Figure 3.2). The formal experiment consisted of 84 audiovisually presented trials. The three congruent stimuli ("Ba," "Da," and "Ga") and the incongruent stimulus ("AbVg") were presented under quiet, 10 dB SNR and -10 dB SNR conditions, each repeated seven times, following a procedure similar to that in Weng et al. (2024). As shown in Figure 4.1(b), each trial included an 800-millisecond fixation, a 1000-millisecond black screen, a 2000-millisecond stimulus, and an infinite response screen, presented using E-Prime 3.0. Participants were instructed to point to one of the three patterns based on their perception, and the examiner recorded their responses by pressing the corresponding button on the Chronos device: the first, third, or fifth button for "Ba," "Da," and "Ga," respectively.

4.2.4 Data Processing

4.2.4.1 Behavioural data

Since the behavioural data were not normally distributed, linear models with permutation methods were adopted for analysis, employing the "*permuco*" package in R (Frossard & Renaud, 2021). For the identification of congruent stimuli, a 5 × 3 × 3 permutation-based repeated measures ANOVA was conducted, with Age Group (4–5-year-olds, 6–7-year-olds, 8–9-year-olds, 10–11-year-olds, and adults) as the between-group factor and Stimulus Type ("Ba," "Da," and "Ga") and Noise Level (quiet, 10 dB SNR, and -10 dB SNR) as within-group factors. Post-hoc pairwise comparisons using Wilcoxon tests with Bonferroni correction were carried out where appropriate. For the perception of incongruent stimuli, a 5 × 3 × 3 repeated measures permutation ANOVA was carried out, with Age Group (4–5-year-olds, 6–7-year-olds, 8–9-year-olds, 10–11-year-olds, and adults) as the between-group factor, Response Type (audio-dominant, audiovisual-integrated, and visual-dominant) and Noise Level (quiet, 10 dB SNR

and -10 dB SNR) as within-group factors. Likewise, post-hoc pairwise comparisons using Wilcoxon tests were performed where appropriate.

Permutation-based linear regression models were built to investigate the correlation between the chronological age of child participants and their behavioural responses. For the congruent stimuli, a series of permutation-based linear regressions were carried out to examine the predictability of Age on the identification Accuracy of each stimulus (i.e., "Ba," "Da," and "Ga") under the three noise levels. For the perception of incongruent stimuli, considering that each response type indexed a distinct perceptual strategy, permutation-based linear regression models were constructed for each response type under varying auditory conditions. In each model, the Percentage of a given response served as the dependent variable, with Age as the predictor. The correlation between language ability and behavioural response patterns was also explored.

4.2.4.2 Eye-tracking data

The original eye-tracking data were processed in R, focusing on data captured during the 2000-ms stimulus. Gaze data validity was assessed for each participant and trial, and no participants or trials were excluded from data processing based on a 75% validity threshold (Grandon et al., 2023). Then, the looking time for three pre-defined AOIs—mouth, eyes and other facial areas (i.e., face areas excluding mouth and eyes)—was calculated by summing the eye fixations falling into these areas and converting them back to looking time (Feng, 2021b; Franco-Watkins & Johnson, 2011). Proportion-looking time was then computed by dividing the gaze duration for each AOI by the total gaze duration fixating on the entire face area (Feng et al., 2021b).

A 5 × 3 × 4 × 3 permutation-based repeated measures ANOVA was performed on the proportion-looking time for each AOI within the 2000-ms stimulus window. The between-group factor was Group (4–5-year-olds, 6–7-year-olds, 8–9-year-olds, 10–11-year-olds, and adults), while Noise Level (quiet, 10 dB SNR, and -10 dB SNR), Video ("Ba," "Da," and "Ga"

and “AbVg”) and AOI (mouth, eyes and other facial areas) were the within-group factors. Wilcoxon tests were adopted for post-hoc pairwise comparisons where appropriate.

To further identify the time windows where significant effect occurred, the eye fixation data, converted to empirical logits, were grouped into 50-ms time bins and analyzed utilizing GAMMs in R. Two GAMMs were constructed for analyzing the time courses of fixation toward speaker’s mouth and other facial areas separately, with empirical logit of fixation on each AOI was treated as the dependent variable (Barr, 2008). Following Grandon et al. (2023), the modelling process for each model began with the simplest model, which included only the smooth term of Time. Random smooths for Item within Group and for Participant were then added sequentially, followed by the fixed-effect factor of Group. The decision to retain each smooth term was based on whether it significantly improved the model fit as assessed by AIC comparisons. If the model goodness was boosted, the more complex model was kept over the simpler one; otherwise the simpler model would be preserved by omitting the added term. Once the model was finalized, the *plot_diff* function in the “*itsadug*” package was used to estimate the windows where the significant effects occurred.

4.3 Results

4.3.1 Behavioural Results

4.3.1.1 Identification accuracy of congruent stimuli

Permutation-based repeated measures ANOVA revealed significant effects of Group ($F(4,98) = 43.64$, permutation $p < .001$, $\eta_p^2 = .14$), Stimulus Type ($F(2,196) = 31.10$, permutation $p < .001$, $\eta_p^2 = .05$) and Noise Level ($F(2,196) = 340.01$, permutation $p < .001$, $\eta_p^2 = .56$). Additionally, significant interactions were observed for Group × Stimulus Type ($F(8,196) = 2.58$, permutation $p = .01$, $\eta_p^2 = .01$), Group × Noise Level ($F(8,196) = 14.55$, permutation $p < .001$, $\eta_p^2 = .09$), Stimulus Type × Noise Level ($F(4,392) = 29.05$, permutation $p < .001$, $\eta_p^2 = .09$), and Group × Stimulus Type × Noise Level ($F(16,392) = 2.59$, permutation $p = .001$, η_p^2

= .02). Post-hoc pairwise comparisons revealed that the significant main effect of Group was driven by the lower identification accuracy observed in 4–5-year-olds relative to all elder groups (all *p*s < .05). Further, 6–7-year-olds exhibited lower accuracy compared to 10–11-year-olds and adults, while 8–9-year-olds lower than adults (all *p*s < .05). For the main effect of Stimulus Type, the accuracy of "Ba" was significantly higher relative to "Da" and "Ga" (both *p*s < .05), while that of "Ga" was higher than "Da" (*p* = .01). The main effect of Noise Level was induced by the highest accuracy achieved in the quiet condition, decreased in the 10 dB SNR condition, and was lowest at -10 dB SNR(all *p*s < .05).

The significant three-way interaction was further examined under Noise Level. In both quiet and 10 dB SNR conditions, only the main effect of Group attained significance (both *p*s < .05), primarily driven by the significantly lower accuracy of 4–5-year-olds relative to any elder group (all *p*s < .05). When the auditory condition was as struggling as -10 dB SNR, the Group × Stimulus Type interaction reached significance ($F(8,196) = 3.09$, permutation $p = .003$, $\eta_p^2 = .08$), which was further analyzed under Stimulus Type. Children aged 4 to 7 years did not achieve adult-like accuracy for "Ba" and "Ga" (all *p*s < .05), whereas 8–9-year-olds demonstrated reduced accuracy for "Ba" and "Da" compared to adults (both *p*s < .05). No significant differences were observed between 10–11 -year-olds and adults (all *p*s > .05). When analyzed under Group, all participants except the youngest group showed a consistent pattern: accuracy was lowest for "Da," while performance on "Ba" and "Ga" was comparable. Figure 4.2 presents the identification accuracy for each congruent stimulus achieved across the five participants groups under varying auditory conditions.

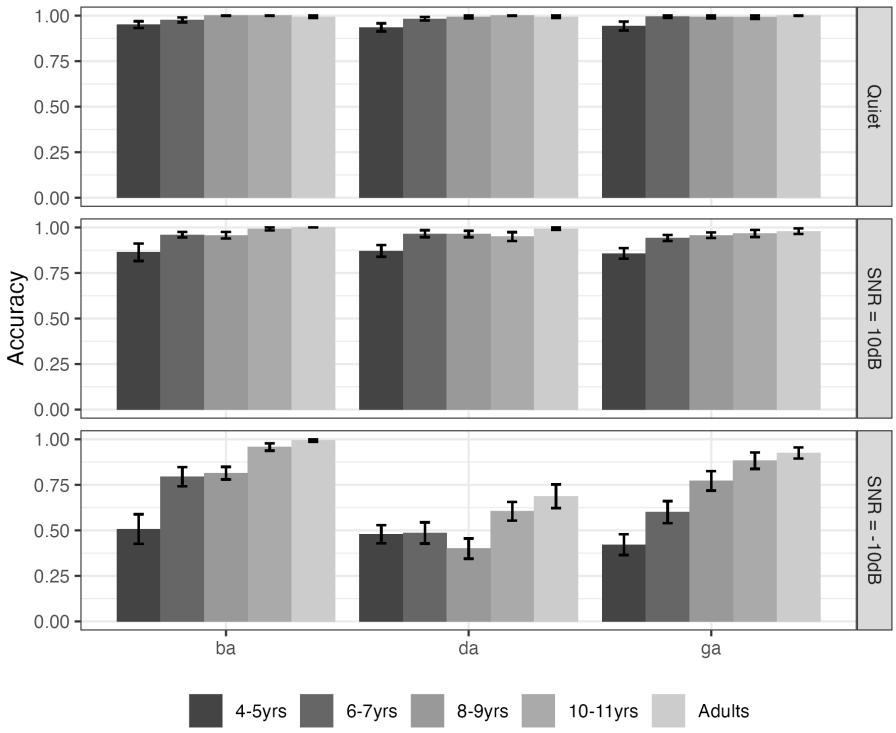


Figure 4.2 The identification accuracy of each congruent stimulus achieved by five groups of participants under varying auditory conditions.

4.3.1.2 The perception of incongruent stimuli

Figure 4.3 illustrates the percentage of the three types of responses made by five age groups in various auditory conditions. Permutation-based repeated measures ANOVA revealed the significant main effects of Response Type ($F(2,196) = 15.64$, permutation $p < .001$, $\eta_p^2 = .21$) and significant interactions, including Group × Response Type ($F(8,196) = 15.79$, permutation $p < .001$, $\eta_p^2 = .21$), Response Type × Noise Level ($F(4, 392) = 32.31$, permutation $p < .001$, $\eta_p^2 = .44$), and Group × Response Type × Noise Level ($F(16, 392) = 8.56$, permutation $p < .001$, $\eta_p^2 = .10$).

When the three-way was further analyzed under Noise Level, the Group × Response Type two-way interactions were detected across Noise Levels. When the auditory condition was free of noise, children aged 4 to 9 years made significantly more audio-dominant responses but fewer audiovisual-integrated responses relative to 10–11-year-olds and adults (all ps $< .05$), with no significant differences in visual-dominant responses. Under the 10 dB SNR condition,

child participants aged 4–9 years continued to make significantly more audio-dominant while fewer audiovisual-integrated responses compared to adult controls (all *p*s > .05). It is noteworthy that the differences in terms of audiovisual-integrated and visual-dominant responses among 6–7-year-old, 8–9-year-old and 10–11-year-old groups were insignificant (all *p*s > .05). When the auditory intelligibility dropped to -10 dB SNR, significant differences emerged only between adults and the younger participants years as children from 4–5-year-old and 6–7-year-old groups remained to exhibit more audio-dominant responses (both *p*s < .05). None of the differences between 10–11-year-olds and adults reached statistical significance (all *p*s > .05).

When analyzing the three-way interaction under Group, significant Response Type by Noise Level interaction emerged within every Group (all *p*s < .05). When the auditory intelligibility fell from quiet to 10 dB SNR, a decrease in audio-dominant responses along with a corresponding elevation in audiovisual-integrated responses could be witnessed in younger child participants aged 4 to 9 years (all *p*s < .05). Whereas such a trend was absent in 10–11-year-olds and adults (all *p*s > .05). When the auditory intelligibility lowered from 10 dB to -10 dB SNR, dramatic growth in visual-dominant responses was obtained across groups (all *p*s < .05). Meanwhile, an accompanying drop in audio-dominant responses was seen in younger participants aged 4 to 9 years (all *p*s < 0.5) but not in the older groups (both *p*s > .05w). Instead, a significant shrink in audiovisual-integrated responses was uncovered in children aged 8 to 11 years together with adult controls (all *p*s < .05).

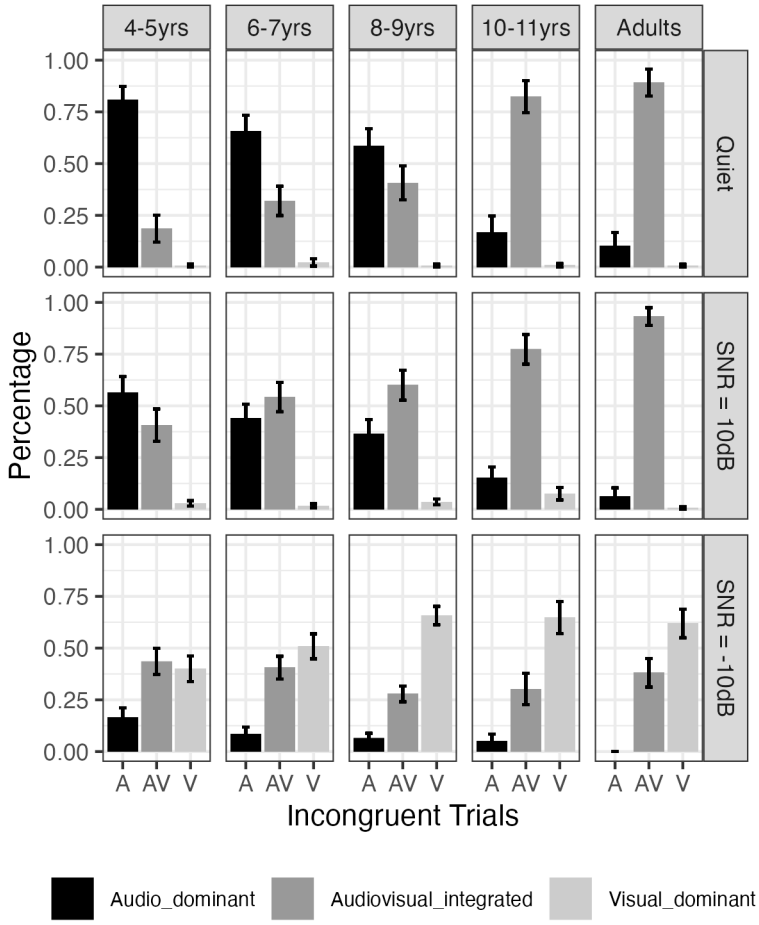


Figure 4.3 The percentage of three types of responses made by five age groups in various auditory conditions.

4.3.1.3 The predictability of chronological age

Figure 4.4 displays the permutation-based linear regression models built for examining the predictability of Age on the identification accuracy of congruent stimuli among child participants. Results demonstrated that increasing Age predicted higher accuracy in identifying congruent stimuli under most circumstances, suggesting the ability to identify CV syllables in various auditory environments was still developing in HK Cantonese-speaking children between the ages of 4 to 11. An exception was noted in the identification of "Da" under the -10 dB SNR condition, where Age did not emerge as a significant predictor (p = .50). This outcome likely stemmed from the relatively low visual saliency of the articulation of "Da", particularly when auditory information was heavily degraded.

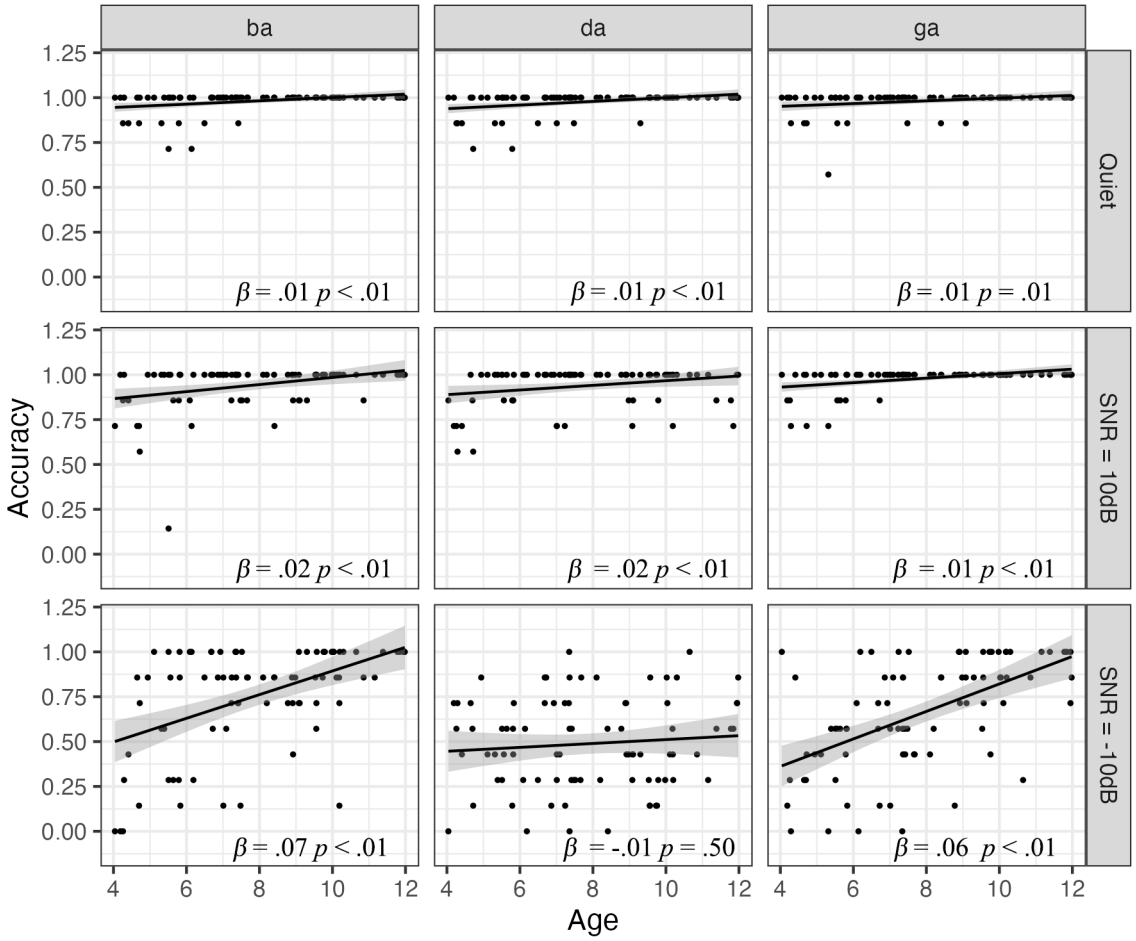


Figure 4.4 Permutation-based linear regression models built for examining the predictability of Age on the identification accuracy of congruent stimuli in child participants.

Figure 4.5 shows the linear regression models constructed for investigating the predictability of Age to different types of responses to incongruent stimuli under three auditory conditions by child participants. Results indicated that Age negatively predicted the occurrence of audio-dominant responses across all noise levels (quiet: β = -.09, SE = .02, t = -5.15, permutation p < .01; 10 dB SNR: β = -.06, SE = .02, t = -3.90, permutation p < .01; -10 dB SNR: β = -.02, SE = .008, t = -2.04, permutation p = .046). Conversely, Age was a positive predictor of audiovisual-integrated responses under both quiet and 10 dB SNR conditions (quiet: β = .09, SE = .02, t = 5.33, permutation p < .01; 10 dB SNR: β = .05, SE = .02, t = 3.22, permutation p < .01). However, under the -10 dB SNR condition, increasing Age was associated with a decrease in audiovisual-integrated responses (β = -.03, SE = .01, t = -2.42,

permutation p = .02) and an increase in elevation in visual-dominant responses (β = .05, SE = .01, t = 3.48, permutation p < .01).

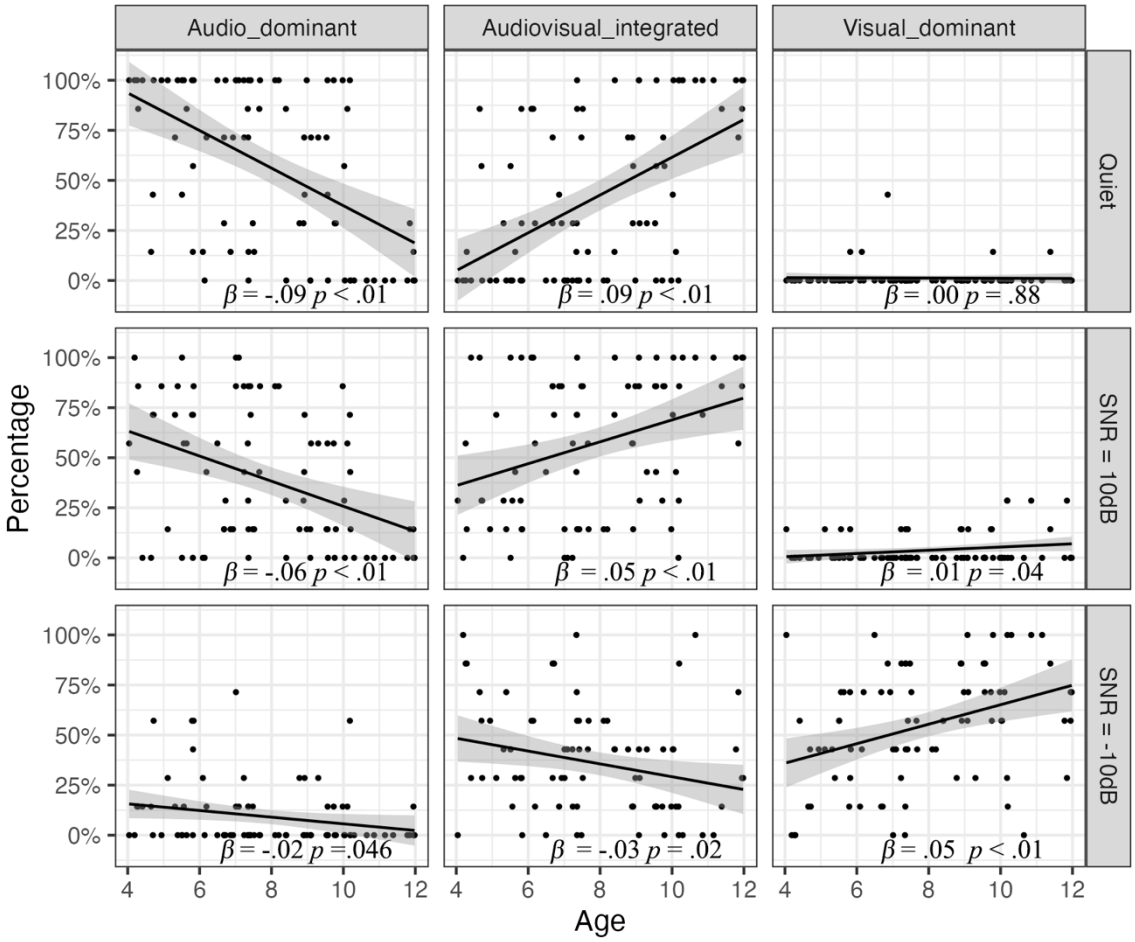


Figure 4.5 Permutation linear regression models constructed for investigating the predictability of Age to different types of responses to incongruent stimuli under three auditory conditions by child participants.

4.3.2 Eye-tracking Results

4.3.2.1 The Distribution of fixation on AOIs

Figure 4.6 displays the proportion-looking time directed to the three AOIs (i.e., mouth, eyes, and other facial areas) across groups under varying auditory conditions. Permutation-based repeated measures ANOVA indicated significant main effects of Group (F(4, 98) = 2.97, permutation p = .02, η_p^2 = .01) and AOI (F(4, 98) = 6115.26, permutation p < .001, η_p^2 = .90),

as well as significant Group × AOI ($F(8, 196) = 166.02$, permutation $p < .001$, $\eta_p^2 = .01$) and Noise Level × AOI ($F(4, 392) = 2.38$, permutation $p = .03$, $\eta_p^2 < .001$) interactions.

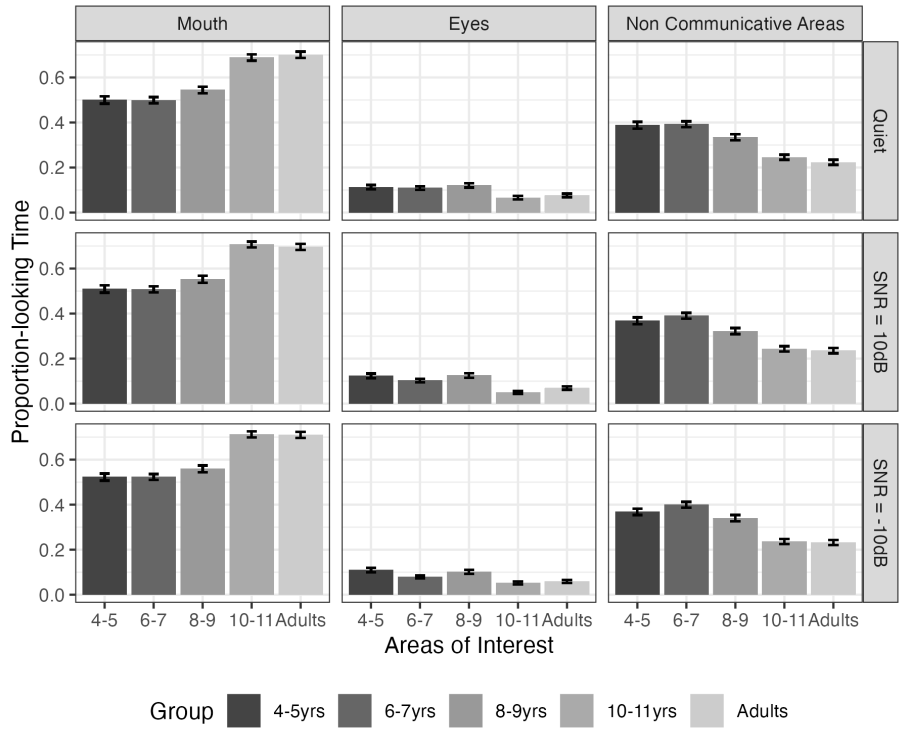


Figure 4.6 Proportion-looking time of three AOIs (i.e., mouth, eyes and other facial areas of the speaker) across groups in three auditory conditions.

When the Group × AOI interaction was further analyzed under Group, participants, regardless of age, were revealed to spend the most time fixating on the mouth area, followed by other facial areas, and the least on the eyes (all ps < .05). When analyzing under AOI, for the mouth area, post-hoc analysis showed that younger children aged 4 to 9 years, who did not differ significantly from each other (all ps > .05), spent less proportion-looking time viewing the speaker's mouth compared to the two eldest groups (all *p*s < .05), 10–11-year-olds and adults, who showed no differences with each other ($p = 1.00$). As for the other two AOIs, on the contrary, child participants aged 4 to 9 years were revealed to spend significantly more proportion time viewing the eyes and other facial areas of the speakers (all *p*s < .05) relative to the children aged 10 to 11 years and adults whose differences were insignificant (both *p*s > .05). When the Noise Level × AOI interaction was analyzed under AOI, only time spent to view eyes

area of the speaker was found to significantly dropped in the -10 dB SNR condition relative the quiet condition (p = .01).

4.3.2.2 Time-course analysis of fixation patterns

Given the results from the permutation-based ANOVA, which showed that participants predominantly directed their eye fixation to the mouth and other facial areas of the speaker's face, two GAMMs were constructed to investigate age-related differences in fixation patterns over time.

Figure 4.7 (A) displays the 50-ms-binned time-course of mouth-directed fixation across the five age groups. The best-fitting GAMM contained the random-effect terms for Participant within each Group and Item within each Group, along with a fixed-effect smooth term capturing the non-linear evolution of fixation across Groups over time. The results, summarized in Table 4.2, revealed that 4–5-year-olds, 6–7-year-olds and 8–9-year-olds exhibited significantly lower average fixation probabilities compared to the baseline level: the adult group, as indicated by the parametric coefficients (all ps < .01). In contrast, the difference between 10–11-year-olds and adults was not significant (p = .90). Furthermore, the smooth terms for all age groups were significant, suggesting that the temporal response of fixation differed meaningfully from zero (all ps < .01).

When visualizing the difference smooths between 4–5-year-old and adult groups, adult controls were found to fixate more on the mouth areas during an early-to-middle window (0–1172 ms) of the 2000-ms stimulus. Likewise, children aged 6 to 7 years also demonstrated lower fixation probabilities than adults during a time window covering 0–1051ms of the 2000-ms window. For the 8–9-year-old group, reduced mouth fixation was observed in two shorter windows (60–606 ms and 869–949 ms). No significant differences were found between 10–11-year-olds and adults throughout the 2000-ms period. Figures 4.8(A) to (D) visualize these significant time windows across age comparisons.

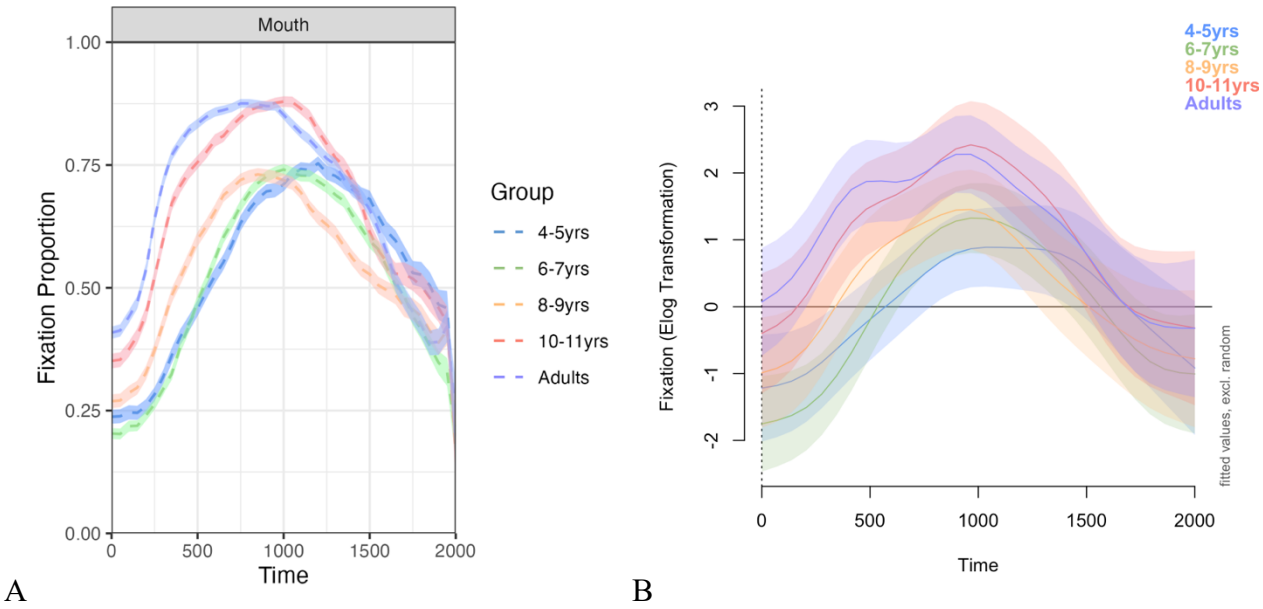


Figure 4.7 (A) 50-ms-binned time-courses of empirical logit of fixation directed to the mouth area of the speaker for five groups of participants, and (B) the estimated temporal courses of fixation (empirical logit-transformed) towards mouth areas of the speaker for five groups of participants derived from GAMM.

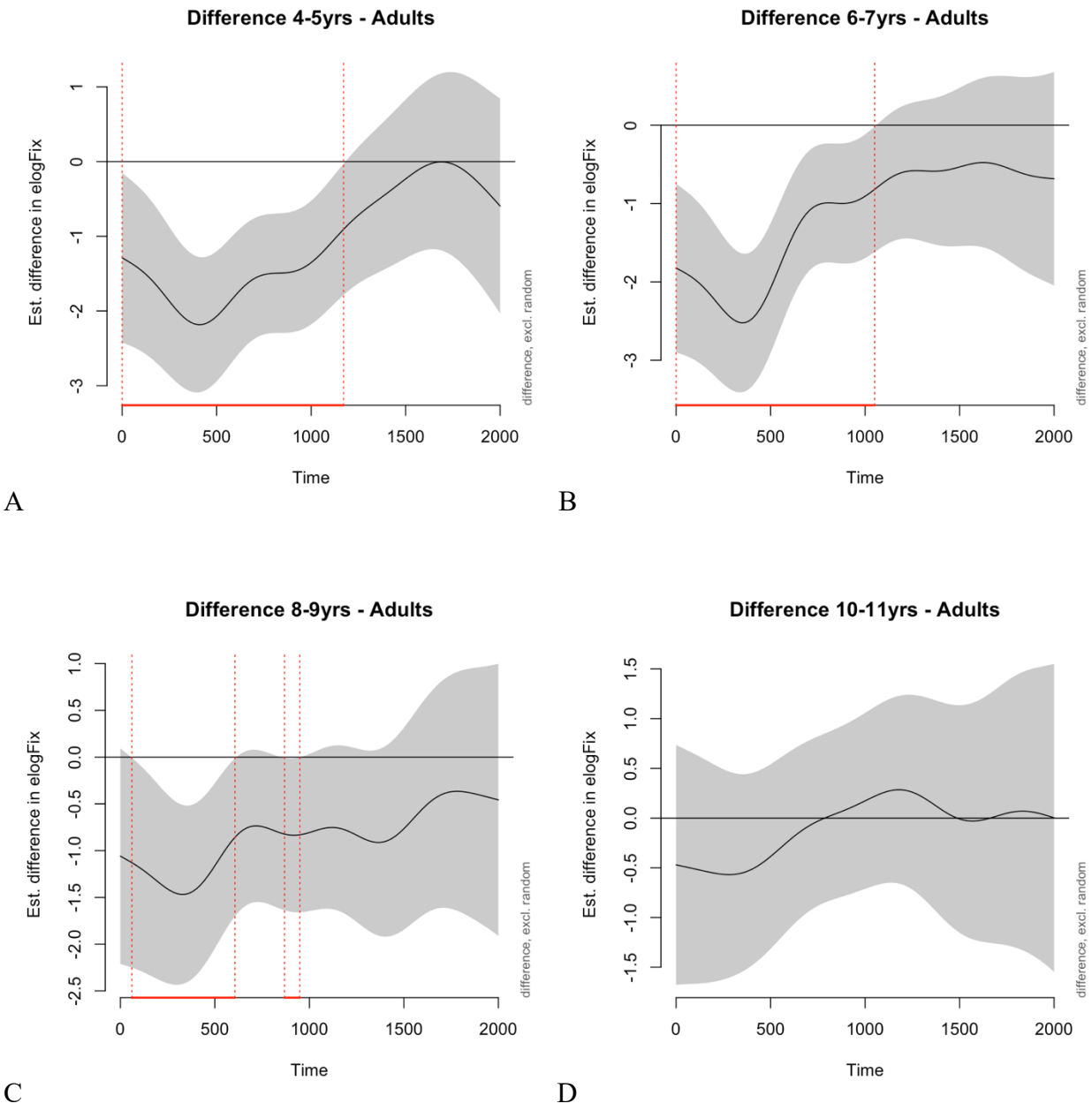


Figure 4.8 The estimated difference in fixation towards the mouth area of the speaker between (A) 4–5-year-olds and adults, (B) 6–7-year-olds and adults (C) 8–9-year-olds and adults, and (D) 10–11-year-olds and adults.

Table 4.2 Model summary for GAMM modelling the fixation (converted to empirical logits) directed to the speaker's mouth area by five groups of participants.

Parametric coefficients				
	Estimate	*SE*	*t values*	*p values*
(Intercept)	-1.76	.21	-8.63	< .01
Group (4-5yrs)	.94	.29	3.19	<.01
Group (6-7yrs)	1.07	.28	3.87	< .01
Group (8-9yrs)	.58	.29	1.95	.05
Group (10-11yrs)	.09	.31	.28	.79
Smooth terms				
	Edf	*Ref.df*	*F*	*p values*
s(Time):Group (4-5yrs)	8.31	8.84	12.39	<.01
s(Time):Group (6-7yrs)	8.30	8.83	13.76	<.01
s(Time):Group (8-9yrs)	8.11	8.75	23.30	<.01
s(Time):Group (10-11yrs)	8.71	8.71	48.51	<.01
s(Time):Group (Adults)	8.80	8.95	102.85	<.01
s(Time, Participant)	265.13	304.00	19.37	<.01
s(Time, Item):Group (4-5yrs)	328.49	419.00	4.97	<.01
s(Time, Item):Group (6-7yrs)	315.15	422.00	3.75	<.01
s(Time, Item):Group (8-9yrs)	325.04	428.00	4.28	<.01
s(Time, Item):Group (10-11yrs)	288.64	425.00	2.76	<.01
s(Time, Item):Group (Adults)	263.13	304.00	1.84	<.01

Conversely, for fixations directed to non-communicative facial areas, the younger child groups (4–5-year-olds, 6–7-year-olds, and 8–9-year-olds) demonstrated significantly higher average fixation probabilities compared to adults, as indicated by the parametric coefficients (all *p*s < .05). No significant difference was detected between 10–11-year-olds and adults (*p* = .79). All groups exhibited significant smooth terms (all *p*s < .05), indicating that the temporal response of fixations directed to other facial areas was meaningfully different from zero. The difference smooths revealed that the younger child groups (4–5-year-olds, 6–7-year-olds, and 8–9-year-olds) showed a significantly higher probability of fixating on face regions excluding mouth and eyes during the early-to-middle period of the stimulus window (4–5-year-olds: 121–1030 ms; 6–7-year-olds: 161–1152 ms; 8–9-year-olds: 282–747 ms). In contrast, the fixation patterns of 10–11-year-olds were comparable to those of adults, as indicated by a flattened difference smooth.

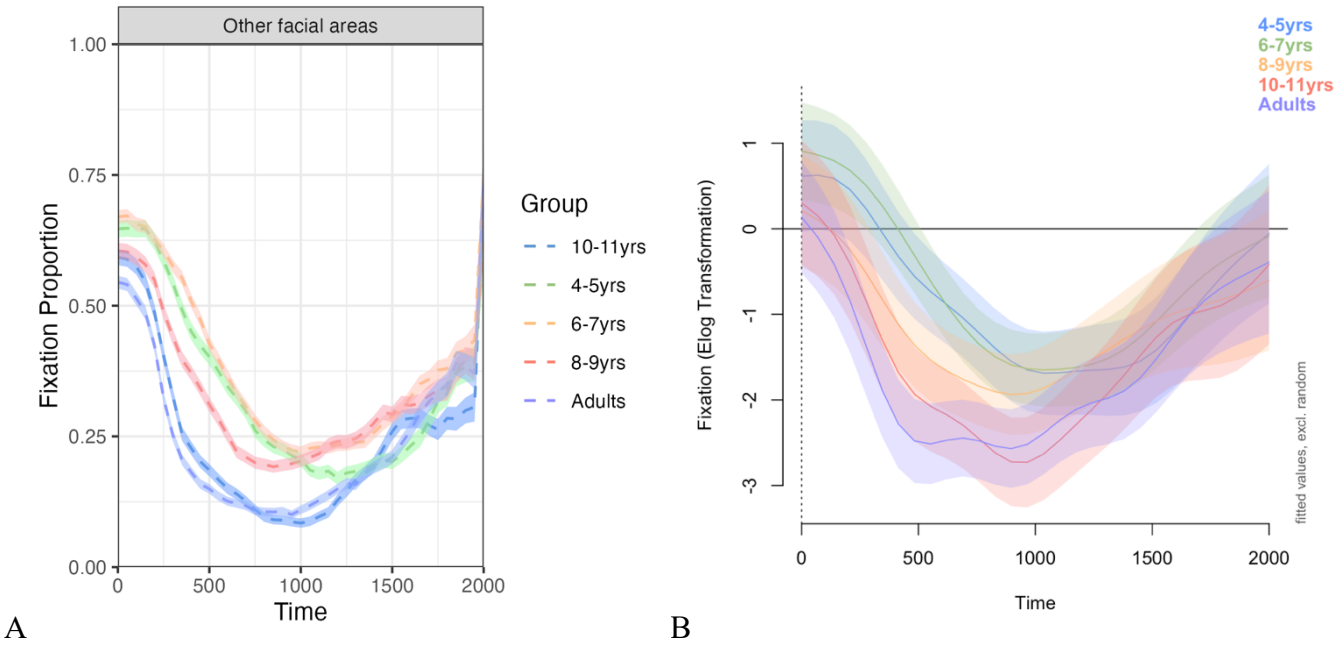


Figure 4.9 The time-courses of empirical logit of fixation directed to the other facial areas of the speaker in 50-ms time bins for five groups of participants, and (B) the estimated temporal courses of fixation towards the other facial areas of the speaker for five groups of participants derived from GAMM.

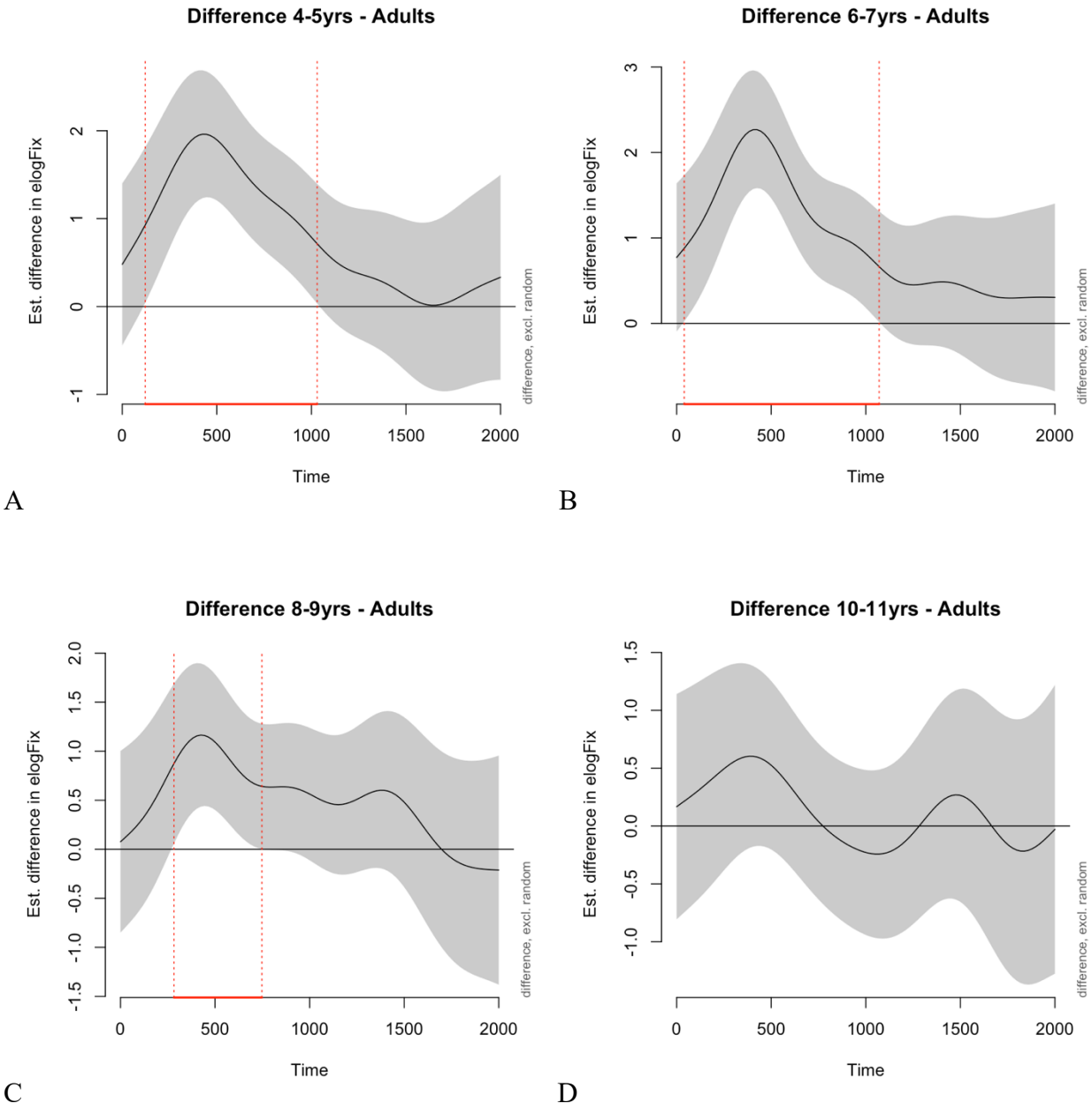


Figure 4.10 The estimated difference of fixation directed to the other facial areas between (A) 4–5-year-olds and adults, (B) 6–7-year-olds and adults (C) 8–9-year-olds and adults, and (D) 10–11-year-olds and adults.

Table 4.3 Summary of GAMM modelling the fixation (converted to empirical logits) directed to the speaker's facial areas other than eyes and mouth by five groups of participants.

Parametric coefficients				
	Estimate	*SE*	*t values*	*p values*
(Intercept)	-1.76	.21	-8.63	< .01
Group (4-5yrs)	.94	.29	3.19	<.01
Group (6-7yrs)	1.07	.28	3.87	< .01
Group (8-9yrs)	.58	.29	1.95	.05
Group (10-11yrs)	.09	.31	.28	.79
Smooth terms				
	Edf	*Ref.df*	*F*	*p values*
s(Time):Group (4-5yrs)	8.31	8.84	12.39	<.01
s(Time):Group (6-7yrs)	8.30	8.83	13.76	<.01
s(Time):Group (8-9yrs)	8.11	8.75	23.30	<.01
s(Time):Group (10-11yrs)	8.71	8.71	48.51	<.01
s(Time):Group (Adults)	8.80	8.95	102.85	<.01
s(Time, Participant)	265.13	304.00	19.37	<.01
s(Time, Item):Group (4-5yrs)	328.49	419.00	4.97	<.01
s(Time, Item):Group (6-7yrs)	315.15	422.00	3.75	<.01
s(Time, Item):Group (8-9yrs)	325.04	428.00	4.28	<.01
s(Time, Item):Group (10-11yrs)	288.64	425.00	2.76	<.01
s(Time, Item):Group (Adults)	263.13	304.00	1.84	<.01

4.4 Discussion

The current study explored the developmental progression of audiovisual speech perception in Cantonese-speaking children aged 4 to 11 years using the McGurk paradigm under various auditory conditions. For the identification of congruent stimuli, the perceptual accuracy was improved along with age, as informed by the main effect of group detected by ANOVA. Children aged 4 to 9 years exhibited lower identification accuracy compared to adults, while those aged 10 to 11 years performed on par with adults. Noise was found to exacerbate the gap between children and adults, with 6–7-year-olds and 8–9-year-olds performing comparably to adults under quiet and 10 dB SNR conditions but not under the more challenging -10 dB SNR condition. In processing incongruent stimuli, a clear developmental shift in sensory dominance was observed. Younger children aged 4–9 years were significantly more likely to exhibit audio-dominant responses and less likely to produce audiovisual-integrated responses compared to 10–11-year-olds and adults, particularly in the quiet condition. Noise seemed to narrow the gap between children and adults as children aged 8–9 years who performed differently from adults in the quiet condition showed no differences with adults under -10 dB SNR condition. Age was a positive predictor of identification accuracy for congruent stimuli in most conditions, and it also predicted shifts in sensory dominance, consistent with the statistically optimal hypothesis in multisensory processing. Regarding face processing, there was a developmental increase in visual attention directed to the speaker's mouth area. Children aged 4 to 9 years allocated significantly more visual attention to non-mouth regions compared to 10–11-year-olds and adults. Time-course analysis further demonstrated that the significant differences in fixation patterns between 4–9-year-olds and adults were primarily located in the early-to-middle stage of the stimulus presentation, with these differences diminishing as age increased. By age 10 to 11, children displayed adult-like patterns in both audiovisual speech perception and face processing, indicating developmental convergence by this age.

4.4.1 Behavioural Findings

4.4.1.1 Factors that impacted the development of identifying congruent stimuli

The significant main effect of Stimulus Type implied that the overall performance of identification varied depending on the visual saliency conveyed by the audiovisual congruent stimuli. Agreeing with Weng et al. (2024), the identification accuracy achieved by participants was the highest in "Ba" trials, followed by "Ga," and lowest in "Da." Such differences likely stem from the varying degrees of identifiability of facial movements triggered by the articulation of different consonants (van Wassenhove et al., 2005). The initial lip closure required to produce "Ba" distinguished itself from the other two stimuli whose manner of articulation initiated with lips being apart, making it most recognizable even with heavy auditory noise (van Wassenhove et al., 2005, Hirst et al., 2018). If considering mouth openness-closure as a continuum, the articulation of "Ga" marked the other end of the continuum by involving the largest openness at the beginning of articulation. On the other hand, "Da" was situated in the ambiguous middle stage, turning out to be difficult to identify based solely on visual cues (Weng et al., 2024). Combining the results from Mandarin-speaking participants, it could be concluded that "Da" was more ambiguous compared to "Ba" and "Ga" for Chinese speakers, which contrasted with the results from English-speaking participants who found it harder to identify "Ga" relative to "Da/Tha" (Hirst et al., 2018), suggesting the visual saliency of a certain phoneme varied by linguistic background.

In quiet and 10 dB SNR conditions, the impact of Stimulus Type was less pronounced compared to the -10 dB SNR condition where the auditory interference was most severe. Specifically, it was not until 9 years that Cantonese-speaking children identify "Ba" as well as adults, and that for "Ga" was 7 years. As for "Da" trials, however, child participants did not significantly underperform their adult controls, except for 8-9-year-olds. This indicates that stimuli with greater visual saliency set a higher accuracy ceiling for adults, creating a more considerable developmental gap for younger children to bridge. Regression analyses further

supported this, showing that age significantly predicted the identification accuracy of "Ba" and "Ga" under the -10 dB SNR condition but not for "Da," which remained consistently low across ages. Consistent with previous studies, noise was uncovered to delay the development of identifying congruent stimuli (Weng et al., 2024).

When the auditory noise was absent or mild, significant developmental progress was observed with 6 years of age, given that Cantonese-speaking 4–5-year-olds encountered difficulties in identifying the three congruent stimuli, while elder children aged 6–10 did not differ from their adult controls in these contexts. However, in the -10 dB SNR condition, it was not until age ten that Cantonese-speaking children attained adult-like accuracy, as children aged 8–9 years still exhibited substantial discrepancies from adults. This delay can be attributed to the collaborative nature of vision and audition in facilitating congruent stimuli identification. The interference of auditory noise disrupts this balance, forcing greater reliance on visual cues, which convey linguistic information at a lower resolution (Kuhl & Meltzoff, 1988). Consequently, children require an extended developmental period to refine their audiovisual integration skills for effective speech decoding under adverse auditory conditions (Elliott, 1979; Gijbels et al., 2021; Johnson, 2000) .

4.4.1.2 The perception of incongruent stimuli

4.4.1.2.1 Perceiving the incongruent stimuli in the quiet condition

The results from perceiving incongruent stimuli in the quiet condition demonstrated a clear developmental shift in sensory dominance in audiovisual speech perception among Cantonese-speaking children. Specifically, children aged 4 to 9 years exhibited significantly more audio-dominant and fewer audiovisual-integrated responses compared to 10–11-year-olds and adults, consistent with findings from prior research (Hirst et al., 2018; Tremblay et al., 2007; Weng et al., 2024). These findings bolstered the universality of the developmental shift in audiovisual speech perception, which initially manifested as a preference for unimodal auditory

information in younger children. The auditory reliance might be the consequence of the earlier functional onset of the auditory system during the embryonic stage (Graven & Browne, 2008). Furthermore, the dynamic and transient nature of auditory signals grants audition an advantage in perceptual dominance during early childhood (Robinson & Sloutsky, 2004, 2010). As children grow, their perceptual and sensory systems undergo constant recalibration among different sensory organs that develop at differing rates, finally achieving a statistically optimal manner in processing sensory inputs (Ernst, 2008). In support of existing studies, our regression results confirmed that the optimal fashion for audiovisual speech perception in noise-free conditions was to take bimodal information into judgment, which was opted for by elder children aged 10 to 11 years and adults (Hirst et al., 2018; McGurk & Macdonald, 1976; Tremblay et al., 2007; Weng et al., 2024).

Combining the results from Mandarin-speaking children, we tended to hold that the developmental shift would not be eliminated by the tonal property of native languages (Weng et al., 2024). The utilization of pitch to contrast lexical meanings, namely, the tonal property common in Japanese pitch accents and Chinese lexical tones, was attributed to the absent shift in the proposal of Sekiyama et al. (2003) and Li et al. (2008) because producing pitch mainly involves the vibrations of vocal folds and evokes limited visible facial movements (Hayes, 2009; Ladefoged & Johnson, 2015). As a result, it was argued that speakers of tonal languages would allocate more attentional resources to auditory input. The lack of a developmental shift among Japanese speakers was linked to the low strength of audiovisual integration shared by both children and adults, which possibly suppressed the opportunity for the developmental shift to emerge (Sekiyama & Burnham, 2008). The weaker audiovisual speech perception was attempted to extend to early studies of Chinese-speaking adults who possibly influenced by shared cultural and social customs with Japanese speakers. Note that the early study by Sekiyama (1997) was conducted in Japan, and the dialectal backgrounds of Mandarin-speaking participants varied, indicating their performance might also be influenced by other linguistic

factors. Moreover, recent studies revealed that the articulation of pitch may also trigger specific and recognizable visual hints (Burnham et al., 2022). Besides, since there is no compelling evidence to suggest that the phonological complexity of Chinese, whether it be Mandarin or Cantonese, is lower than in English, the demand for visual compensation in speech processing among Chinese speakers may not necessarily be lower than their English-speaking counterparts (Hazan et al., 2010; Weng et al., 2024). Instead, studies with larger sample sizes uncovered comparable or even enhanced audiovisual integration from Mandarin- and Cantonese-speaking participants relative to Indo-European language speakers (de Gelder et al., 1995; Hazan et al., 2010; Magnotti et al., 2015). As a result, the gap between the high dependence on audiovisual-integrated strategy in adults and the natural bias for unimodal auditory information among children made the developmental shift in Chinese-speaking children necessary.

4.4.1.2.2 Perceiving the incongruent stimuli in noisy conditions

When noise as mild as 10 dB SNR was introduced to the auditory modality, another evident development shift could be observed, which closely resembled the pattern observed in the quiet condition. Specifically, children aged 4 to 9 years made significantly more audio-dominant while fewer audiovisual-integrated responses as opposed to their adult controls, while elder children aged 10 to 11 years did not significantly differ from adults. Clearly, this mild noise was not sufficient for Cantonese-speaking 10–11-year-olds and adults to shift their perceptual strategy, leaving the audiovisual-integrated one remaining to be the optimal choice, which could be seen by their similar response pattern in both quiet and 10 dB SNR conditions. As for younger children aged 4 to 9 years, who have not fully completed the development shift, despite their preference for the audio-dominant strategy over the audiovisual-integrated one relative to adults, the 10 dB SNR condition significantly boosted their audiovisual integration, aligning with previous findings (Hirst et al., 2018; Sekiyama & Burnham, 2008; Weng et al., 2024). At this moment, the differences between 6–9-year-olds and 10–11-year-olds in terms of audio-dominant and audiovisual-integrated responses were insignificant. These results suggest that

mild noise can also promote audiovisual integration by releasing the hyper-reliance on audition in children and directing more attention to visual cues. This interplay between noise and age was further corroborated by the regression results, which demonstrated that increasing age significantly predicted fewer audio-dominant and more audiovisual-integrated responses in the 10 dB SNR condition, pointing to a developmental direction aligning with the prediction of the statistically optimal hypothesis (Ernst & Banks, 2002; Ernst & Bülthoff, 2004; Weng et al., 2024).

When the auditory intelligibility was significantly compromised at -10 dB SNR, a different developmental shift was observed: from the audiovisual-integrated strategy to a visual-dominant one. Unlike the shifts seen under quiet and 10 dB SNR conditions, this shift was driven by severe auditory noise, which drastically reduced the reliability of auditory information while elevating the importance of visual cues in perceptual processing. Across all age groups, visual-dominant responses increased as auditory reliability plummeted. For younger children aged 4 to 7 years, their audio-dominant responses were further suppressed, implying their auditory bias was further inhibited. For elder children aged 8 to 11 years and adults who mainly adopted the audiovisual-integrated strategy under 10 dB SNR condition, the lowered auditory intelligibility drove them to depend more on unimodal visual modality, which was informed by their decreasing audiovisual-integrated responses accompanying the increasing visual-dominant ones. In this context, the gap between child and adult participants was significantly narrowed since no group differences were obtained regarding audiovisual-integrated and visual-dominant response, which was probably due to the intensity of noise that forced children to disengage with auditory modality and to switch to a more adult-like visual-dominant strategy (Hirst et al., 2018; Weng et al., 2024). This trend was also confirmed by the regression models, as children increasingly adopted the visual-dominant strategy while reducing their use of the audio-dominant or audiovisual-integrated strategies with age. Again, the direction of development was towards the statistically optimal option.

4.4.1.3 The later development of audiovisual speech perception in Cantonese-speaking children

Compared to their Mandarin-speaking counterparts, Cantonese-speaking seemed to take a longer journey to experience the developmental shift, confirming that the time point where the developmental shift takes place is subject to language and social background (Weng et al., 2024). There are several explanations for the later development of Cantonese-speaking children. First, Cantonese is more complex in both segmental and suprasegmental phonology compared to Mandarin (Zhang et al., 2018), which contains more rimes, tones and base syllables and allows for more complicated syllable structures (Bauer & Benedict, 1997; Lee et al., 2002). With such a rich phonological inventory, the phonological space of a language would become crowded (Peng, 2006), the auditory proximity would be heightened (Zhang et al., 2018), while the constraints between phonological elements would become stricter (Manuel, 1987, 1990). Hence, children with limited linguistic experience might allocate more auditory attention to overcome the intensified auditory ambiguity and distinguish speech sounds. On the other hand, higher phonological density could inherently complicate visual speech perception (Bernstein, 2012). Given the comparatively lower resolution of visual speech (Kuhl & Meltzoff, 1988), the development of this competence may therefore be extended. Accordingly, the higher phonological complexity in Cantonese might postpone the disengagement with the initial auditory preference in children and lead to the later occurrence of developmental shift. Second, the later shift might be linked to the phonological training received by Cantonese-speaking children. Since Cantonese-speaking children would not acquire any phonemic coding system (e.g., *Pinyin*) as their Mandarin-speaking peers did, they exhibited weaker syllable and phoneme onset awareness, indicating a delayed capability to identify, distinguish and manipulate these units (McBride-Chang et al., 2004). Notably, the current design mainly involves syllable and phoneme onset processing. Additionally, given that previous findings further suggested that children with impaired phonological awareness presented stronger auditory bias in the McGurk paradigm (Dodd et al., 2008), the later established phonological

awareness might be conjectured to underly the slower development in audiovisual speech perception among Cantonese-speaking children. Also, the weaker visual utilization in Cantonese-speaking children might owing to the strict regulation on mask-wearing and social distancing during the COVID-19 pandemic. During the approximately 1000 days when the mask mandate was in effect, children at school were required to wear masks all the time (Cheng, 2023). Masks covered the oral region where visual linguistic cues were the richest and finest, which cut the visual inputs at their source. With the evidence of reduced visual sensitivity brought about by the COVID-19 pandemic (Charney et al., 2021; Chládková et al., 2021), the later developmental shift in Cantonese-speaking children might thus be explained.

4.4.2 Eye-tracking Findings

4.4.2.1 The development of talking face processing

As indicated by results from ANOVA, both child and adult participants exhibited a pronounced fixation preference for the speaker's mouth area, as evidenced by the highest proportion-looking time towards mouth across all age groups, noise levels, and presented videos, which likely attributable to the objective of the current study that involved speech processing (Yamamoto et al., 2019). However, the mouth-looking time measured in younger children aged 4 to 9 years was less than 10–11-year-olds and adults, while the fixation proportions directed to the eyes and other facial areas were significantly greater, indicating children might also experience developmental changes in processing talking faces. Specifically, this developmental course began with scanning talking faces in a more extensive way and gradually grew into an adult-like mouth-centered manner as children mature. This developmental trend echoed the observations in previous studies. For instance, Yamamoto et al. (2019) documented a gradual increase in mouth-directed fixation in Japanese-speaking children, while Irwin et al. (2017) reported similar developmental changes among English-speaking children. However, unlike Irwin et al. (2017), our study, similar to Yamamoto et al. (2019), did not observe a inverted U-shaped developmental pattern. Tentatively, we held that the reduced visual attention allocated

to mouth by adults in Irwin et al. (2017) might be yielded by the simplicity of tasks which loosened the demand for finer visual cues from lip movement, as a ceiling was achieved by both children and adults in their behavioural test. Combining the existing findings, we tended to suggest that, after the developmental shifts happening in infancy (Lewkowicz & Hansen-Tift, 2012), strategic visual attention allocation in processing talking faces continuously evolves in the later stages of lifespan, where children converge their fixation to the mouth area of speakers during speech perception tasks (Irwin et al., 2017; Yamamoto et al., 2019). This process appears mature in adolescence, seemingly to prepare children for acquiring sophisticated and nuanced language skills.

The impact of auditory noise on proportion fixation was limited in the current study. There were no significant differences driven by noise level regarding proportion fixation towards mouth and other facial areas onto which most of the eye gaze from participants fell. This might probably be due to the random order in stimulus presentation, and participants might, therefore, maintain their attention tightly concentrated on the mouth area of the speaker for handling adverse conditions (Król, 2018). However, a significant drop in eyes-looking time induced by auditory intelligibility falling from quiet to -10 dB SNR condition was observed, consistent with the proposal that eyes indexing social cues will pay the price for increasing attention directed toward mouth areas to achieve successful decoding speech in noisy conditions (Król, 2018).

To further estimate the specific time window for significant effect, the time-course of proportion-looking time directed to AOIs was analyzed using GAMM. Results indicated that the curves modelling temporal changes in mouth-looking time for 4–5-year-olds, 6–7-year-olds and 8–9-year-olds, instead of 10 to 11 years, was significantly lower in height relative to adults, aligning with results from ANOVA. When locating time windows of significance, it was observable that the developmental effect mainly occurred during the early-to-middle period of the 2000-ms stimulus window, corresponding with Irwin et al. (2017). During this period,

stimuli were presenting the transition from the speaker's still face to speaking with maximum mouth openness. It is noteworthy that this time window was where linguistic cues were most densely distributed for the current design, as stimuli contrasted in the consonant part whose articulation took place at the onset of a syllable. Accordingly, with their lower likelihood of looking at the mouth area, younger children aged 4 to 9 years might miss out on the key linguistic cues for speech decoding probably underlying their underperformance in audiovisual speech perception. Meanwhile, the scale of the time window of significance was witnessed to shrink progressively until it was eventually eliminated at 10 years of age, signalling the direction of development was to concentrate on the mouth area sooner during speech perception task. There are two potential consequences that trigger this progression. First, elder participants might direct their gaze on the speaker's mouth quicker as they were aware that the mouth area was the richest in finer visual speech cues (Skipper et al., 2007). Second, with the ability to segment syllables sharpening with age (Ho & Bryant, 1997), experienced participants might notice that stimuli only differed in phoneme onset, and therefore, they might consciously attend more to the speaker's mouth earlier, which could be illustrated by the left skewness of the curve for the mouth-looking time in adults.

Considering facial areas other than eyes and mouth of the speaker's face also captured a large proportion of fixation, the temporal course of fixation towards this AOI was also explored. An entirely opposite pattern to the mouth area was disclosed, given that the curve for younger children aged 4 to 9 years was higher than that for adults. When estimating when the difference emerged, time windows identified during at a similar period, namely, the first half of a stimulus, were observed, indicating that younger children might be focusing facial components sparse in linguistic or social cues (e.g., noise, cheek). Likewise, the duration of this window shortened with development and disappear by 10 years of age, implying to a more strategic face-processing fashion alongside the maturation of visual attention and linguistic skills.

4.4.3 Synchronous Developmental Courses Shared by Audiovisual Speech Perception and Face Processing

With the McGurk paradigm in various auditory conditions, we observed synchronous developmental trajectories in behavioural response and talking face processing. Specifically, most of the statistically significant group differences were found between adults and younger participants aged 4 to 9 years rather than 10 to 11 years. Our findings supported the link between language development and face processing with data from Cantonese-speaking preschool and school-aged children (Pascalis et al., 2014). To the best of our knowledge, this study may, for the first time, provide behavioural evidence corresponding with eye-tracking findings. Though this association has been apparent in the pioneering works by Irwin et al. (2017) and Yamamoto et al. (2019), the previous behavioural results might still require further validation due to the insufficient task difficulty or the limited developmental change among participants. Introducing noise to the auditory modality, enlarging the sample size, and including younger child participants enabled us to better elucidate a clear, interconnected, and synchronized developmental process involving audiovisual speech perception and talking-face processing. This synchronization is highly likely driven by the shared neural underpinning between audiovisual speech perception and face processing, especially mouth processing, as indicated by the fMRI study employing the McGurk paradigm by Rennig and Beauchamp (2018) where the anterior subregion of the posterior STS (pSTS) was found to not only prefer trials in which participants fixated the mouth but also responded more strongly to audiovisual speech. Following this line, the maturation of the anterior pSTS may facilitate the reciprocal development of audiovisual speech perception and talking-face processing at the same pace, rendering a fuller picture of multisensory development.

4.5 Conclusion

Behavioural responses and eye movements from HK Cantonese-speaking 4–5-year-old, 6–7-year-old, 8–9-year-old and 10–11-year-old children as well as adults were documented to

portray the development trajectories for audiovisual speech perception and talking face processing. According to behavioural results, the ability to identify congruent stimuli in Cantonese-speaking children developed along with age and matured around ten years of age. As for the perception of incongruent stimuli, a clear but later developmental shift was observed to occur around ten years of age, as children aged 4 to 9 years, instead of 10–11-year-olds, made significantly more audio-dominant and fewer audiovisual-integrated responses compared to adults, supporting the notion that the tonal properties of the native background do not eliminate the developmental shift. While auditory noise postponed the development of the ability to identify congruent stimuli, it led children to perceive incongruent stimuli in a manner more similar to adults at earlier ages. Regression results confirmed that the direction of this developmental process was aligned with the prediction from the statistically optimal hypothesis. Eye-tracking data recorded during audiovisual speech perception revealed fewer fixations on the mouth area but more to other facial areas by younger children aged 4 to 9 years relative to both 10–11-year-old and adult groups. Time-course analysis of fixation revealed that these developmental effects mainly took place during the first half of the stimulus, where linguistic information was densely distributed, indicating the development of talking face processing might be driven by linguistic experience. A synchronous development shared by audiovisual speech perception and talking face processing was observed, reaching the adult level at around the age of ten.

Chapter 5. Deficient Attention allocation towards Human Faces Hampers Audiovisual Speech Perception in Children with Autism Spectrum Disorder

5.1 Introduction

Autism Spectrum Disorder (ASD) has been characterized by persistent challenges with social communication and interaction, as well as restricted and repetitive behaviours, interests, or activities. According to *The Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (DSM-5, American Psychiatric Association, 2013), sensory disturbance remains one of the vital diagnostic criteria for ASD. Difficulties in integrating multisensory cues among autistic individuals remain one of the significant aspects of sensory disturbance, as they struggle, or even are unable, to combine information from multiple sensory organs to form a holistic and coherent perception of their surrounding environment (Stevenson et al., 2014a; Rong et al., 2023). Since sensory integration is considered to provide a solid foundation for the development of higher-order cognitive activities (Marco et al., 2011), such integration difficulties may hamper cognitive development in autistic individuals and place them at social and communicative disadvantages. The differences between autistic and neurotypical individuals appear even more pronounced when processing social stimuli, as in the McGurk paradigm, which involves two vital dimensions of interpersonal interaction: speech decoding and talking face processing (Klin et al., 2002; Stevenson et al., 2014a). However, little is known about the extent to which language ability and face-processing manner might account for the atypical audiovisual speech perception in autistic individuals.

5.1.1 The Magnitude of the McGurk Illusion in Children with ASD

Previous studies on audiovisual speech perception in autistic individuals showed considerable inconsistency. On the one hand, there are substantial findings suggesting that autistic individuals adopt atypical perceptual strategies in perceiving audiovisual stimuli as they were more likely to make audio-dominant responses in the McGurk paradigm (DePape et al., 2012;

Feldman et al., 2022; Feng et al., 2021a, 2021b; Iarocci et al., 2010; Irwin et al., 2011; Stevenson et al., 2014a, 2014b, 2016), which was supported by recent systematic reviews and meta-analysis (Feldman et al., 2018; Zhang et al., 2019). For instance, a reduced degree of audiovisual speech perception with enhanced audio-dominant responses was observed among 26 preschool-aged autistic children aged from 4 to 7 years (mean age = 6.07) compared with their TD peers matched on age, full-scale intelligence quotient (IQ) and verbal IQ when the speaker's eyes were open in Feng et al., (2021a). Consistently, in another study, Feng et al. (2021b) also found a weaker McGurk effect at an overall level in the autistic group in a similar age range. When examining autistic individuals with a broader age range, Irwin et al. (2011) included 13 autistic children aged from 5 to 15 years (mean age = 9.08) and 13 age-matched TD for comparison and consistently obtained significantly fewer visually influenced responses from the autistic group in the audiovisual incongruent condition, in parallel with their reduced visual gain in the speech-in-noise perception task. On the one hand, it cannot be ignored that several studies also measured a comparable performance in experiencing the McGurk illusion between the autistic population with their TD counterparts, suggesting that audiovisual integration may remain intact in some autistic individuals (Keane et al., 2010; Saalasti et al., 2012; Woynaroski et al., 2013).

Taken together, while autistic individuals were frequently reported to show a perceptual preference for audio-dominant responses, even after controlling for age and IQ, there exists variability in the extent of audiovisual integration indexed by the McGurk illusion across the literature. To address these conflicting findings, some studies have explored potential explanations by examining the heterogeneity within the autism spectrum population. Feldman et al. (2018) reviewed existing research and identified correlations between the strength of audiovisual integration, language/communication abilities, and/or autism symptom severity. In a subsequent empirical study, Feldman et al. (2022) found that the strength of the McGurk illusion positively predicted social communication skills and negatively correlated with autism

features in a sample of 18 autistic children (mean age = 12 years and 7 months). These findings suggest that between-group discrepancies observed in previous studies may be linked to uncontrolled language abilities.

5.1.2 Audiovisual Speech Perception in Noise among Children with ASD

Noise is a common factor affecting audiovisual speech perception in everyday life, as our environment is inherently noisy. In auditory noisy circumstances, visual information often serves as a crucial compensatory source, especially for face-to-face communication. Existing studies have reported an elevation in visual reliance among perceivers when auditory noise is introduced to stimuli, regardless of their language background (Hazan et al., 2010). However, if autistic individuals have difficulty processing social stimuli such as the human face (Kanner, 1943; Klin et al., 2002; Madipakkam et al., 2017), they are unlikely to effectively utilize visual information to assist with speech processing. As a result, they may benefit less from the facilitative effect of visual compensation on speech comprehension (Irwin et al., 2011). For the identification of congruent stimuli, autistic children have been observed to underperform their TD peers in perceiving various linguistic components, including phonemes (Stevenson et al., 2017, 2018), syllables (Irwin et al., 2011), words (Foxe et al., 2015; Smith & Bennetto, 2007; Stevenson et al., 2017). This weaker audiovisual integration seemed to stem from insufficient visual intake and/or utilization. Irwin et al. (2011) found that autistic children only underperform their TD counterparts when perceiving audiovisual speech stimuli in auditory noise (-10 dB, -15dB, and -20dB SNRs), rather than in unimodal auditory context, suggesting that autistic individuals benefit less from visual assistance. Findings concerning the perception of incongruent stimuli in noise also support that autistic children may be less influenced by visual cues. For instance, Liu et al. (2020) measured the McGurk effect in Mandarin-speaking autistic children under various auditory conditions, observing an enhanced McGurk effect as auditory noise increased from no noise to -6dB SNR, alongside a reduced McGurk effect in the autistic group. Notably, the diagnostic group × auditory condition two-way interaction did not

reach statistical significance, implying that autistic remained weaker in integrating audiovisual information in noisy conditions, despite increased visual utilization.

In summary, the disadvantage experienced by autistic children in the noise-free condition seems to persist under noisy conditions. This disparity appears to stem from their distinctive manner of visual speech processing, particularly when it comes to processing talking faces.

5.1.3 Face-viewing Pattern in Children with ASD

Given that the human face is the most important source of visual information during audiovisual speech processing (Meltzoff & Kuhl, 1994), how one processes facial cues may explicitly influence the output of audiovisual speech perception. A tendency to avoid looking at human faces has been observed among autistic individuals (Baranek, 1999; Grossman, 2015; Osterling & Dawson, 1994), potentially reducing their intake of visual speech cues. For example, when performing the McGurk paradigm, Irwin et al. (2011) found that autistic children were less likely to fixate on the speaker's face. This reduced focus on facial features may stem from the "eye avoidance" behaviour often noted in autistic individuals (Kirchner et al., 2011). Using a dynamic social scene viewing task, Klin et al. (2002) found that autistic individuals fixated significantly more on the human body and objects rather than the face regions of the speakers. Notably, the suppressed face-directed fixation seemed to result from a marked decrease in eyes-directed fixation, as autistic individuals spent twice as much time focusing on the speakers' mouths. Crucially, it was mouth-looking time that positively predicted social adjustment and verbal ability of autistic participants, while it was negatively associated with autistic social impairment. Accordingly, the abnormal processing of faces and facial features in autistic individuals might sacrifice the visual intake for audiovisual speech perception from facial movement, which conveys densely and finely represented linguistic cues.

In addition to the whole face, the frequency of looking at the mouth region has also been closely linked to audiovisual speech perception and language decoding (Feng et al., 2021b,

2022; Gurler et al., 2015). However, previous studies comparing the processing of internal facial components between autistic and neurotypical individuals during the McGurk task have yielded inconsistent findings. For example, using the McGurk paradigm, Saalasti et al. (2012) found that adults with Asperger's exhibited a face-viewing pattern similar to that of their neurotypical counterparts, with both groups primarily fixating on the mouth area and the least on the eyes. Similar findings were reported by Yi et al (2013). In contrast, Feng et al. (2021) found that autistic participants were less likely to fixate on the mouth when the speaker's eyes were open, but their fixation patterns were comparable to their TD peers when the speaker's eyes were closed. With the time-course analysis of mouth-looking time, the "eye avoidance" manner at the expense of attention to the mouth area during audiovisual speech perception tasks was further confirmed. In a more recent study, Feng et al. (2022) observed significantly higher McGurk responses—indicative of better audiovisual speech integration—when the visual attention of autistic children was directed to the mouth, echoing Klin et al. (2002) and underscoring the importance of mouth-directed attention for absorbing visual speech information.

Collectively, an atypical face-processing manner in the autistic population has been disclosed in previous studies, with reduced attention directed towards speakers' faces in audiovisual speech processing. On top of that, whether there are abnormalities in how autistic individuals view the internal regions of faces remains controversial.

5.1.4 The Current Study

Based on existing studies, whether the underlying mechanism of audiovisual integration, as indexed by the strength of the McGurk effect, is deficient in autistic individuals remains controversial. Previous studies have pointed out that the frequency of experiencing the McGurk effect tends to be associated with the verbal abilities and skills of autistic individuals, possibly accounting for the mixed results on this issue (Feldman et al., 2018, 2022). Besides, the atypical

face processing pattern observed among autistic populations, including reduced attention directed to the human face and an abnormal face scanning manner, might also contribute to their disadvantages in audiovisual speech perception (Feng et al., 2021b, 2022; Irwin et al., 2011; Klin et al., 2002). Therefore, the current study sought to investigate the role of language ability in audiovisual speech perception in autistic children by including a group of TD children whose scores in language tests matched the autistic group on a one-by-one basis (language-ability-matched TD group, LA-matched TD group) in addition to a chronological-age-matched TD (CA-matched TD) group. Also, the face processing pattern was examined through the simultaneous recording of eye movements from the participants, with the likelihood of viewing the speaker's face and the mouth area of special interest. Through the current design, we aimed to explore the following research questions: 1) whether autistic children behave differently in making responses to the McGurk paradigm compared to their chronological-age-matched TD groups, and 2) to what degree the abnormal behavioural responses can be explained by the language ability of autistic children through the comparison between their language-ability-matched TD group, 3) whether the autistic group shows atypical visual attention allocation to the speaker's face and the speaker's mouth area compared to the other two TD groups, and 4) to what extent the atypicalities manifested in audiovisual speech perception can be accounted for by the abnormal visual attention allocation in autistic children?

5.2 Methods

5.2.1 Participants

Sixty-six Cantonese-speaking children were initially recruited for the current study, including 24 diagnosed with ASD (mean age = 9.41 years, *SD* = 1.02 years), and 42 were TD children. Children matched with autistic groups in terms of chronological age (n = 20) or language ability (n = 20) were allocated to two distinct control groups, giving rise to CA-matched TD and LA-matched TD groups. Language ability was measured using the Test of Cantonese Grammar, a

subtest of Hong Kong Cantonese Oral Language Assessment Scale (HKCOLAS, T'sou et al., 2006), a widely used standardized test of oral Cantonese, whose norms were based on a representative sample of 1,120 Hong Kong children aged between 4 years 10 months and 12 years 1 month (Klee et al., 2008). The subtest measured the grammar knowledge of children from both receptive and expressive perspectives, with 83 items in total. The raw scores of the subtest, instead of age-equated scores, were adopted in the current study when matching the autistic group with their language-ability-matched TD peers on a one-to-one basis, considering that the norms established in 2006 might not fully align with the current situation.

All the participants were recruited through the departmental website and the official social media account. They were born in HK, had Cantonese as their first language, and attended mainstream schools. Children from the autistic group were diagnosed by qualified paediatricians or child psychiatrists in local hospitals or Child Assessment Centres managed by HK Government based on the fourth or the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, 2000, 2013). According to the diagnostic certificates provided for children in the autistic group, their intelligence had been assessed using the Fourth Edition of Wechsler Intelligence Scale for Children (WISC-IV-HK, Wechsler, 2010) or Wechsler Preschool and Primary Scale of Intelligence (WPPSI-IV-HK, Wechsler, 2012). Two autistic children were excluded due to intelligence disability (intelligence quotient under 70). Another two child autistic participants were excluded due to aversion to auditory noise (see *Stimulus*). TD children reported no physical, cognitive or hearing impairment, and their behavioural traits were further confirmed using the Autism Spectrum Quotient—Children's Version (AQ-Child, traditional Chinese version, Auyeung et al., 2008). Two children were excluded from LA-matched TD group with one scoring higher than cut-off in AQ-Child, and one suspected for language delay indicated by HKCOLAS. All participants had normal or corrected-to-normal vision. Written consent was obtained from their

caregivers together with verbal consent from child participants. Participants and caregivers would receive monetary compensation and a small gift upon completion.

Table 5.1 Descriptive characteristics of autistic and non-autistic children.

Group	*N* (Female/Male)	Chronological Ages (Range, in year)		Language Score (Range)		AQ-Child (Range, in month)	
		Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*
Autistic Group	20 (5/15)	9.41 (8.10–11.27)	1.02	7.30 (40–81)	1.95	-	-
Chronological-age-matched TD group	20 (7/13)	9.69 (8.36–11.97)	.93	76.1 (70–82)	4.78	55.60 (20–74)	17.76
Language-ability-matched TD group	20 (6/14)	8.21 (5.11–11.38)	1.83	7.35 (41–80)	1.76	63.65 (38–75)	9.86

5.2.2 Stimulus & Procedure

Stimuli and procedure adopted in the current chapter were identical to those used in Chapter 4 (see Section 4.2.2 & 4.2.3 for reference).

5.2.3 Data Analysis

5.2.3.1 Behavioural data

Similar to Chapter 4, given that the behavioural data were not normally distributed, permutation-based ANOVAs were adopted for analysis using the package "*permuco*" in R. For the identification of congruent stimuli, a 3 × 3 × 3 ANOVA was conducted with Group (autistic, chronological-age-matched TD, and language-ability-matched TD) as the between-group factors, and Stimulus Type ("Ba," "Da," and "Ga") and Noise Level (quiet, 10 dB SNR, and -10 dB SNR) as within-group factors to examine how identification accuracy was affected. For

the perception of incongruent stimuli, another 3 × 3 × 3 ANOVA was performed with Group (autistic, chronological-age-matched TD, and language-ability-matched TD) as between-group factors, and Response Type (audio-dominant, audiovisual-integrated, and visual dominant) and Noise Level (quiet, 10 dB SNR and -10 dB SNR) as within-group factors to explore how these factors impacted the percentage of different responses. Post-hoc pairwise comparisons using Wilcoxon tests with Bonferroni correction were applied where appropriate.

5.2.3.2 Eye-tracking Data

The obtained raw eye gaze data were processed in R, and only the data from the 2000-ms stimulus window were extracted for processing. Data validity was checked for participants and trials, with no participants excluded based on a 75% threshold (Grandon et al., 2023). Subsequently, we calculated the number of eye fixations falling into three fixed pre-defined areas-of-interest (AOIs): mouth, eyes and other facial areas, and converted them to looking time (Franco-Watkins & Johnson, 2011). Proportion-looking time was obtained by dividing the duration of looking at the corresponding AOI by the total looking time directed at the whole face area (Feng et al., 2021b). The Face-looking Ratio was calculated by dividing the total fixation duration directed towards the face area (i.e., all three AOIs) by the total duration spent looking at the screen.

Two key variables associated with visual attention allocation to the human face were particularly examined: Face-directed fixation duration and Mouth-looking time. Face-directed fixation duration, calculated by summing all the fixations directed towards the mouth area of the speaker, was fitted into a linear mixed model (LMM) where the fixed effects were Group (autistic, CA-matched TD, and LA-matched TD), Noise Level (quiet, 10 dB SNR, and -10 dB SNR) and Congruency (congruent vs. incongruent) of the stimuli while random effect was Participant. Logit-transformed Mouth-looking time, namely the proportion-looking time directed at the mouth area, was also fitted into another LMM model to examine how it was impacted by Group, Noise Level, and the Congruency of stimuli. The initial model included

all main and interaction effects of the variables, and model comparisons were conducted using LRTs. The best-fitting models were determined by removing all the terms that did not significantly improve model goodness. Post-hoc pairwise comparisons with Bonferroni correction were carried out based on the best-fitting model using the "*emmeans*" package in R.

Given that mouth-looking time and face-looking ratio were likely to impact the strength of audiovisual integration in the McGurk design, mixed-effects models were employed for further investigation using the "*lmer4*" package in R. Specifically, GLMMs with a logistic link function were constructed with Mouth-looking Time, Face-looking Ratio, Group, Noise Level and their interaction, excluding the interaction between Mouth-looking Time and Face-looking Ratio, as fixed effects and the random intercept of Participant (formula: Response Type – Group * Mouth-looking Time * Noise Level + Group * Face-looking Ratio * Noise Level + (1| Participant)) to examine their impact on a certain type of response to the incongruent stimuli under varying auditory conditions. Model construction and comparison processes were similar to those for LMMs.

To further analyze how fixation towards the speaker's face and mouth area changed over time, the empirical logit of eye fixations was assigned to 50-ms time bins and analyzed using GAMMs in R. GAMMs better handle the autocorrelation of eye-tracking data, apply smoothing to capture underlying trends, and control for within-group variability. First, a model was constructed with the empirical logit of fixation on the speaker's whole face treated as the dependent variable (Barr, 2008). Second, another GAMM modelling the empirical logit of fixation on the speaker's mouth area was also built. Both models followed a forward stepwise process, which began with the simplest model containing the smooth term of Time. Subsequently, random smooths for the Item within Group and for Participant were added to the model step-by-step, followed by the fixed-effect factor of Group (Grandon et al., 2023). Model comparisons based on AIC differences were carried out to confirm whether the added smooth

term significantly improved the goodness of fit. If so, the more complex model was kept over the simpler model, or the simpler one would be retained by excluding the added factor. Furthermore, we utilized the *plot_diff* function from the "*itsadug*" package to detect the time window where significant effects occurred.

5.3 Results

5.3.1 Behavioural Results

5.3.1.1 The identification of congruent stimuli

Figure 5.1 displays the identification accuracy achieved by three groups of participants under three auditory conditions. Permutation-based repeated measures ANOVA detected the significant main effects of Group ($F(2,57) = 3.93$, permutation $p = .02$, $\eta_p^2 = .01$), Stimulus Type ($F(2,114) = 17.90$, permutation $p < .001$, $\eta_p^2 = .05$), Noise Level ($F(2,114) = 265.68$, permutation $p < .001$, $\eta_p^2 = .79$), and a Stimulus Type × Noise Level interaction ($F(4,228) = 16.03$, permutation $p < .001$, $\eta_p^2 = .09$). The main effect of Group was mainly driven by the comparatively lower accuracy attained by the autistic group ($M = .84$, $SE = .002$) relative to their CA-matched ($M = .88$, $SE = .001$) and LA-matched ($M = .88$, $SE = .001$) TD counterparts. For the Stimulus Type × Noise Level interaction, participants identified the three congruent syllables with a comparable accuracy under quiet and 10 dB SNR conditions. When the auditory intelligence dropped to -10 dB SNR, participants achieved the highest accuracy in identifying "Ba", followed by "Ga", and the lowest in "Da" (all ps $< .05$).

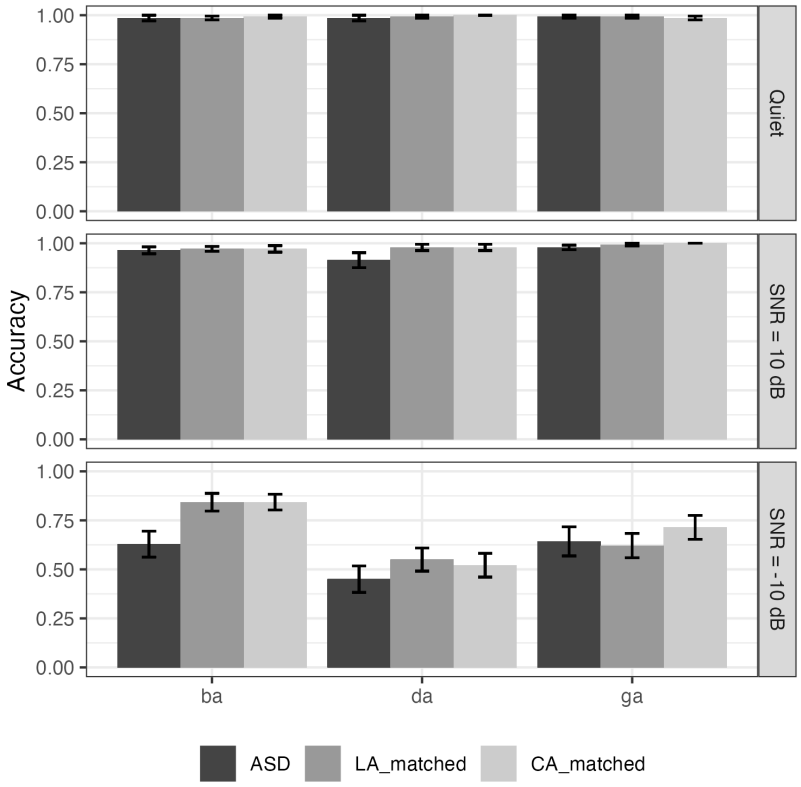


Figure 5.1 The identification accuracy achieved by three groups of child participants under three auditory conditions.

5.3.1.2 The perception of incongruent stimuli

Figure 5.2 presents the percentage of responses to the incongruent stimuli by three groups of participants across auditory conditions. Repeated measures permutation ANOVA reported the significant Group × Response Type × Noise Level three-way interaction ($F(8,228) = 4.23$, permutation $p < .001$, $\eta_p^2 = .06$), which was further analyzed based on Noise Level. In quiet and 10 dB SNR conditions, the autistic group was found to make significantly more audio-dominant responses while fewer audiovisual-integrated responses relative to their CA-matched TD counterparts (all ps $< .05$), while such differences were insignificant between the autistic and LA-matched TD groups. No between-group differences reached significance regarding visual-dominant responses. In the -10 dB SNR, however, no significant differences were observed among groups across responses (all ps $> .05$).

When it comes to sensory dominance in percept generation, which could be indexed by the response type to the incongruent stimuli, three groups of participants exhibited distinctive patterns. In the quiet condition, the number of audio-dominant responses significantly exceeded that of audiovisual-integrated ones in the autistic group (p = .04), whereas, in stark contrast, it was lower than that of audiovisual-integrated ones in the CA-matched TD group (p = .048). While in the LA-matched group, the number of these two types of responses were comparable (p = 1.00). In the 10 dB SNR condition, both the autistic and LA-matched TD groups made a comparable number of audio-dominant and audiovisual-integrated responses (both ps > .05), whilst the CA-matched TD group still made the most audiovisual-integrated responses (both ps < .05). In the -10 dB SNR condition, all groups of participants made a comparable number of audiovisual-integrated and visual-dominant responses, which were significantly higher than that of audio-dominant responses.

When the auditory intelligibility dropped from quiet to 10 dB SNR, only LA-matched group observed a significant descent in audio-dominant responses (p = .02). With the auditory condition dropping from quiet to -10 dB SNR, there witnessed a significant decrease in audio-dominant responses along with the significant increase in visual-dominant responses regardless of groups (all ps < .05). Besides, the percentage of audiovisual-integrated responses significantly shrunk in the CA-matched group (p = .003).

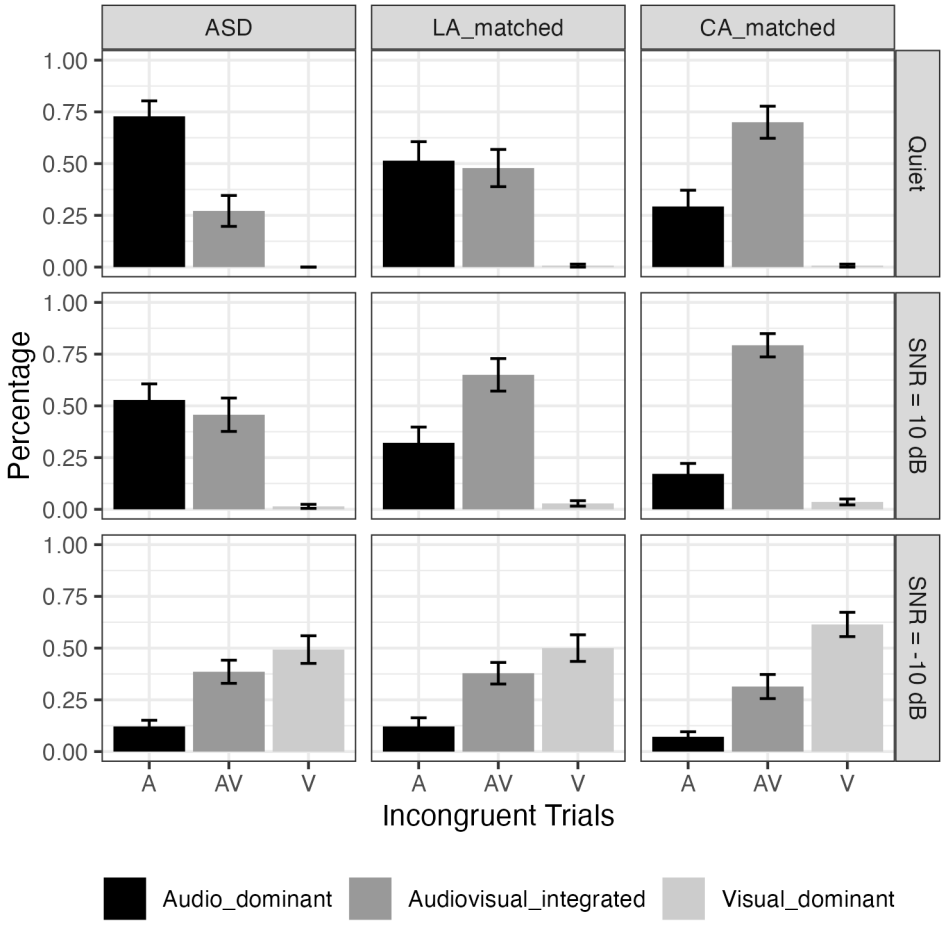


Figure 5.2 Percentage of responses to the incongruent stimuli by three groups of participants across auditory conditions.

5.3.2 Eye-tracking Results

5.3.2.1 Face-directed fixation duration and Mouth-looking time

The LMM investigating face-directed fixation duration revealed the significant main effect of Group (F= 5.30, p = .01) and Noise Level (F= 3.25, p = .04). Post-hoc pairwise comparisons uncovered that the autistic group showed a significantly shorter face-directed fixation duration relative to both TD groups (both $p < .05$). For the main effect of Noise Level which was likely induced by the lengthened face-directed fixation with lower SNR (Quiet: M = 1078.40, SE = 12.95; 10 dB: M = 1092.88, SE = 8.79; -10 dB: M = 1109.00, SE = 9.12), post-hoc analysis failed to detect any differences with pairwise comparisons. For the LMM examining Mouth-looking Time, no main effects or interactions reached the significance level (all $p > .05$).

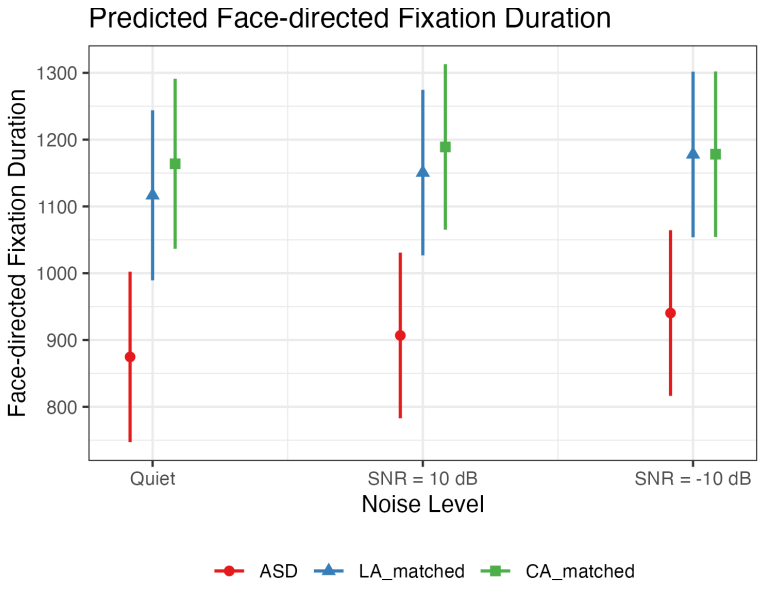


Figure 5.3 The predicted face-directed duration for autistic, LA-matched TD and CA-matched-TD groups in processing audiovisual speech stimuli.

5.3.2.2 The perception of incongruent stimuli

5.3.2.2.1 Audio-dominant responses

The GLMM constructed on the probability of making audio-dominant responses revealed that the main effects of Face-looking Ratio ($\chi^2(1) = 15.84$, $p < .01$), Group ($\chi^2(2) = 10.35$, $p = .01$) and Noise Level ($\chi^2 = 122.45$, $p < .01$), together with the Mouth-looking Time × Noise Level ($\chi^2(2) = 7.57$, $p = .02$) and Group × Noise Level ($\chi^2(4) = 14.05$, $p = .01$) interactions significantly improved model goodness. Face-looking Ratio was found to negatively predict the likelihood of audio-dominant responses with a moderate effect size ($OR = .23$). The significant Mouth-looking Time × Noise Level interaction was induced by negative predictability of Mouth-looking Time was only revealed in the quiet ($p = .03$, $OR = .32$) and 10 dB SNR ($p = .002$, $OR = .25$) conditions with small-to-medium effect sizes but not in the -10 dB SNR condition ($p = .84$). Results of post-hoc pairwise comparison regarding the Group × Noise Level interaction were consistent with the ANOVA, as the likelihood of making an audio-dominant response in the autistic group appeared to be higher than in the chronological-

age-matched TD groups under quiet and 10 dB SNR conditions (all *p*s < .05), except for the -10 dB SNR one (all *p*s > .05).

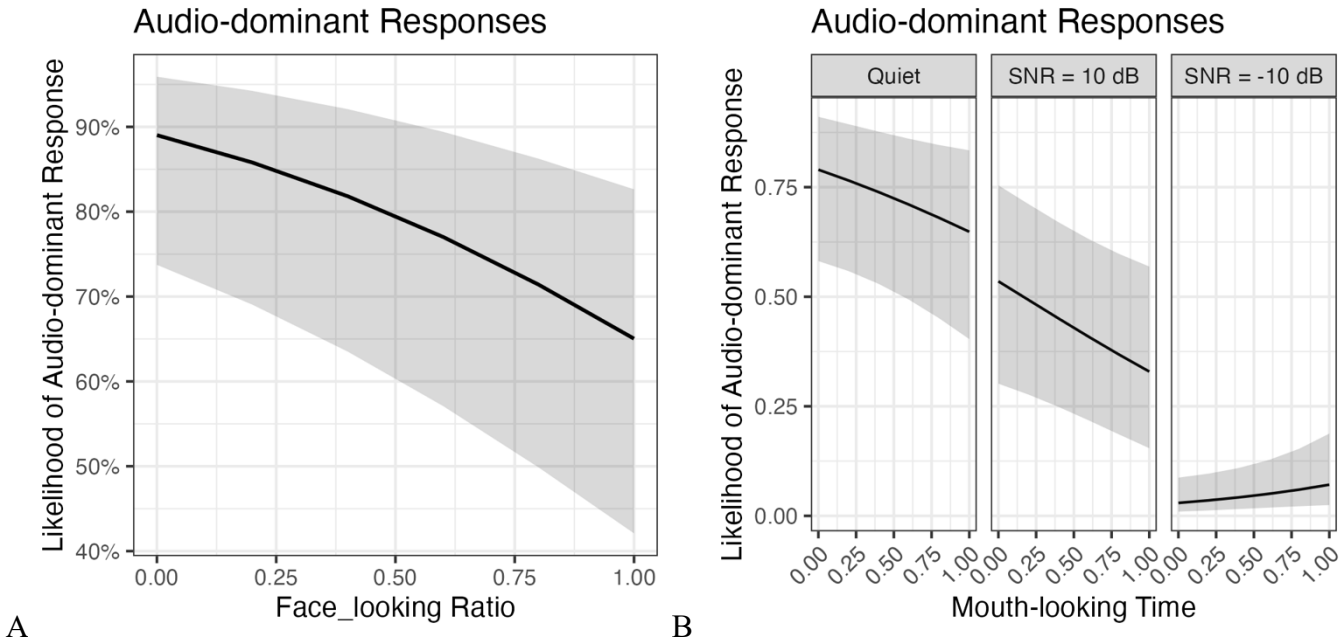


Figure 5.4 (A) The estimated main effect of Face-looking ratio, and (B) the estimated interaction effect between Mouth-looking time and Noise level on the likelihood of making audio-dominant responses by GLMM.

5.3.2.2.2 Audiovisual-integrated responses

As for the model for the likelihood of making audiovisual-integrated responses, the main effects of Face-looking Ratio (χ^2 (1) = 11.09, p < .01) and Noise Level (χ^2(2) = 64.71, p < .01), the Noise Level × Mouth-looking Time (χ^2(2) = 9.59, p = .01) and Group × Noise Level (χ^2(4) = 44.72, p < .01) two-way interactions, and the Mouth-looking Time × Group × Noise Level three-way interaction (χ^2(2) = 11.29, p = .02) significantly enhanced the model fit. Face-looking Ratio was found to be positively associated with the likelihood of making audiovisual-integrated responses with a moderate effect (p < .01, *OR* = 2.91). To better understand the Mouth-looking Time × Group × Noise Level interaction, it was further analyzed under Noise Level. In quiet and 10 dB SNR conditions, the two-way Mouth-looking Time × Group interaction was insignificant (both *p*s > .05). Mouth-looking Time was found to positively

predict the occurrence of audiovisual-integrated responses in the 10 dB SNR condition (p = .01, OR = 4.03), which was marginally significant in the quiet condition (p = .07, OR = 3.56). Consistent with the permutation ANOVA, the autistic group was found to make fewer audiovisual-integrated responses relative to their CA-matched TD controls in both conditions (both ps < .05). In the -10 dB SNR condition, since the Mouth-looking Time × Group interaction reached marginal significance (χ^2 = 5.52, p = .06), the predictability of Mouth-looking Time was further examined within groups. This interaction was seemingly due to the significant negative predictability of Mouth-looking Time in the LA-matched TD group with a medium effect size (p = .048, OR = .33), instead of the other two groups (both ps > .05).

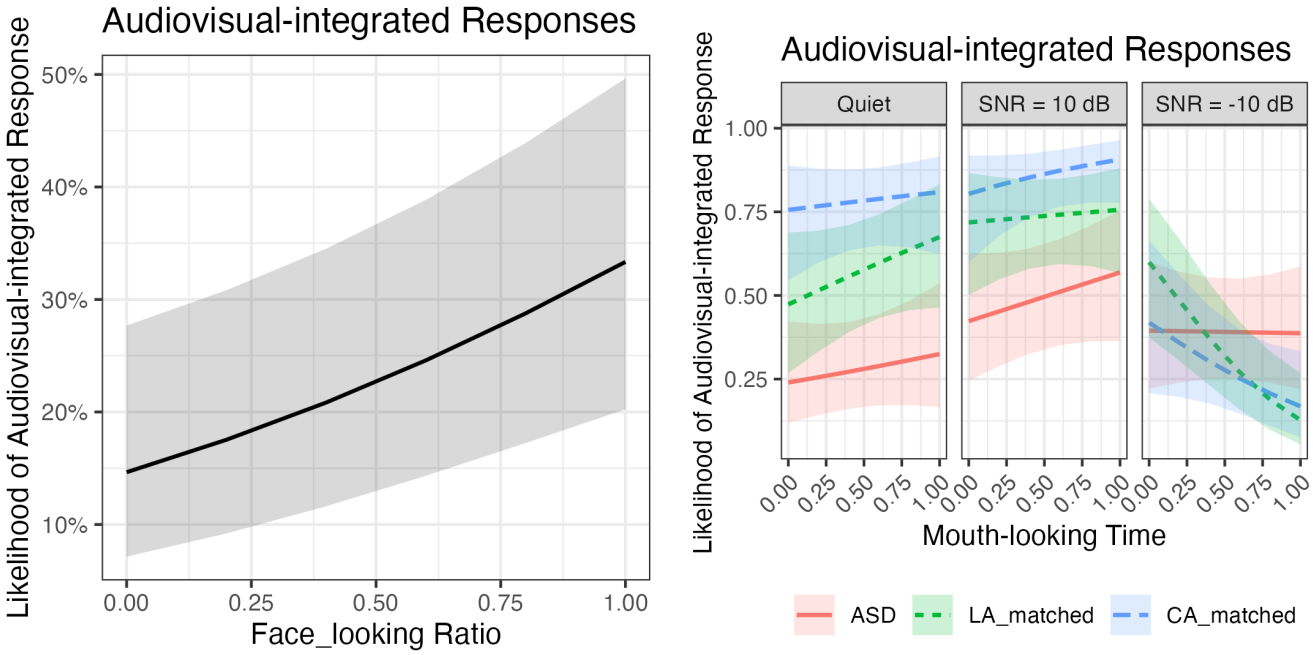


Figure 5.5 (A) The estimated main effect of Face-looking ratio, and (B) the estimated interaction effect among Group, Mouth-looking time and Noise level on the likelihood of making audiovisual-integrated responses by GLMM.

5.3.2.2.3 Visual-dominant responses

Considering the issue of data sparsity for visual-dominant responses under quiet and 10 dB SNR conditions, a GLMM model with Mouth-looking Time, Face-looking Ratio, Group, and their interaction as fixed effects was only constructed for the probability under -10 dB SNR

condition. Model comparisons detected that Group × Mouth-looking Time significantly boosted model goodness ($\chi^2(2) = 11.71, p < .01$), which was driven by the significant positive predictability of Mouth-looking Time in the LA-matched TD group ($p = .01$, $OR = 6.10$), but marginally negative in the autistic group ($p = .07$, $OR = .36$) and insignificant in the CA-matched TD group ($p = .71$).

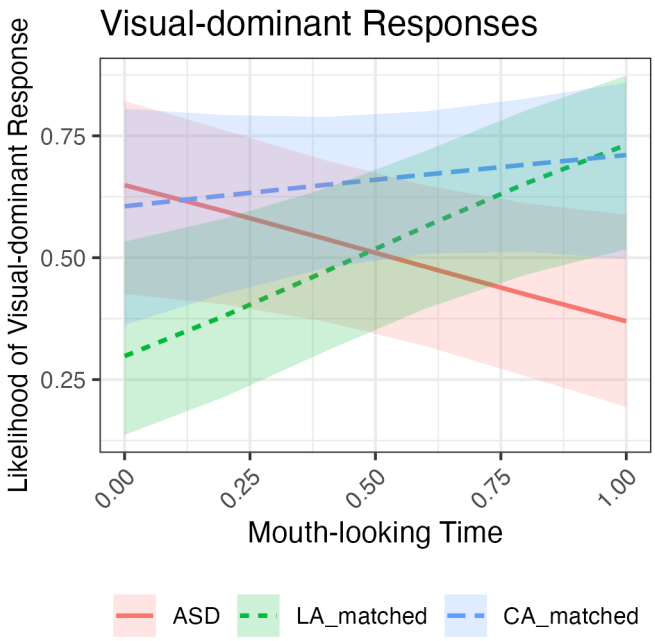


Figure 5.6 The estimated main effect of Mouth-looking time on the likelihood of making audiovisual-integrated responses in -10 dB SNR condition by GLMM.

5.3.2.4 Time-course analysis on fixation towards the mouth area

Figure 5.7 (A) exhibits the time course of fixation (empirical-logit transformed) to the face area of speakers in 50-ms time bins by three groups of participants. A GAMM was fitted to examine whether the temporal evolution of the face-looking ratio over time varied across groups, which contained smooths for the predictor of Group as well as random smooths for Participant and Item within each Group. Guided by the parametric coefficients measured by the GAMM, the curve of the face-looking ratio modelled for autistic children was found to be significantly lower in height compared with both TD groups (both $p < .05$). For the smooth terms, the fixed effect smooth for each group was highly significant (all ps $< .01$), indicating the temporal

evolution of the face-looking ratio within each group was significantly different from zero. The random effects of Participant as well as Item within each Group also reached the significance level (both *p*s < .01). When visualizing the difference curves, autistic children were revealed to show a significantly lower likelihood of fixating on the speaker's face when looking at the screen than their LA-matched TD peers during a time window covering 0 to 1818 ms of the 2000-ms stimulus. When comparing their CA-matched TD counterparts, autistic children also showed a lower probability of looking at the speaker's face during a time window situated at the early-to-middle period of the stimulus (i.e., from 222ms to 1111ms). No time window of significant differences was detected between the two TD groups.

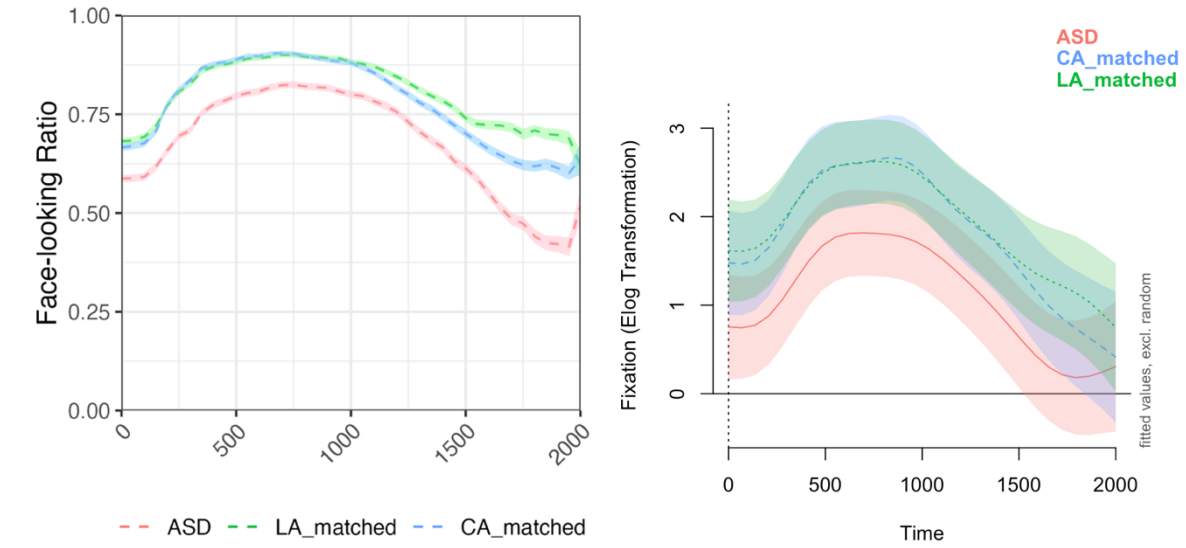


Figure 5.7 (A) The time course of the face-looking ratio (empirical-logit transformed) in 50-ms time bins by three groups of participants, and (B) the estimated temporal courses of fixation (empirical logit-transformed) towards the face areas of the speaker for three groups of participants derived from GAMM.

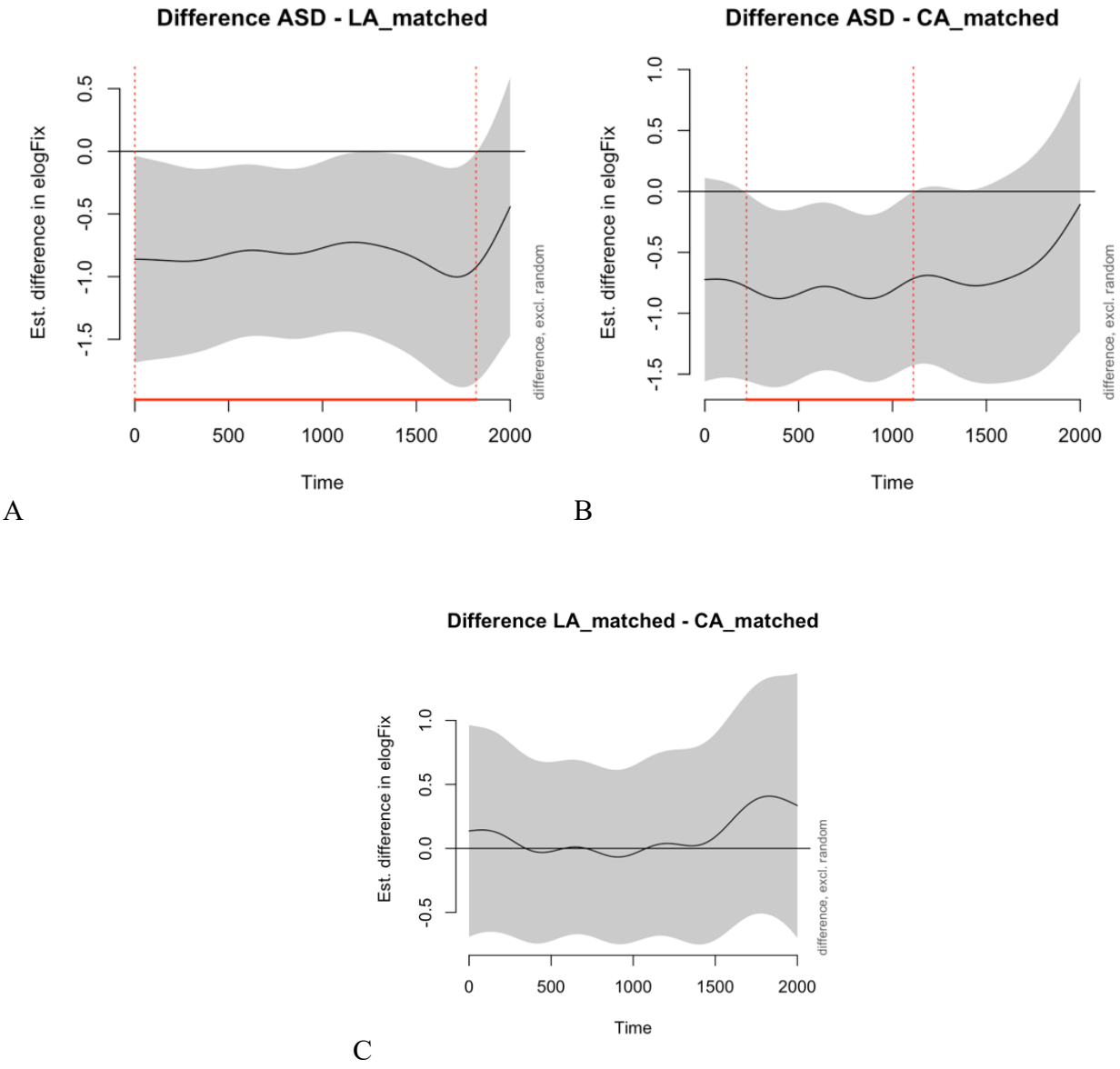


Figure 5.8 The estimated difference in the probability of fixating the speaker's face between (A) autistic and LA-matched TD groups, (B) autistic and CA-matched TD groups, and (C) LA-matched and CA-matched TD groups.

Table 5.2 Model summary for GAMM regarding fixation towards the face of the speakers.

Face-looking Ratio				
Parametric Coefficients	*Estimate*	*SE*	*t values*	*p values*
(Intercept)	1.18	0.22	5.37	0.00
Group (CA-matched TD)	0.76	0.31	2.45	0.01
Group (LA-matched TD)	0.83	0.31	2.67	0.01
Approximate significance of smooth terms:	*Edf*	*Ref.df*	*F*	*p values*
s(Time) × Group (ASD)	7.27	7.56	8.19	<.01
s(Time) × Group (CA-matched TD)	7.66	7.89	10.29	<.01
s(Time) × Group (LA-matched TD)	7.14	7.44	7.31	<.01
s(Time,Subject)	457.44	597.00	10.83	<.01
s(Time,Item) × Group (ASD)	799.64	1409.00	1.80	<.01
s(Time,Item) × Group (CA-matched TD)	758.56	1429.00	1.51	<.01
s(Time,Item) × Group (LA-matched TD)	691.11	1409.00	1.26	<.01

Figure 5.9 (A) presents the time course of fixation (transformed into empirical logit) towards the mouth area of the speaker by participants with and without ASD. Similarly, a GAMM containing fixed effect smooths for each group as well as random smooths for Participant and for Item within each Group was fitted to investigate the temporal evolution of mouth-looking time across groups. According to the parametric coefficients, the average height of the curves for both TD groups did not differ from the reference level: the autistic group (both *p*s > .05). Guided by the significance of the smooth terms, the probability of fixation directed at the speaker's mouth area changed significantly over time for all three groups of children (all *p*s < .05), indicating the probability of viewing the speaker's mouth area changed significantly over time within each group. However, when visualizing the difference waves,

the comparisons between any two groups showed no significant differences in any time window, suggesting that the patterns of the time course of mouth-looking time did not significantly differ among groups.

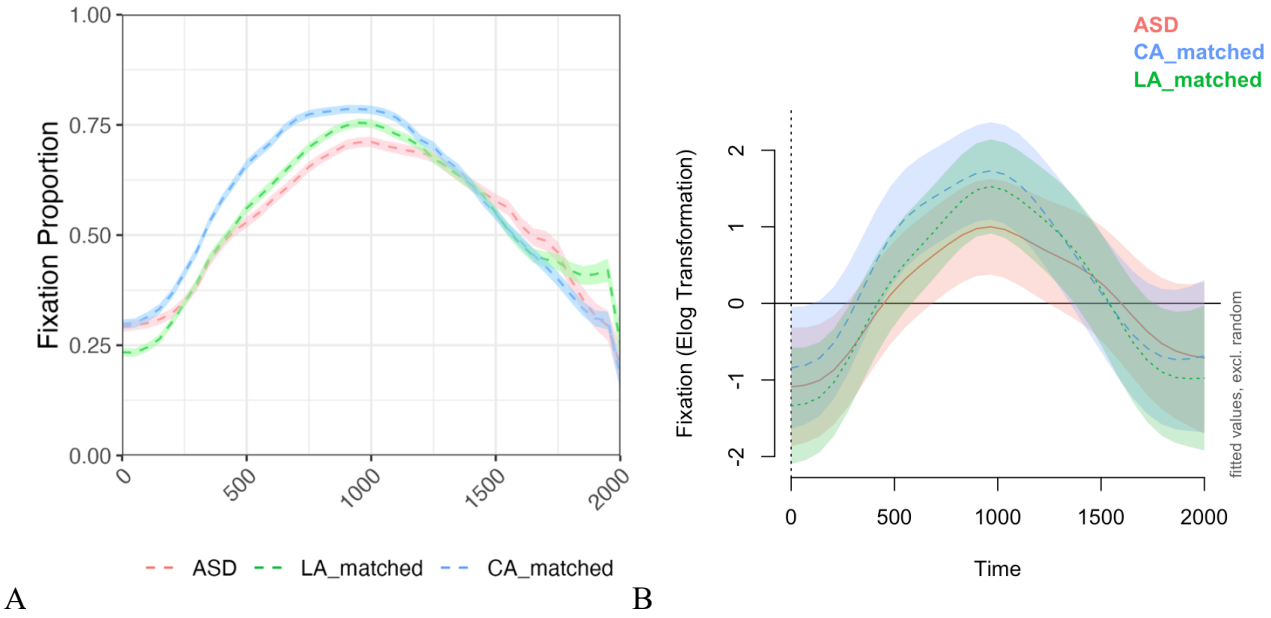


Figure 5.8 (A) The time course of mouth-looking time (empirical-logit transformed) in 50-ms time bins by three groups of participants, and (B) the estimated temporal courses of fixation (empirical logit-transformed) towards the mouth areas of the speaker for three groups of participants derived from GAMM.

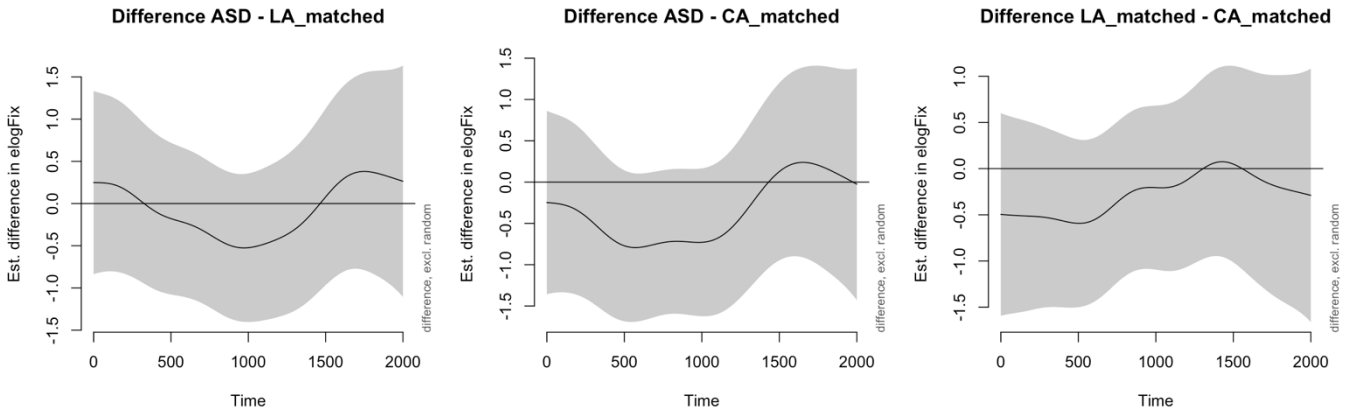


Figure 5.9 The estimated difference in the probability of fixating the speaker's mouth area between (A) autistic and LA-matched TD groups, (B) autistic and CA-matched TD groups, and (C) LA-matched and CA-matched TD groups.

Table 5.3 Model summary for GAMM regarding fixation towards the mouth area of the speaker.

Mouth-looking Time				
Parametric Coefficients	*Estimate*	*SE*	*t values*	*p values*
(Intercept)	0.15	0.28	0.53	0.60
Group (CA-matched TD)	0.45	0.40	1.12	0.26
Group (LA-matched TD)	0.13	0.40	0.34	0.74
Approximate significance of smooth terms:	*Edf*	*Ref.df*	*F*	*p values*
s(Time) × Group (ASD)	6.69	7.03	7.77	<.01
s(Time) × Group (CA-matched TD)	7.49	7.72	12.79	<.01
s(Time) × Group (LA-matched TD)	7.63	7.86	15.49	<.01
s(Time,Subject)	466.00	597.00	11.26	<.01
s(Time,Item) × Group (ASD)	822.62	1409.00	2.20	<.01
s(Time,Item) × Group (CA-matched TD)	947.71	1429.00	2.94	<.01
s(Time,Item) × Group (LA-matched TD)	863.44	1399.00	2.39	<.01

5.4 Discussion

The current study compared the behavioural responses and eye movements during audiovisual speech perception tasks using the classic McGurk paradigm among autistic children and their LA-matched and CA-matched TD peers. For behavioural responses, the autistic group exhibited lower accuracy compared to the two TD groups when identifying the audiovisual congruent stimuli. Regarding the perception of incongruent stimuli, significant group differences were detected, with the autistic group showing stronger auditory reliance and weaker audiovisual integration compared to their CA-matched TD peers in quiet and 10 dB conditions. Furthermore, weaker visual utilization in the autistic group than in their LA-

matched TD counterparts could be captured by the within-group differences. In terms of the talking-face processing manner, a lower likelihood of looking at the human face was revealed in the autistic group as opposed to both TD groups, while mouth-looking time was comparable among three groups of child participants. Generally, the face-looking ratio and mouth-looking time negatively predicted audio-dominant responses while positively correlated with audiovisual-integrated responses, regardless of group. Time-course analysis for the face-looking ratio successfully estimated that the time window of group differences occurred in the first half of the time window, while the analysis for mouth-looking time failed to detect any group differences over time.

5.4.1 Atypical Audiovisual Speech Processing in Children with ASD

Permutation-based repeated measures ANOVA examining the identification accuracy uncovered a significant group effect, which was mainly driven by the relatively less accurate identification of the autistic group compared to both TD groups, indicating a weaker capability among autistic children in identifying audiovisually-presented CV syllables regardless of auditory conditions. This finding was consistent with previous studies, where the autistic group benefited less from the facilitative effect of visual speech information (Foxe et al., 2015; Irwin et al., 2011; Smith & Bennetto, 2007; Stevenson et al., 2017, 2018).

When we further explored the strength of audiovisual integration across groups using the McGurk stimuli, significant between-group effects were observed between the autistic group and their CA-matched TD counterparts. In particular, significantly more audio-dominant responses, while fewer audiovisual-integrated responses, were recorded from the autistic group relative to their CA-matched TD counterparts in both quiet and 10 dB SNR conditions, supporting the preference for auditory unimodal processing instead of an audiovisual-integrated strategy in generating perceptual outcomes among autistic children. Another set of evidence came from the stark contrast in the dominant perceptual strategies adopted by the respective groups, as the audio-dominant strategy was overwhelming dominant in the autistic

group, while the audiovisual-integrated strategy prevailed in their CA-matched TD peers when auditory noise was absent or mild. Taken together, autistic children who lagged behind their peers in language ability could not achieve the competence commensurate with their chronological age in terms of audiovisual speech perception. This weaker integrative mechanism might be explained by the Weak Central Coherence (Frith, 1989; Happé & Frith, 2006), where autistic individuals were hypothesized to fragment their perception into parts and struggle to integrate multiple cues to form a holistic picture. On the other hand, the weaker audiovisual integration might be related to the wider temporal binding window observed in autistic individuals, which disrupted multisensory integration and led to independent processing rather than an integrated percept (Stevenson, 2016; Feng et al., 2021a).

With the comparison between the autistic group and their LA-matched TD peers, it can be concluded that language ability partially explained the behavioural abnormalities of the autistic group in perceiving incongruent stimuli. The autistic group exhibited behaviour more similar to their LA-matched TD peers, as none of their between-group differences attained the statistical significance, regardless of noise levels. These results provided an alternative perspective that underscored the association between audiovisual integration, as indexed by the McGurk response, and language ability, confirming that atypical sensory processing and integration might disrupt higher-order language skills (Feldman et al., 2022). When it came to the within-group differences, observable distinctions still existed. In the quiet condition, the percentage of audio-dominant responses significantly surpassed that of audiovisual-integrated ones in the autistic group, while this difference was insignificant in the LA-matched TD group. From this perspective, the autistic group was less adept in utilizing visual information when processing audiovisual speech stimuli in quiet and 10 dB SNR. Such differences might be due to the atypical manner of processing talking faces in the autistic group, as observed by previous studies (Irwin et al., 2011; Klin et al., 2002) and the current findings, which will be discussed in Section 5.4.2.

Differences observed between the two TD groups with a difference exceeding one year in mean age demonstrated the developmental effect on utilizing visual speech information in audiovisual speech perception. In quiet and 10 dB SNR conditions where the CA-matched TD group made significantly more audiovisual-integrated responses, this difference failed to reach statistical significance in the LA-matched TD group. In the -10 dB SNR condition where the visual-dominant strategy was adopted as a statistically optimal option for deriving perceptual outcomes by the CA-matched TD group (audiovisual-integrated vs. visual-dominant = 61.42% vs. 31.42%), the language-ability-matched TD group made a comparable number of these two types of responses (audiovisual-integrated vs. visual-dominant = 50.00% vs. 37.86%). To conclude, for typically developing children, the development of audiovisual speech perception progresses with a disengagement from auditory reliance and an enhancement in visual utilization.

The shifting effect of auditory noise was reflected in the perceptual strategy transitions observed across the three groups of participants. In particular, the performance of the autistic group tended to approach that of the two TD groups under noisy conditions. When auditory noise as mild as 10 dB SNR was introduced, the autistic group did not significantly differ from their LA-matched TD peers, in either between-group or within-group comparisons. When the auditory condition was as challenging as -10 dB SNR, the between-group differences between the autistic and CA-matched TD groups failed to reach significance, regardless of response types. Such a narrowed gap between autistic and TD children contradicted previous studies where audiovisual integration in autistic individuals remained at a lower level in noisy conditions (e.g., Liu et al., 2020). On the contrary, autistic children in the current study exhibited the ability to automatically adjust their perceptual strategies in response to changes in auditory noise. Furthermore, with the increase in auditory noise, the optimal strategy for the autistic group gradually shifted from audiovisual-integrated to a visual-dominant one, and the demand for audiovisual integration diminished. One possible explanation was that, at this point,

the optimal visual-dominant strategy might somewhat eliminate the disadvantages brought by the weakened integration mechanisms and/or the widened temporal binding window in autistic individuals, manifesting as a reduction in between-group differences.

5.4.2 Atypical Face-viewing Pattern in Autistic Children during Audiovisual Speech Perception

5.4.2.1 Face-looking Ratio

For the face-viewing pattern, a significant association between the likelihood of looking at the speaker's face and response type to incongruent stimuli was recognized by the GLMMs. Specifically, the face-looking ratio negatively predicted the occurrence of audio-dominant responses while positively correlating with that of audiovisual-integrated responses, regardless of group or auditory condition. In accordance with previous findings (Irwin et al., 2011; Yi et al., 2013), the duration of fixation on the speaker's face recorded from the autistic group was significantly suppressed relative to both TD groups, further indicating that autistic children were more likely to avoid directly gazing at the human face, even when the audiovisual stimuli were in a two-dimensional video setting. Moreover, when trying to portray the temporal changes of the face-looking ratio among three groups of children, the autistic group was estimated to have a lower likelihood of fixating on the speaker's face most of the time compared to their LA-matched-TD counterparts. Interestingly, although the time window with significant differences compared to the CA-matched TD group was narrower, it was precisely situated in the early-to-middle stage of the stimulus, which was the primary period of stimulus contrast in the current study (i.e., the initials of the CV syllables).

The avoidance of faces was hypothesized to be due to autistic individuals misinterpreting eye gaze information as social threats (Feng, et al., 2021; Tanaka & Sung, 2016). Such atypical mismapping might partially explain the previous finding that audiovisual integrative deficit seemed to be speech-specific (Stevenson et al., 2014a, 2016; Stewart et al., 2016), as speech-related tasks often require face processing. Considering the significant

predictability of the face-looking ratio, this atypical face avoidance appeared to profoundly impact their audiovisual speech perception. On the one hand, given that facial movement provides rich and comprehensive visible linguistic cues for speech decoding (Kuhl & Meltzoff, 1988; Skipper et al., 2005), increased visual attention directed to the speaker's face benefited audiovisual speech integration by allowing for more precise auditory and visual alignment (Fleming et al., 2021). Autistic children, who encounter greater challenges in temporally binding audiovisual information, might be placed in a more unfavourable situation with reduced visual hints (Stevenson et al., 2016). On the other hand, decreased visual attention to the human face might hamper audiovisual speech perception in the autistic group by preventing visual intake at its source, leading to a tendency to process audiovisual stimuli relying on auditory unimodal information, especially in the face-to-face communication contexts. Temporal analysis further disclosed that the disparity occurred during the time window where key linguistic information was most densely distributed, potentially leading autistic children to miss the critical time point to intake and bind essential linguistic cues, and therefore, diminishing the influence of visual cues in generating perceptual outcomes. Such an effect was particularly evident when auditory information was clear and prominent. Therefore, it could account for the shrunken audiovisual speech integration of the autistic group even when compared with their LA-matched TD counterparts whose chronological age was younger.

5.4.2.2 Mouth-looking time

For the mouth-looking time, consistent with Saalasti et al. (2012) and Yi et al. (2013), no differences between the autistic group and the two TD groups regarding mouth-looking time reached statistical significance. Temporal course analysis using GAMM further confirmed that mouth-looking time among the three groups of participants showed no differences throughout the 2000-ms stimulus time window, indicating their performance remained consistently comparable over time. Furthermore, the current findings enhanced the body of evidence suggesting that mouth-looking time is associated with audiovisual speech perception as

discussed in Chapters 3 and 4, as we found that intensified attention to the speaker's mouth area predicted reduced audio-dominant responses while increasing audiovisual-integrated responses in general.

Mouth-looking time reflected the processing of visual information in a finer grain size, which supports the disambiguation of speech signals (Baron et al., 2023; Lusk & Mitchel, 2016). In the current study, mouth-looking time was defined as the proportion of time spent fixating on the speaker's mouth relative to the total fixation duration looking at the speaker's whole face (Barr, 2008; Feng et al., 2021b), inferring the probability of eye gaze falling within the mouth AOI when viewing the speaker's face. The comparable mouth-looking time between the autistic and both TD groups that indicated the attention allocated to the mouth area in the autistic group might not necessarily be impaired. This finding was consistent with Yi et al. (2013), where the autistic group did not differ in the proportional fixation directed to the internal facial components when looking time at the whole face was controlled. Additionally, the time-course analysis of mouth-looking time did not detect significant between-group differences, with fixation on the mouth area increasing as articulation commenced and decreasing once the linguistic information presentation was complete, suggesting that autistic individuals were aware of the criticality of the mouth area to speech decoding and managed to appropriately adjust their visual attention distribution according to the density of linguistic cues conveyed by the stimuli.

The role of mouth-looking time in predicting response types to the incongruent stimuli seemed to be subject to noise levels. In quiet and 10 dB SNR conditions, where the intelligibility of auditory information was high, increased mouth-looking time predicted a reduced likelihood of making audio-dominant responses and was positively associated with the occurrence of audiovisual-integrated responses. Results regarding audio-dominant responses accorded with Feng et al. (2022), where autistic children were more likely to get rid of their hyper-reliance on unimodal auditory processing with more visual attention directed to the

speaker's mouth region. On the other hand, direct foveal fixation on the speaker's mouth region might give rise to enhanced audiovisual speech integration by lowering sensory noise, as mouth movements offered visual speech information that was direct and precise, as suggested by Gurler et al. (2015). When the auditory condition was as challenging as -10 dB SNR, the predictability of mouth-looking time was revealed to negatively predict audiovisual-integrated responses and was positively correlated with visual-dominant responses only in the LA-matched TD group. The prominent role of auditory noise in influencing the perceptual strategy adopted for responding to incongruent stimuli could be observed across groups, as the reliability of the auditory modality was undermined while visual reliability remained constant. Under this circumstance, perceptual outcomes would be generated with more weight placed on visual information in typical perceivers, and the statistically optimal strategy shifted from the audiovisual-integrated strategy to the visual-dominant one (Ernst & Banks, 2002; Hirst et al., 2018; Weng et al., 2024). The significant predictability of mouth-looking time in the LA-matched TD group could be attributed to the developmental process that TD children undergo to acquire an adult-like optimal manner (Weng et al., 2024). This younger TD group, whose audiovisual perceptual mechanism had not yet matured to be automatic, did not make significantly more visual-dominant over audiovisual-integrated responses under the -10 dB SNR condition, indicating that they still required explicit assistance from mouth movements to opt for the optimal strategy in deriving percepts. As for the CA-matched TD group, the facilitating role of mouth-looking time might be difficult to pinpoint, but we observed a tendency similar to the LA-matched TD group (see Figures 5.4 (B) and 5.5 (B)). The role of mouth-looking time could not be inferred from the autistic group, likely due to their greater difficulties in interpreting linguistic information from lip movement (Iarocci et al., 2010; Smith & Bennetto, 2007).

5.5 Conclusion

The behavioural responses and eye movements during the McGurk paradigm of Cantonese-speaking autistic children aged 8 to 11 years were compared with their TD counterparts matched on chronological age and language ability, respectively. Behavioural results revealed a weaker capability in audiovisual speech perception among autistic children compared to the CA-matched TD group, which could be illustrated by their lower identification accuracy in the congruent trials, together with a bias for unimodal auditory processing in incongruent trials. Language ability could partially account for the atypicality, as autistic children only demonstrated differences in response patterns in within-group comparisons with their LA-matched TD peers. The disparity in behavioural responses could be explained by the atypical eye movements recorded from autistic children, as they were less likely to allocate visual attention to the human face during speech perception tasks, especially during the early-to-middle stage of stimulus presentation. However, there were no differences in fixations on the speaker's mouth, suggesting that their visual intake during audiovisual speech perception was mainly constrained by a lack of general social interest, as reflected in the face-looking ratios, rather than an inability to recognize the importance of the mouth area for speech decoding.

Chapter 6. General Discussion and Conclusions

The current thesis explored the behavioural responses and eye movements in perceiving audiovisual speech stimuli among various groups of tonal language speakers, especially those native to HK Cantonese. In accordance with previous studies focusing on Indo-European language speakers, language background, developmental stage, and neurotypicality pose critical impacts on the audiovisual speech processing of perceivers from the East. Additionally, auditory noise and the manner of face processing also profoundly influence the strategies adopted for audiovisual speech perception across different groups of tonal language speakers. In this chapter, we will summarize and discuss the general findings from Chapters 3 to 5.

6.1 Audiovisual Speech Perception across Tonal-language-speaking Populations

The impacts of three group-level factors on audiovisual speech perception were examined through three main studies in this thesis. Results showed that all these factors contributed to variability in perceptual strategy, which was manifested by adjusting the perceptual reliance on respective modality cues.

6.1.1 Identifying Congruent Stimuli in Noise-free Condition

Overall, all participants across the three studies demonstrated a considerably high accuracy in identifying the congruent stimuli. Apart from the simplicity of the task, this high performance was also because the ability to clearly distinguish the three CV stimuli was one of the criteria for participant screening. However, younger participants were found to underperform in this task as the identification accuracy achieved by Cantonese-speaking children aged 4–5 years was significantly lower than that of any elder group. Given that this group of children had also passed both training sessions, we propose that this was due to the fact that focused attention is under development during the first five years of life (Ruff & Lawson, 1990). Consequently, this group of young children might encounter difficulty in maintaining consistent attention to

the task throughout the study. Thus, the ability to identify the three audiovisual congruent stimuli did not seem to vary dramatically among the several groups of tonal-language-speaking participants studied.

6.1.2 Perception of Incongruent Stimuli in Quiet Condition

The investigation of the perception of speech stimuli comprising conflicting audiovisual information in a quiet condition examined the weighing on audition and vision when both modalities were free of noise. In this case, we observed greater variation among child participants in Chapters 4 and 5 compared to the adult participants in Chapter 3.

Two factors were identified as strongly influencing the perceptual outcomes of audiovisual speech perception within Cantonese-speaking children: developmental stage and neurotypicality. First, guided by the results from Chapter 4, the strategy adopted for processing audiovisual speech stimuli changed with development. In line with developmental studies centering on Indo-European-language-speaking children (Hirst et al., 2018; Tremblay et al., 2007), a clear developmental shift in sensory dominance in audiovisual speech perception was tracked by the current cross-sectional study. Specifically, this developmental process began with the common preference for auditory unimodal information among young children and gradually transitioned into an adult-like strategy that integrates bimodal information when generating perceptual outcomes. Consistent with Weng et al. (2024), these findings contradicted the hypothesis that tonal-language-speaking children could be exempted from this developmental shift, evidenced by the significant gap in the favoured strategies between adults and children. However, the impact of age indeed appeared to interact with language background, affecting the timing of this developmental shift. Combining findings from the literature, age remains a critical aspect contributing to the diversity in perceptual strategies in audiovisual speech processing.

Second, according to findings from Chapter 5, autistic children were likely to derive perceptual outcomes using unimodal auditory information compared to their neurotypical counterparts. Autistic children included in the current study ranged in age from 8 to 11 years (mean = 9.41). Referring to the development trajectory depicted in Chapter 4, it can be inferred that the majority of TD children in this age range had already undergone the developmental shift in audiovisual speech perception. Data from the CA-matched TD group lent support for this statement, as this group of children primarily employed the audiovisual-integrated strategy to respond to the audiovisual incongruent stimuli. Considering language ability had been raised to be correlated with audiovisual integration, a group of TD children matched on language ability was also included. Results showed that language ability indeed partially accounted for the differences, yet considerable differences still persisted in within-group comparisons. Consistent with prior review papers, individuals with ASD exhibited atypicality in the mechanism for processing audiovisual stimuli, manifested in an over-reliance on unimodal auditory information.

Collectively, for the noise-free condition, the effects of developmental stage and neurotypicality on audiovisual speech perception among children were aligned, as both of them promoted the emergence of the audiovisual-integrated strategy. Thus, the perceptual patterns exhibited by two groups of children who benefit less from these two factors—namely the younger children and autistic children—were strikingly similar, with both of them showing a bias for unimodal auditory information. This explains why Zhang et al. (2019) noted in their review paper that the group differences between TD and autistic groups widened across the studies reviewed. However, it is noteworthy that the mechanisms leading to similar behavioural patterns in these two groups are likely different. For instance, Chapter 5 uncovered an atypical eye movement trajectory in the autistic group, specifically a distinct distribution of visual attention when processing the human face compared to the two TD groups (see discussion in section 6.2). Therefore, future studies are encouraged to consider the developmental shift in

audiovisual speech perception in children, as participants whose age falls within the range before undergoing the developmental shift may exhibit behaviour patterns similar to autistic children, despite fundamentally different underlying mechanisms.

From another perspective, both developmental stage and neurotypicality drive the evolution of perceptual strategies towards the adult-like audiovisual-integrated approach. In Chapter 3, two groups of adults from distinct tonal language backgrounds were tested, both of whom demonstrated high levels of audiovisual integration with no significant group differences, contrary to the hypothesis that the tonal aspect of languages might lead to a low level of audiovisual integration (Sekiyama, 1994; Sekiyama & Burnham, 2008). These findings also indicated that the audiovisual-integrated strategy predominates in adulthood when responding to audiovisual stimuli, regardless of whether participants' native language is Cantonese or Mandarin. In addition, regression results from Chapter 4 pointed out that such an audiovisual-integrated strategy marked the direction of the development of audiovisual speech perception. Also, in Chapter 5, both TD groups adopted this strategy more frequently relative to the autistic group. Taken together, the tendency to integrate audiovisual bimodal cues appears to be a natural development at the population level. If this natural progression is disrupted by factors associated with neuroatypicality, such as ASD, it may result in delayed or even altered patterns of perception development, which is less probable to be predicted by typical developmental trajectories.

6.2 Role of Auditory Noise in Audiovisual Speech Perception

The findings of the current thesis also revealed the profound impact of auditory noise on audiovisual speech processing. As auditory noise increased, participants showed a gradual decrease in their reliance on auditory information and a corresponding increase in dependence on visual input. The observed noise-induced effect followed the statistically optimal hypothesis in multisensory processing (Ernst & Banks, 2002; Weng et al., 2024). Across the various groups

of tonal language speakers involved in the study—despite differences in language background, developmental stage, and neurotypicality—the impact of auditory noise on perception strategies was consistently observed. Given this pervasive influence, the central issue was whether the introduction of auditory noise brought the different groups closer in performance or exacerbated the disparities between them. The answer varied depending on the audiovisual consistency of the stimulus.

6.2.1 Higher Susceptibility to Auditory Noise in Child Participants during Congruent Stimuli Identification

Findings from Chapter 3 suggested that the introduction of auditory noise did not lead to significant group differences between the two groups of adult participants with varying language background, indicating that auditory noise did not cause differences in their ability to identify the three CV syllables. However, this discussion is limited to the two tonal language backgrounds involved, where the three syllables in both languages were real words and were produced by similar places and manners of articulation.

With the results from Chapter 4, it could be observed that noise might widen the gap between children and adults in identifying congruent stimuli, as younger children from 6–7- and 8–9-year-old groups, who did not behave differently from their adult counterparts under quiet conditions, failed to achieve comparable accuracy with their adult counterparts under noisy conditions. Such a degradative role of auditory noise was tentatively attributed to the cooperative relationship between audition and vision in congruent trials (Weng et al., 2024). In this case, when the intelligibility of one modality (e.g., audition) is compromised by noise, the overall support for speech decoding weakens accordingly. Although the visual modality should receive greater reliance from perceivers as predicted by the statistically optimal hypothesis, children —whose language abilities were still developing—might struggle to fully utilize the less linguistically distinguishable information provided by vision (Gijbels et al., 2021; Kuhl & Meltzoff, 1988). As a result, such an underdeveloped capacity was likely to underlie the

widened gap between children and adults when auditory noise disrupted the dominant modality (e.g., audition) in such conditions.

This trend could also be observed in Chapter 5. Although the Group × Noise Level interaction did not reach significance, it could be inferred from the descriptive data that the differences between the autistic group and their TD counterparts appeared more pronounced under noisy conditions compared to the quiet condition. For instance, under the -10 dB SNR condition, the mean identification accuracy was lower in the autistic group (Mean = .57, *SE* = .04) than in the CA-matched TD group (Mean = .69, *SE* = .04) and the LA-matched TD group (Mean = .67, SE = .04). In contrast, under the quiet condition, the identification accuracy was similar across all groups (autistic: Mean = .99, *SE* = .01; CA-matched TD: Mean = .99, *SE* = .00; LA-matched TD: Mean = .99, SE = .00). The degradative effect of auditory noise appeared similar to that in Chapter 4, yet the underlying causes might not be necessarily consistent. As revealed by the eye-tracking data (see discussion in Section 6.2), the poorer performance in the autistic group might be linked to their lack of interest in social stimuli, and the resulting insufficient visual input might, therefore, lead to a tendency towards unimodal auditory information processing. Following this line, their unimodal bias that posed less impact in the quiet condition would contribute to the widened gap relative to TD groups when the intelligibility of primary information source (i.e., audition) was disrupted by noise.

6.2.2 Perceptual Strategies for Incongruent Stimuli Were Varied by Noise in Adults but Unified in Children

Stimuli with incongruent audiovisual information allow us to observe how the statistically optimal strategy in audiovisual speech processing shifts as auditory noise increases. Interestingly, in Chapter 3, auditory noise led to a significant divergence in perceptual strategies between two groups of adult participants from different language backgrounds who did not show significant differences in the quiet condition. In contrast, in Chapters 4 and 5, auditory noise narrowed the significant between-group differences observed in the quiet condition.

In Chapter 3, introducing auditory noise appeared to magnify the differences between the two groups of participants, as the impact of language background only manifested in noisy conditions. Specifically, Cantonese, which is phonologically more complex at both segmental and suprasegmental levels, might guide its native speakers to maintain considerable attention to the auditory modality even when the auditory conditions become less reliable (Zhang et al., 2018). This tendency made their perceptual strategy for audiovisual processing less susceptible to auditory noise compared to their Mandarin-speaking counterparts, as evidenced by the greater amount of auditory noise required for Cantonese speakers to shift to a visual-dominant strategy. The preference for an audiovisual-integrated strategy in Cantonese-speaking participants under noisy conditions was potentially driven by the joint influence of greater segmental and suprasegmental complexity, which demanded more attention to both modalities and required a higher degree of bimodal integration, particularly when auditory modality was disrupted by noise.

On the contrary, we observed that the introduction of auditory noise blurred the group-level differences in perceptual strategies for audiovisual speech processing among typical HK Cantonese-speaking children in Chapter 4. For instance, child participants aged from 8–9 years, who made significantly more audio-dominant responses than adults under quiet and 10 dB SNR conditions, did not exhibit significant differences across all responses when SNR dropped to -10 dB. In other words, in noisy conditions, an earlier emergence of adult-like patterns could be observed. This might be induced by the deprivation of audition by noise forced younger children to loosen their over-reliance on the auditory modality and, instead, to focus on visual information. However, the younger groups aged 4–7 years still displayed a clear auditory bias compared to adults, indicating that the disengagement with auditory preference also followed a developmental sequence. Regression results further confirmed the interplay between development and noise, as the development of audiovisual speech perception, including under noisy conditions, pointed to the direction of the statistically optimal strategy (Weng et al., 2024).

Similar to Chapter 4, in Chapter 5, it was also observed that the differences between the autistic group and their TD counterparts in noisy conditions were less pronounced. In the quiet condition, autistic participants showed a stronger preference for auditory information compared to the TD groups. Specifically, significant group differences were observed between the autistic group and their CA-matched counterparts. Albeit there were no significant group differences between the autistic group and their LA-matched TD peers, who were younger in chronological age, differences were revealed by within-group comparisons. However, when the auditory noise was at 10 dB SNR, both between- and within-group differences between the autistic group and their LA-matched TD group were eliminated. When the SNR fell to -10 dB, even the differences between their CA-matched TD group became insignificant. These findings indicate that the autistic group was capable of adjusting their strategies for processing audiovisual speech stimuli according to the level of auditory noise. When the auditory condition was as challenging as -10 dB SNR, their strategy might not significantly differ from that of the TD groups.

6.3 Role of Face Processing in Audiovisual Speech Perception

Audiovisual speech perception serves as a convergence point between speech perception and face processing, which is particularly interesting given that these two domains have been traditionally studied separately (Pascalis et al., 2014, 2020). Based on findings from the current thesis, it could be deduced that the relationship between audiovisual speech perception and talking face processing is much closer than expected, primarily evidenced in two key aspects. Firstly, different groups of participants in the current thesis exhibited distinct face-processing manners. Secondly, variations in face processing could, at least to some extent, account for the observed discrepancies in their behavioural responses to audiovisual speech stimuli. Specifically, participants' visual allocation towards the speakers' mouth area was of interest—not only because the mouth area is where fine linguistic cues are densely distributed (Meltzoff

& Kuhl, 1994; Rennig & Beauchamp, 2018; Skipper et al., 2005, 2007), but also due to the proposed relationship between audiovisual integration and the visual attention directed to the mouth area (Feng et al., 2021b, 2022; Gurler et al., 2015). Incorporating temporal analysis, all the three group-level factors, namely, language background, developmental stage, and neurotypicality, have been shown to profoundly influence the manner of processing talking faces.

Firstly, language background appeared to modulate the manner of talking face processing, as suggested by Chapter 3. Specifically, Cantonese-speaking participants, who were from a language background more complex in phonology, exhibited a higher dependence on the visual cues offered by speakers' mouth area, and this dependence could be illustrated by two perspectives. First, when identifying congruent stimuli, mouth-looking time positively predicted the accuracy in the Cantonese group but not in the Mandarin group. Second, analysis of the temporal evolution of mouth-looking time revealed that the Cantonese group demonstrated a higher probability of looking at the mouth area during the mid-to-late period of the 2000-ms stimulus window compared to their Mandarin-speaking counterparts. Taken together, these findings suggest that the greater phonological complexity led the Cantonese group to rely more heavily on the fine linguistic information provided by mouth movements when processing audiovisual speech stimuli, resulting in a talking face processing pattern distinctive from that of the Mandarin group. Moreover, this face-processing manner was likely driven by segmental, rather than suprasegmental, complexity, as the production of segments involved richer, distinguishable visual cues to aid speech decoding (Sekiyama, 1994). Coupled with the intensified reliance on the auditory modality due to the more complex suprasegmental phonology in Cantonese, Cantonese speakers thus developed a behavioural preference for an audiovisual-integrated strategy while processing audiovisual speech stimuli in noisy conditions.

From Chapter 4, it was observed that the manner of talking face processing during speech perception tasks becomes nuanced with development. Specifically, younger children aged 4–9 years exhibited a lower probability of looking at the speaker's mouth area during the early-to-middle period of the 2000-ms stimulus window compared to older children and adults. Such a difference diminished with age, and by the age of 10–11, children performed comparably with adults. These findings strongly supported the close relationship between audiovisual speech perception and talking face processing (Irwin et al., 2017; Pascalis et al., 2014; Yamamoto et al., 2019). Firstly, their developmental processes displayed striking synchrony in their development trajectories, as children aged 4–9 exhibited differences from adults in both behavioural and eye movement data, while 10–11-year-olds aligned with adults in both measurements. Secondly, unlike the window of significance observed between two adult groups in Chapter 3, the development effects between children and adults were primarily evident during the early-to-mid period of stimulus presentation, where consonant contrasts occurred. These findings indicate that the development of talking face processing during speech perception tasks is significantly shaped by language knowledge and experience (Lewkowicz & Hansen-Tift, 2012). Elder children and adult participants, who manage to swiftly direct visual attention towards the speaker's mouth area once stimulus presentation begins, are more susceptible to the interference of visual information, while younger children, who inherently prefer auditory information due to a less mature integrative mechanism (Robinson & Sloutsky, 2010), are likely to miss the critical visual assistance, contributing to their less pronounced audiovisual integration.

Results from Chapter 5 revealed that the impact of autism on talking face processing was not mediated by visual attention directed to the speaker's mouth area but rather by attention to the human face, suggesting that the core abnormality in individuals with ASD lies in the lack of general interest in social engagement (Grelotti et al., 2002), despite the fact that they showed awareness of the importance of mouth movements for audiovisual speech processing (Yi et al.,

2013). Such avoidance of social stimuli aligns with the critical challenges faced by this clinical group, which might directly and overtly affect audiovisual speech perception by cutting off the visual intake at its source and eventually resulting in weakened audiovisual integration (Irwin et al., 2011). Therefore, the underperformance of the autistic group when identifying audiovisual congruent stimuli in noisy conditions was likely due to their over-reliance on unimodal auditory information and avoidance of visual contact with the human face (Klin et al., 2002; Tanaka & Sung, 2016), which led them to benefit less from the facilitation offered by the visual modality. On the other hand, given that the mouth-looking time was calculated by dividing the time spent looking at the speaker's mouth area by the time spent on her whole face, the comparable mouth-looking time between autistic and TD groups indicated that, when individuals with ASD did direct their visual attention toward the speaker's face, the probability of looking at the mouth area was comparable to that of their TD peers. Accordingly, the primary difficulty encountered by autistic individuals did not appear to stem from an inability to recognize or utilize the fine linguistic information provided by mouth movements but rather from their reduced interest in social engagement, as reflected by their lower face-looking ratio.

With the assistance of eye-tracking technology, this study explored the relationship between the manner of processing talking faces during audiovisual speech perception. Results indicate that different groups of tonal-language-speaking perceivers may employ distinct strategies in processing talking faces, which quite possibly underlies the behavioural discrepancies in audiovisual speech perception, supporting the idea that speech perception and face processing are not as distant as traditionally believed.

Chapter 7. Significance and Limitation of the Study

In this chapter, we shall summarize the findings of the current thesis. On top of that, the significance of our study, together with some limitations, will be concluded. Additionally, we will propose several future directions that we find interesting for upcoming research to explore.

7.1 Summary of Findings

First, we found that tonal language backgrounds, especially the difference in phonological complexity, significantly impacted audiovisual speech perception. From the behavioural perspective, Cantonese- and Mandarin-speaking participants, who showed no differences in identifying congruent stimuli across noise levels, demonstrated a comparably high degree of audiovisual integration when perceiving incongruent stimuli in the quiet condition, contrary to previous studies where the tonal aspect was argued to lead to weaker integration due to over-reliance on auditory modality. However, different tonal language backgrounds might influence the perceptual strategy for audiovisual speech stimuli in noisy conditions as native speakers of Cantonese, which is more complex in both segmental and suprasegmental aspects, made significantly more audiovisual-integrated responses and fewer visual-dominant ones compared to their Mandarin-speaking counterparts. Moreover, as revealed by eye-tracking data, greater phonological complexity might also lead Cantonese speakers to rely more on fine-grained visual linguistic cues offered by the mouth area of speakers. Taken together, the preference for audiovisual-integrated strategy in noisy conditions seems to arise from the combination of this featural visual attention allocation pattern alongside their intensified attention to the auditory modality due to the higher aural ambiguity introduced by the complex nature of Cantonese phonology.

Second, Cantonese-speaking children were found to adopt different strategies to process audiovisual speech stimuli until they reached 10 years of age, with both behavioural and eye-tracking evidence. When identifying audiovisual congruent stimuli, Cantonese-

speaking children aged 4 to 9 years could not achieve comparable accuracy to adults. For the perception of incongruent stimuli, consistent with previous findings, the current study confirmed that tonal language background would not eliminate the experience of the developmental shift in sensory dominance in audiovisual speech perception, as children aged 4 to 9 years made significantly more audio-dominant responses and fewer audiovisual-integrated responses compared to adults in the quiet condition. No significant differences were detected between the 10–11-year-old group and adults. Eye-tracking data uncovered a synchronized developmental course in visual allocation on talking faces, as younger children aged 4 to 9 years showed reduced probability of directing their eye gaze on the mouth area of the speakers during the early-to-middle stage of stimuli presentation compared to adults, while the 10–11-year-old group, again, did not significantly differ from adults. Findings from the current study indicate that the tonal property of languages may not exempt its young speakers from undergoing the developmental shift in audiovisual speech perception, while the timing of this shift may be influenced by language background. Moreover, the link between speech perception and talking face processing is strengthened by the current findings of a synchronous developmental course shared by these two interacting processes.

Finally, we found that autistic individuals showed differences in processing audiovisual speech stimuli. In terms of identifying congruent stimuli, autistic individuals could not achieve a comparable accuracy compared with two groups of TD participants. For the perception of incongruent stimuli, the autistic group was found to make significantly more audio-dominant responses compared to their CA-matched TD counterparts. When compared with a group of TD controls matched on language ability, they still exhibited differences in within-group comparisons. The bias towards the audio-dominant strategy in autistic individuals might be due to their atypical processing of talking faces stemming from the avoidance of social stimuli, which resulted in reduced visual intake when processing bimodal information. However, the differences in mouth-looking time between the autistic group and their TD counterparts did not

reach significance, indicating autistic individuals may be aware of and capable of utilizing the visual linguistic cues offered by mouth movements.

7.2 Significance of Findings

The current studies make a significant contribution to the field by offering a novel perspective on audiovisual speech processing. By integrating behavioral and eye-tracking data, it provides a more comprehensive understanding of how individuals dynamically allocate sensory attention during audiovisual speech perception. This multi-level approach enhances the ecological validity of speech processing research and offers deeper insights into the interaction between visual and auditory modalities.

To the best of our knowledge, this is the first comprehensive study investigating audiovisual speech perception among tonal language speakers on a large scale, with a particular focus on Cantonese speakers, whose situation has been understudied. Traditionally, research in speech perception has placed a particular emphasis on unimodal auditory processing, with comparatively little exploration of an audiovisual bimodal setting, despite the fact that speech communication in real-world scenarios frequently engages audiovisual processing. Findings from tonal language speakers in the current study provide an essential complement to existing research on audiovisual speech processing, which has predominantly focused on Indo-European language speakers, and hence, contribute to a more holistic understanding of the mechanisms underlying audiovisual speech perception across language backgrounds.

Furthermore, this is the first large-scale study to integrate behavioural and eye-tracking data in the exploration of audiovisual speech perception, which has allowed us to better assess the role of talking face processing in speech decoding. The findings from the current study have highlighted a close relationship between audiovisual speech perception and talking face processing, two domains conventionally studied separately. In particular, the manner of scanning talking faces effectively predicted the behavioural responses in many cases, stressing

the interconnected underpinnings shared by these two key components of social communications. Most importantly, our study provides mutually corroborative evidence from both behavioural and eye-tracking data supporting the synchronous development of audiovisual speech perception and talking face processing for the first time, which has enabled us to further clarify a multimodal developmental process that incorporates speech decoding and talking face processing.

Moreover, the findings of this study may shed light on the rehabilitation and intervention strategies designed for children with ASD. Our findings has shown that autistic children may be less likely to fixate on the human face when processing audiovisual speech stimuli, suggesting a reduced general interest in social cues. In this context, further research could focus on understanding the underlying mechanisms of such social avoidance, while also developing more tailored and targeted intervention programs for this clinical population. From a clinical perspective, clinicians and caregivers can play a crucial role in supporting autistic children in overcoming challenges related to the misreading or misinterpretation of facial information. Encouraging increased visual attention to the human face may enhance the ability of autistic individuals to integrate visual speech cues, thereby improving speech decoding and comprehension in social interactions.

7.3 Limitations of the Study

The current study faces the following limitations.

First, the nature of the stimuli may constrain the interpretability and generalizability of findings. While the classic McGurk paradigm has been adopted to investigate audiovisual speech perception, it involves a restricted set of stimuli. More specifically, only syllabic stimuli were adopted in the current study, leaving a broader range of more complex speech stimuli unexamined. Moreover, though the classic McGurk stimuli, namely, the incongruent stimuli, have been frequently adopted as the measurement of audiovisual speech integration, they lack

naturality because they are rarely seen in the real-world scenario. Additionally, due to attentional constraints in young participants, only stimuli generated from a single female speaker were used in Chapters 4 and 5 of the current thesis, which may also affect the external validity of the findings.

Second, for participant inclusion, more detailed demographic variables in more depth (including IQ and executive functions) would be warranted for a refined characterization of child participants, particularly for the autistic group in Chapter 5. The measurement of additional demographic indices was mainly prevented by practical constraints, potentially reducing the granularity of insights into this population.

Third, regarding data analysis, our findings were mainly derived from group-level comparisons to capture core features or differences, which might not fully account for the substantial heterogeneity within individuals with ASD.

In addition, it should be noted that our findings regarding face processing patterns are specific to the speech perception tasks employed in this study. Previous research has shown that face processing trajectories can vary significantly depending on the tasks used (e.g., Yamamoto et al., 2019), and thus, it should be treated with caution when generalizing our observations to other contexts. It is also important to acknowledge that the face processing patterns recorded in this study were observed under strictly controlled laboratory settings, which may limit the ecological validity of the findings.

7.4 Future Directions

First, future research may consider expanding the stimulus variety and complexity by incorporating a broader array of speech stimuli, including those with greater linguistic complexity (e.g., phrase, sentence, and even discourse) and varying phonetic features, to better stimulate real-world social interaction. Incorporating a more extensive and diverse set of

stimuli could help improve the generalizability of findings and offer deeper insights into audiovisual speech perception mechanisms across different types of linguistic inputs.

Second, for a more nuanced understanding of audiovisual speech processing in children, especially those diagnosed with ASD, future studies should consider assessing additional demographic and cognitive factors, including but not limited to IQ and executive functions. Including these variables would provide a more detailed profile of ASD participants and help refine interpretations regarding individual differences in audiovisual speech perception.

Third, in future research, it would be important to move beyond group-level comparisons and explore the within-group heterogeneity more thoroughly. By considering individual characteristics as continuous variables within models, we can better understand how these variables contribute to outcomes and explore how heterogeneity per se influences the results. This will provide more nuanced insights into the complex nature of these traits and their effects on the broader population.

Additionally, to explore the variability of face-processing strategies beyond laboratory settings, future studies may consider including a wider range of tasks and settings with wearable eye-tracking devices. Moving beyond desktop-fixed eye-tracking methods would allows future investigations to reveal task-dependent variations in face processing and clarify how experimental conditions and real-world contexts influence face-processing strategies.

References

American Psychiatric Association. (2000). Diagnostic and Statistical Manual of Mental Disorders (4th ed., text rev.)

American Psychiatric Association. (2013). Diagnostic and Statistical Manual of Mental Disorders. In *American Psychiatric Publishing*.

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. https://doi.org/10.1044/1092-4388(2007/080)

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

Baranek, G. T. (1999). Autism During Infancy: A Retrospective Video Analysis of Sensory-Motor and Social Behaviors at 9–12 Months of Age. *Journal of Autism and Developmental Disorders*, *29*(3), 213–224. https://doi.org/10.1023/A:1023080005650

Baron, A., Harwood, V., Kleinman, D., Campanelli, L., Molski, J., Landi, N., & Irwin, J. (2023). Where on the face do we look during phonemic restoration: An eye-tracking study. *Frontiers in Psychology*, *14*, 1005186. https://doi.org/10.3389/fpsyg.2023.1005186

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism, Males and Females, Scientists and Mathematicians. *Journal of Autism and Developmental Disorders*, *31*(1), 5–17. https://doi.org/10.1023/A:1005653411471

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*(4), 457–474. https://doi.org/10.1016/j.jml.2007.09.002

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Basu Mallick, D., F. Magnotti, J., & S. Beauchamp, M. (2015). Variability and stability in the McGurk effect: Contributions of participants, stimuli, time, and response type. *Psychonomic Bulletin & Review*, *22*(5), 1299–1307. https://doi.org/10.3758/s13423-015-0817-4

Bauer, R. S. (2016). The Hong Kong Cantonese language: Current features and future prospects. *Global Chinese*, *2*(2), 115–161. https://doi.org/10.1515/glochi-2016-0007

Bauer, R. S., & Benedict, P. K. (1997b). *Modern Cantonese Phonology:* De Gruyter Mouton. https://doi.org/10.1515/9783110823707

Belin, P. (2017). Similarities in face and voice cerebral processing. *Visual Cognition*, *25*(4–6), 658–665. https://doi.org/10.1080/13506285.2017.1339156

Belin, P., Bestelmeyer, P. E. G., Latinus, M., & Watson, R. (2011). Understanding Voice Perception. *British Journal of Psychology*, *102*(4), 711–725. https://doi.org/10.1111/j.2044-8295.2011.02041.x

Bernstein, L. E. (2012). Visual speech perception. In *Audiovisual speech processing* (pp. 21–39).

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

Birulés, J., Goupil, L., Josse, J., & Fort, M. (2023). The Role of Talking Faces in Infant Language Learning: Mind the Gap between Screen-Based Settings and Real-Life Communicative Interactions. *Brain Sciences*, *13*(8), 1167. https://doi.org/10.3390/brainsci13081167

Burnham, D., Vatikiotis-Bateson, E., Vilela Barbosa, A., Menezes, J. V., Yehia, H. C., Morris, R. H., Vignali, G., & Reynolds, J. (2022). Seeing lexical tone: Head and face motion

in production and perception of Cantonese lexical tones. *Speech Communication*, *141*, 40–55. https://doi.org/10.1016/j.specom.2022.03.011

Burr, D., & Gori, M. (2012). Multisensory Integration Develops Late in Humans. In *The Neural Bases of Multisensory Processes*.

Charney, S. A., Camarata, S. M., & Chern, A. (2021). Potential Impact of the COVID-19 Pandemic on Communication and Language Skills in Children. *Otolaryngology–Head and Neck Surgery*, *165*(1), 1–2. https://doi.org/10.1177/0194599820978247

Chen, Y., & Hazan, V. (2009). Developmental factors and the non-native speaker effect in auditory-visual speech perceptiona). *The Journal of the Acoustical Society of America*, *126*(2), 858. https://doi.org/10.1121/1.3158823

Cheng, S. (2023). Hong Kong Lifts Citywide Mask Mandate After Almost Three Years. *The Wall Street Journal*. https://www.wsj.com/articles/hong-kong-lifts-citywide-mask-mandate-after-almost-three-years-da759c3a

Chládková, K., Podlipský, V. J., Nudga, N., & Šimáčková, Š. (2021). The McGurk effect in the time of pandemic: Age-dependent adaptation to an environmental loss of visual speech cues. *Psychonomic Bulletin & Review*, *28*(3), 992–1002. https://doi.org/10.3758/s13423-020-01852-2

Crystal, D. (2008). *A dictionary of linguistics and phonetics* (6th ed). Blackwell Pub.

de Gelder, B., Bertelson, P., Vroomen, J., & Chen, H. J. (1995). Interlanguage differences in the McGurk effect for Dutch and Cantonese listeners. *Proceedings of the Fourth European Conference on Speech Communication and Technoloy*, 1699–1702.

de Boer M. J., Başkent D., & Cornelissen F. W. (2020). *Eyes on Emotion: Dynamic Gaze Allocation During Emotion Perception From Speech-Like Stimuli*. https://doi.org/10.1163/22134808-bja10029

Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015). Functional Organization of Social Perception and Cognition in the Superior Temporal Sulcus. *Cerebral Cortex*, *25*(11), 4596–4609. https://doi.org/10.1093/cercor/bhv111

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

DePape, A. M. R., Hall, G. B. C., Tillmann, B., & Trainor, L. J. (2012). Auditory Processing in High-Functioning Adolescents with Autism Spectrum Disorder. *PLOS ONE*, *7*(9), e44084. https://doi.org/10.1371/JOURNAL.PONE.0044084

Department of Chinese Language and Literature of Peking University. (2004). *Modern Chinese*. The Commercial Press. [In Chinese]

Dodd, B., Mcintosh, B., Erdener, D., & Burnham, D. (2008). Perception of the auditory-visual illusion in speech perception by children with phonological disorders. *Clinical Linguistics & Phonetics*, *22*(1), 69–82. https://doi.org/10.1080/02699200701660100

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

Eberhard, D. M., Simons, G. F., Fennig, C. D., & Summer Institute of Linguistics (Eds.). (2022). *Ethnologue: Languages of the Americas and the Pacific* (Twenty-fifth edition). SIL.

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*(3), 651–653. https://doi.org/10.1121/1.383691

Ernst, M. O. (2006). A Bayesian view on multimodal cue integration. In *Human body perception from the inside out* (pp. 105–131). Oxford University Press.

Ernst, M. O. (2008). Multisensory Integration: A Late Bloomer. *Current Biology*, *18*(12), 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*(6870), 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*(4), 162–169. https://doi.org/10.1016/J.TICS.2004.02.002

Farroni, T., Csibra, G., Simion, F., & Johnson, M. H. (2002). Eye contact detection in humans from birth. *Proceedings of the National Academy of Sciences*, *99*(14), 9602–9605. https://doi.org/10.1073/pnas.152159999

Feldman, J. I., Conrad, J. G., Kuang, W., Tu, A., Liu, Y., Simon, D. M., Wallace, M. T., & Woynaroski, T. G. (2022). Relations Between the McGurk Effect, Social and Communication Skill, and Autistic Features in Children with and without Autism. *Journal of Autism and Developmental Disorders*, *52*(5), 1920–1928. https://doi.org/10.1007/s10803-021-05074-w

Feldman, J. I., Dunham, K., Cassidy, M., Wallace, M. T., Liu, Y., & Woynaroski, T. G. (2018). Audiovisual multisensory integration in individuals with autism spectrum disorder: A systematic review and meta-analysis. *Neuroscience and Biobehavioral Reviews*, *95*, 220–234. https://doi.org/10.1016/j.neubiorev.2018.09.020

Feng, S., Lu, H., Fang, J., Li, X., Yi, L., & Chen, L. (2021a). Audiovisual speech perception and its relation with temporal processing in children with and without autism. *Reading and Writing*. https://doi.org/10.1007/s11145-021-10200-2

Feng, S., Lu, H., Wang, Q., Li, T., Fang, J., Chen, L., & Yi, L. (2021b). Face-viewing patterns predict audiovisual speech integration in autistic children. *Autism Research*, *14*(12), 2592–2602. https://doi.org/10.1002/aur.2598

Feng, S., Wang, Q., Hu, Y., Lu, H., Li, T., Song, C., Fang, J., Chen, L., & Yi, L. (2022). Increasing audiovisual speech integration in autism through enhanced attention to mouth. *Developmental Science*. https://doi.org/10.1111/desc.13348

Fleming, J. T., Maddox, R. K., & Shinn-Cunningham, B. G. (2021). Spatial alignment between faces and voices improves selective attention to audio-visual speech. *The Journal of the Acoustical Society of America*, *150*(4), 3085–3100. https://doi.org/10.1121/10.0006415

Foxe, J. J., Molholm, S., Del Bene, V. A., Frey, H.-P., Russo, N. N., Blanco, D., Saint-Amour, D., & Ross, L. A. (2015). Severe Multisensory Speech Integration Deficits in High-Functioning School-Aged Children with Autism Spectrum Disorder (ASD) and Their Resolution During Early Adolescence. *Cerebral Cortex*, *25*(2), 298–312. https://doi.org/10.1093/cercor/bht213

Franco-Watkins, A. M., & Johnson, J. G. (2011). Decision moving window: Using interactive eye tracking to examine decision processes. *Behavior Research Methods*, *43*(3), 853–863. https://doi.org/10.3758/s13428-011-0083-y

Frith, U. (1989). *Autism: Explaining the enigma*. Blackwell Scientific Publications.

Fromkin, V. (1978). *Tone: A linguistic survey*. Academic press, Harcourt Brace Jovanovich.

Frossard, J., & Renaud, O. (2021). Permutation Tests for Regression, ANOVA, and Comparison of Signals: The permuco Package. *Journal of Statistical Software*, *99*(15). 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*(12), 5022–5040. https://doi.org/10.1044/2021_JSLHR-21-00196

Goldstein, L., & Fowler, C. A. (2003). Articulatory Phonology: A phonology for public language use. In N. O. Schiller & A. S. Meyer (Eds.), *Phonetics and Phonology in*

Language Comprehension and Production (pp. 159–208). De Gruyter Mouton. https://doi.org/10.1515/9783110895094.159

Gori, M., Del Viva, M., Sandini, G., & Burr, D. C. (2008). Young Children Do Not Integrate Visual and Haptic Form Information. *Current Biology*, *18*(9), 694–698. https://doi.org/10.1016/j.cub.2008.04.036

Grandon, B., Schlechtweg, M., & Ruigendijk, E. (2023). Processing of noun plural marking in German-speaking children: An eye-tracking study. *Journal of Child Language*, 1–28. https://doi.org/10.1017/S0305000923000521

Graven, S. N., & Browne, J. V. (2008). Auditory Development in the Fetus and Infant. *Newborn and Infant Nursing Reviews*, *8*(4), 187–193. https://doi.org/10.1053/j.nainr.2008.10.010

Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & Psychophysics*, *50*(6), 524–536. https://doi.org/10.3758/BF03207536

Grelotti, D. J., Gauthier, I., & Schultz, R. T. (2002). Social interest and the development of cortical face specialization: What autism teaches us about face processing. *Developmental Psychobiology*, *40*(3), 213–225. https://doi.org/10.1002/dev.10028

Grieser, D. L., & Kuhl, P. K. (1988). Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese. *Developmental Psychology*, *24*(1), 14–20. https://doi.org/10.1037/0012-1649.24.1.14

Grossman, R. B. (2015). Judgments of social awkwardness from brief exposure to children with and without high-functioning autism. *Autism*, *19*(5), 580–587. https://doi.org/10.1177/1362361314536937

Gurler, D., Doyle, N., Walker, E., Magnotti, J., & Beauchamp, M. (2015). A link between individual differences in multisensory speech perception and eye movements.

Attention, Perception, and Psychophysics, *77*(4), 1333–1341. https://doi.org/10.3758/s13414-014-0821-1

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

Hasegawa, Y., & Hata, K. (1992). Fundamental Frequency as an Acoustic Cue to Accent Perception. *Language and Speech*, *35*(1–2), 87–98. https://doi.org/10.1177/002383099203500208

Hayes, B. (2009). *Introductory phonology*. Wiley-Blackwell.

Hazan, V., Kim, J., & Chen, Y. (2010). Audiovisual perception in adverse conditions: Language, speaker and listener effects. *Speech Communication*, *52*(11–12), 996–1009. https://doi.org/10.1016/j.specom.2010.05.003

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., Tiippana, K., Loberg, O., & Leppänen, P. H. T. (2018). Neural Processing of Congruent and Incongruent Audiovisual Speech in School-Age Children and Adults: Audiovisual Speech Processing in Children. *Language Learning*, *68*, 58–79. https://doi.org/10.1111/lang.12266

Hein, G., & Knight, R. T. (2008). Superior Temporal Sulcus—It's My Area: Or Is It? *Journal of Cognitive Neuroscience*, *20*(12), 2125–2136. https://doi.org/10.1162/jocn.2008.20148

Helbig, H. B., & Ernst, M. O. (2007). Optimal integration of shape information from vision and touch. *Experimental Brain Research*, *179*(4), 595–606. https://doi.org/10.1007/s00221-006-0814-y

Hirst, R. J., McGovern, D. P., Setti, A., Shams, L., & Newell, F. N. (2020). What you see is what you hear: Twenty years of research using the Sound-Induced Flash Illusion. *Neuroscience & Biobehavioral Reviews*, *118*, 759–774. https://doi.org/10.1016/j.neubiorev.2020.09.006

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). https://doi.org/10.1038/s41598-018-30798-8

Hisanaga, S., Sekiyama, K., Igasaki, T., & Murayama, N. (2016). Language/Culture Modulates Brain and Gaze Processes in Audiovisual Speech Perception. *Scientific Reports*, *6*(1), 35265. https://doi.org/10.1038/srep35265

Ho, C. S. H., & Bryant, P. (1997). Development of Phonological Awareness of Chinese Children in Hong Kong. In *Journal of Psycholinguistic Research* (Vol. 26, Issue 1).

Holm, A., & Dodd, B. (1996). The effect of first written language on the acquisition of English literacy. In *Cognition* (Vol. 59, Issue 2, pp. 119–147). https://doi.org/10.1016/0010-0277(95)00691-5

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

Irwin, J., Brancazio, L., & Volpe, N. (2017). The development of gaze to a speaking face. *The Journal of the Acoustical Society of America*, *141*(5), 3145–3150. https://doi.org/10.1121/1.4982727

Irwin, J. R., Tornatore, L. A., Brancazio, L., & Whalen, D. H. (2011). Can Children With Autism Spectrum Disorders “Hear” a Speaking Face?: Audiovisual Speech Perception in ASD. *Child Development*, *82*(5), 1397–1403. https://doi.org/10.1111/j.1467-8624.2011.01619.x

Johnson, C. E. (2000). Childrens' Phoneme Identification in Reverberation and Noise. *Journal of Speech, Language, and Hearing Research*, *43*(1), 144–157. https://doi.org/10.1044/jslhr.4301.144

Kanner, L. (1943). Autistic disturbances of affective contact. *Nervous Child*, *2*(3), 217–250.

Keane, B. P., Rosenthal, O., Chun, N. H., & Shams, L. (2010). Audiovisual integration in high functioning adults with autism. *Research in Autism Spectrum Disorders*, *4*(2), 276–289. https://doi.org/10.1016/j.rasd.2009.09.015

Keil, J., Müller, N., Ihssen, N., & Weisz, N. (2012). On the Variability of the McGurk Effect: Audiovisual Integration Depends on Prestimulus Brain States. *Cerebral Cortex*, *22*(1), 221–231. https://doi.org/10.1093/cercor/bhr125

Kirchner, J. C., Hatri, A., Heekeren, H. R., & Dziobek, I. (2011). Autistic Symptomatology, Face Processing Abilities, and Eye Fixation Patterns. *Journal of Autism and Developmental Disorders*, *41*(2), 158–167. https://doi.org/10.1007/s10803-010-1032-9

Klee, T., Wong, A. M.-Y., Stokes, S. F., Fletcher, P., & Leonard, L. B. (2008). Assessing Cantonese-speaking Children with Language Difficulties from the Perspective of Evidencebased Practice: Current Practice and Future Directions. In S.-P. Law, B. Weekes, & A. M.-Y. Wong (Eds.), *Language Disorders in Speakers of Chinese* (pp. 89–111). Multilingual Matters. https://doi.org/10.21832/9781847691170-008

Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D. (2002). Visual Fixation Patterns During Viewing of Naturalistic Social Situations as Predictors of Social Competence in Individuals With Autism. *Archives of General Psychiatry*, *59*(9), 809. https://doi.org/10.1001/archpsyc.59.9.809

Kong, J. (2007). *Laryngeal dynamics and physiological models: High speed imaging and acoustical techniques* (1st ed). Peking University Press.

Król, M. E. (2018). Auditory noise increases the allocation of attention to the mouth, and the eyes pay the price: An eye-tracking study. *PLOS ONE*, *13*(3), e0194491. https://doi.org/10.1371/journal.pone.0194491

Kubozono, H. (1995). Perceptual evidence for the mora in Japanese. In *Phonology and phonetic evidence: Papers in laboratory phonology* (pp. 141–156).

Kuhl, P. K., & Meltzoff, A. N. (1988). Perceptual Development in infancy. In *Speech as an Intermodal Object of Perception* (pp. 235–266).

Kyle, F. E., Campbell, R., Mohammed, T., Coleman, M., & MacSweeney, M. (2013). Speechreading Development in Deaf and Hearing Children: Introducing the Test of Child Speechreading. *Journal of Speech, Language, and Hearing Research*, *56*(2), 416–426. https://doi.org/10.1044/1092-4388(2012/12-0039)

Ladefoged, P., & Johnson, K. (2015). *A Course in phonetics* (Seventh edition). Cengage Learning.

Lalonde, K., & Holt, R. F. (2015). Preschoolers benefit from visually salient speech cues. *Journal of Speech, Language, and Hearing Research*, *58*(1), 135–150. https://doi.org/10.1044/2014_JSLHR-H-13-0343

Lalonde, K., & McCreery, R. W. (2020). Audiovisual enhancement of speech perception in noise by school-age children who are hard of hearing. *Ear and Hearing*, *41*(4), 705. https://doi.org/10.1097/AUD.0000000000000830

Lalonde, K., & Werner, L. A. (2021). Development of the Mechanisms Underlying Audiovisual Speech Perception Benefit. *Brain Sciences*, *11*(1), 49. https://doi.org/10.3390/brainsci11010049

Lee, T., Lo, W. K., Ching, P. C., & Meng, H. (2002). Spoken language resources for Cantonese speech processing. *Speech Communication*, *36*(3–4), 327–342. https://doi.org/10.1016/S0167-6393(00)00101-1

Lee, W.-S., & Zee, E. (2003). Standard Chinese (Beijing). *Journal of the International Phonetic Association*, *33*(1), 109–112. https://doi.org/10.1017/S0025100303001208

Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences*, *109*(5), 1431–1436. https://doi.org/10.1073/pnas.1114783109

Li, Y., Mei, L., & Dong, Q. (2008). The characteristics and development of audiovisual speech perception in native Chinese speakers. *Psychological Development and Education*, *3*(24), 43–47.

Lin, Y., Lin, Y.-J., Wang, F., Wu, X., & Kong, J. (2020). The development of phonological awareness and Pinyin knowledge in Mandarin-speaking school-aged children. *International Journal of Speech-Language Pathology*, *22*(6), 660–668. https://doi.org/10.1080/17549507.2020.1819417

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*(26), 232–239

.Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., & Bishop, S. (2012). Autism Diagnostic Observation Schedule: ADOS–2. Los Angeles, CA: Western Psychological Services.

Lusk, L. G., & Mitchel, A. D. (2016). Differential Gaze Patterns on Eyes and Mouth During Audiovisual Speech Segmentation. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00052

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-Reading Aids Word Recognition Most in Moderate Noise: A Bayesian Explanation Using High-Dimensional Feature Space. *PLoS ONE*, *4*(3), e4638. https://doi.org/10.1371/journal.pone.0004638

Macdonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics*, *24*(3), 253–257. https://doi.org/10.3758/BF03206096

Madipakkam, A. R., Rothkirch, M., Dziobek, I., & Sterzer, P. (2017). Unconscious avoidance of eye contact in autism spectrum disorder. *Scientific Reports*, *7*(1), 13378. https://doi.org/10.1038/s41598-017-13945-5

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*(9), 2581–2586. https://doi.org/10.1007/s00221-015-4324-7

Manuel, S. Y. (1987). Output constraints and cross-language differences in coarticulation. *The Journal of the Acoustical Society of America*, *82*(S1), S115–S115. https://doi.org/10.1121/1.2024600

Manuel, S. Y. (1990). The role of contrast in limiting vowel-to-vowel coarticulation in different languages. *The Journal of the Acoustical Society of America*, *88*(3), 1286–1298. https://doi.org/10.1121/1.399705

Marco, E. J., Hinkley, L. B. N., Hill, S. S., & Nagarajan, S. S. (2011). Sensory Processing in Autism: A Review of Neurophysiologic Findings: *Pediatric Research*, *69*(5 Part 2), 48R-54R. https://doi.org/10.1203/PDR.0b013e3182130c54

Marian, V., Hayakawa, S., Lam, T. Q., & Schroeder, S. R. (2018). Language Experience Changes Audiovisual Perception. *Brain Sciences*, *8*(5), Article 5. https://doi.org/10.3390/brainsci8050085

Massaro, D. W. (1984). *Children's Perception of Visual and Auditory Speech*.

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. https://doi.org/10.1016/0022-0965(86)90053-6

Maurer, D., Grand, R. L., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences*, *6*(6), 255–260. https://doi.org/10.1016/S1364-6613(02)01903-4

McBride-Chang, C., Bialystok, E., Chong, K. K. Y., & Li, Y. (2004). Levels of phonological awareness in three cultures. *Journal of Experimental Child Psychology*, *89*(2), 93–111. https://doi.org/10.1016/j.jecp.2004.05.001

McGovern, D. P., Roudaia, E., Stapleton, J., McGinnity, T. M., & Newell, F. N. (2014). The sound-induced flash illusion reveals dissociable age-related effects in multisensory integration. *Frontiers in Aging Neuroscience*, *6*. https://doi.org/10.3389/fnagi.2014.00250

McGurk, H., & Macdonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. https://doi.org/10.1038/264746a0

Meltzoff, A. N., & Kuhl, P. K. (1994). Faces and speech: Intermodal processing of biologically relevant signals in infants and adults. In *The Development of Intersensory Perception* (pp. 335–369). Psychology Press.

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*(3), 793–804. https://doi.org/10.1044/2014_JSLHR-S-12-0395

Mishra, J., Martinez, A., & Hillyard, S. A. (2008). Cortical processes underlying sound-induced flash fusion. *Brain Research*, *1242*, 102–115. https://doi.org/10.1016/j.brainres.2008.05.023

Mishra, J., Martinez, A., Sejnowski, T. J., & Hillyard, S. A. (2007). Early Cross-Modal Interactions in Auditory and Visual Cortex Underlie a Sound-Induced Visual Illusion. *Journal of Neuroscience*, *27*(15), 4120–4131. https://doi.org/10.1523/JNEUROSCI.4912-06.2007

Miyamoto, Y., Yoshikawa, S., & Kitayama, S. (2011). Feature and Configuration in Face Processing: Japanese Are More Configural Than Americans. *Cognitive Science*, *35*(3), 563–574. https://doi.org/10.1111/j.1551-6709.2010.01163.x

Mohammed, T., Campbell, R., Macsweeney, M., Barry, F., & Coleman, M. (2006). Speechreading and its association with reading among deaf, hearing and dyslexic individuals. *Clinical Linguistics & Phonetics*, *20*(7–8), 621–630. https://doi.org/10.1080/02699200500266745

Nardini, M., Jones, P., Bedford, R., & Braddick, O. (2008). Development of Cue Integration in Human Navigation. *Current Biology*, *18*(9), 689–693. https://doi.org/10.1016/j.cub.2008.04.021

National Research Council (US) Committee on Disability Determination for Individuals with Hearing Impairments. (2004). Basics of Sound, the Ear, and Hearing. In *Hearing Loss: Determining Eligibility for Social Security Benefits*. National Academies Press.

Osterling, J., & Dawson, G. (1994). Early recognition of children with autism: A study of first birthday home videotapes. *Journal of Autism and Developmental Disorders*, *24*(3), 247–257. https://doi.org/10.1007/BF02172225

Otake, T. (2015). 12 Mora and mora-timing. In H. Kubozono (Ed.), *Handbook of Japanese Phonetics and Phonology* (pp. 493–524). De Gruyter. https://doi.org/10.1515/9781614511984.493

Paré, M., Richler, R. C., ten Hove, M., & Munhall, K. G. (2003). Gaze behavior in audiovisual speech perception: The influence of ocular fixations on the McGurk effect. *Perception & Psychophysics*, *65*(4), 553–567. https://doi.org/10.3758/BF03194582

Parker, J. L., & Robinson, C. W. (2018). Changes in multisensory integration across the life span. *Psychology and Aging*, *33*(3), 545–558. https://doi.org/10.1037/PAG0000244

Pascalis, O., de Viviés, X. de M., Anzures, G., Quinn, P. C., Slater, A. M., Tanaka, J. W., & Lee, K. (2011). Development of Face Processing. *Wiley Interdisciplinary Reviews. Cognitive Science*, *2*(6), 666–675. https://doi.org/10.1002/wcs.146

Pascalis, O., Fort, M., & Quinn, P. C. (2020). Development of face processing: Are there critical or sensitive periods? *Current Opinion in Behavioral Sciences*, *36*, 7–12. https://doi.org/10.1016/j.cobeha.2020.05.005

Pascalis, O., Loevenbruck, H., Quinn, P. C., Kandel, S., Tanaka, J. W., & Lee, K. (2014). On the Links Among Face Processing, Language Processing, and Narrowing During Development. *Child Development Perspectives*, *8*(2), 65–70. https://doi.org/10.1111/cdep.12064

Peng, G. (2006). Temporal and Tonal Aspects of Chinese Syllables: A Corpus-Based Comparative Study of Mandarin and Cantonese. *Journal of Chinese Linguistics*, *34*(1), 134–154.

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Rennig, J., & Beauchamp, M. S. (2018). Free viewing of talking faces reveals mouth and eye preferring regions of the human superior temporal sulcus. *NeuroImage*, *183*, 25–36. https://doi.org/10.1016/j.neuroimage.2018.08.008

Rennig, J., Wegner-Clemens, K., & Beauchamp, M. S. (2020). Face viewing behavior predicts multisensory gain during speech perception. *Psychonomic Bulletin & Review*, *27*(1), 70–77. https://doi.org/10.3758/s13423-019-01665-y

Robinson, C. W., & Sloutsky, V. M. (2004). Auditory Dominance and Its Change in the Course of Development. *Child Development*, *75*(5), 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*(1), 135–141. https://doi.org/10.1002/WCS.12

Rong, Y., Weng, Y., Chen, F., & Peng, G. (2023). Categorical perception of Mandarin lexical tones in language-delayed autistic children. *Autism*, *27*(5), 1426-1437. https://doi.org/10.1177/13623613221138687

Ruff, H. A., & Lawson, K. R. (1990). Development of sustained, focused attention in young children during free play. *Developmental Psychology*, *26*(1), 85–93. https://doi.org/10.1037/0012-1649.26.1.85

Saalasti, S., Kätsyri, J., Tiippana, K., Laine-Hernandez, M., Von Wendt, L., & Sams, M. (2012). Audiovisual speech perception and eye gaze behavior of adults with asperger syndrome. *Journal of Autism and Developmental Disorders*, *42*(8), 1606–1615. https://doi.org/10.1007/s10803-011-1400-0

Sato, M., Troille, E., Ménard, L., Cathiard, M.-A., & Gracco, V. (2013). Silent articulation modulates auditory and audiovisual speech perception. *Experimental Brain Research*, *227*(2), 275–288. https://doi.org/10.1007/s00221-013-3510-8

Sekiyama, K. (1994). Differences in auditory-visual speech perception between Japanese and Americans: McGurk effect as a function of incompatibility. *Journal of the Acoustical Society of Japan (E)*, *15*(3), 143–158. https://doi.org/10.1250/ast.15.143

Sekiyama, K. (1997). Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. In *Perception & Psychophysics* (Vol. 59, Issue 1, pp. 73–80).

Sekiyama, K., & Burnham, D. (2008). Impact of language on development of auditory-visual speech perception. *Developmental Science*, *11*(2), 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., & Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics*, *21*(4), 427–444. https://doi.org/10.1016/s0095-4470(19)30229-3

Shams, L., Iwaki, S., Chawla, A., & Bhattacharya, J. (2005). Early modulation of visual cortex by sound: An MEG study. *Neuroscience Letters*, *378*(2), 76–81. https://doi.org/10.1016/j.neulet.2004.12.035

Shams, L., Kamitani, Y., & Shimojo, S. (2000). What you see is what you hear. *Nature*, *408*(6814), 788–788. https://doi.org/10.1038/35048669

Shams, L., Kamitani, Y., Thompson, S., & Shimojo, S. (2001). Sound alters visual evoked potentials in humans. *NeuroReport*, *12*(17), 3849.

Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2005). Listening to talking faces: Motor cortical activation during speech perception. *NeuroImage*, *25*(1), 76–89. https://doi.org/10.1016/j.neuroimage.2004.11.006

Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing Lips and Seeing Voices: How Cortical Areas Supporting Speech Production Mediate Audiovisual Speech Perception. *Cerebral Cortex*, *17*(10), 2387–2399. https://doi.org/10.1093/cercor/bhl147

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

Sommers, M. S., & Phelps, D. (2016). Listening effort in younger and older adults: A comparison of auditory-only and auditory-visual presentations. *Ear and Hearing*, *37*, 62S-68S. https://doi.org/10.1097/AUD.0000000000000322

Soto-Faraco, S., & Alsius, A. (2009). Deconstructing the McGurk–MacDonald illusion. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(2), 580–587. https://doi.org/10.1037/a0013483

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 2021 82:7*, *82*(7), 3544–3557. https://doi.org/10.3758/S13414-020-02042-X

Stevenson, R. A., Baum, S. H., Segers, M., Ferber, S., Barense, M. D., & Wallace, M. T. (2017). Multisensory speech perception in autism spectrum disorder: From phoneme to whole-word perception: Multisensory speech in noise perception in autism. *Autism Research*, *10*(7), 1280–1290. https://doi.org/10.1002/aur.1776

Stevenson, R. A., Segers, M., Ferber, S., Barense, M. D., Camarata, S., & Wallace, M. T. (2016). Keeping time in the brain: Autism spectrum disorder and audiovisual temporal processing. *Autism Research*, *9*(7), 720–738. https://doi.org/10.1002/aur.1566

Stevenson, R. A., Segers, M., Ncube, B. L., Black, K. R., Bebko, J. M., Ferber, S., & Barense, M. D. (2018). The cascading influence of multisensory processing on speech perception in autism. *Autism*, *22*(5), 609–624. https://doi.org/10.1177/1362361317704413

Stevenson, R. A., Siemann, J. K., Schneider, B. C., Eberly, H. E., Woynaroski, T. G., Camarata, S. M., & Wallace, M. T. (2014a). *Brief Communications Multisensory Temporal Integration in Autism Spectrum Disorders*. *37232*, 7. https://doi.org/10.1523/JNEUROSCI.3615-13.2014

Stevenson, R. A., Siemann, J. K., Woynaroski, T. G., Schneider, B. C., Eberly, H. E., Camarata, S. M., & Wallace, M. T. (2014b). Brief report: Arrested development of audiovisual speech perception in autism spectrum disorders. *Journal of Autism and*

Developmental Disorders, *44*(6), 1470–1477. https://doi.org/10.1007/s10803-013-1992-7

Stewart, C. R., Sanchez, S. S., Grenesko, E. L., Brown, C. M., Chen, C. P., Keehn, B., Velasquez, F., Lincoln, A. J., & Müller, R.-A. (2016). Sensory Symptoms and Processing of Nonverbal Auditory and Visual Stimuli in Children with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, *46*(5), 1590–1601. https://doi.org/10.1007/s10803-015-2367-z

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. https://doi.org/10.1121/1.1907309

T'sou B, Lee T, Tung P, Man Y, Chan A, To CK, Chan Y. Hong Kong Cantonese oral language assessment scale. Hong Kong: City University of Hong Kong. 2006.

Taitelbaum-Swead, R., & Fostick, L. (2016). Auditory and visual information in speech perception: A developmental perspective. *Clinical Linguistics & Phonetics*, *30*(7), 531–545. https://doi.org/10.3109/02699206.2016.1151938

Tanaka, J. W., & Sung, A. (2016). The "Eye Avoidance" Hypothesis of Autism Face Processing. *Journal of Autism and Developmental Disorders*, *46*(5), 1538–1552. https://doi.org/10.1007/s10803-013-1976-7

Tremblay, C., Champoux, F., Voss, P., Bacon, B. A., Lepore, F., & Théoret, H. (2007). *Speech and Non-Speech Audio-Visual Illusions: A Developmental Study*. https://doi.org/10.1371/journal.pone.0000742

Tsujimura, N. (Ed.). (1999). *The handbook of Japanese linguistics*. Blackwell Publishers.

Ujiie, Y., Asai, T., & Wakabayashi, A. (2015). The relationship between level of autistic traits and local bias in the context of the McGurk effect. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00891

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. https://doi.org/10.1073/pnas.0408949102

Vanderbilt Assessment Scale (Parent Informant, American Academy of Pediatrics and National Initiative for Children's Healthcare Quality, 2002)

Vatakis, A., & Spence, C. (2006). Audiovisual synchrony perception for music, speech, and object actions. *Brain Research*, *1111*(1), 134–142. https://doi.org/10.1016/j.brainres.2006.05.078

Vecera, S. P., & Johnson, M. H. (1995). Gaze detection and the cortical processing of faces: Evidence from infants and adults. *Visual Cognition*, *2*(1), 59–87. https://doi.org/10.1080/13506289508401722

Wang, Y., Thomas, J., Weissgerber, S. C., Kazemini, S., Ul-Haq, I., & Quadflieg, S. (2015). The Headscarf Effect Revisited: Further Evidence for a Culture-Based Internal Face Processing Advantage. *Perception*, *44*(3), 328–336. https://doi.org/10.1068/p7940

Weng, Y., & Peng, G. (2023). The development of audiovisual speech perception in Mandarin-speaking children. In: R. Skarnitzl & J. Volín (Eds.) *Proceedings of the 20th International Congress of Phonetic Sciences – ICPhS 2023* (pp. 4155-4159). International Phonetic Association.

Weng, Y., Rong, Y., & Peng, G. (2024). The development of audiovisual speech perception in Mandarin-speaking children: Evidence from the McGurk paradigm. *Child Development*, *95*(3), 750–765. https://doi.org/10.1111/cdev.14022

Wing, L., National Autistic Society, & University of Kent (Eds.). (1988). *Aspects of autism: Biological research; proceedings of a conference held at the University of Kent, 18 - 20 September, 1987*. Gaskell.

Woynaroski, T. G., Kwakye, L. D., Foss-Feig, J. H., Stevenson, R. A., Stone, W. L., & Wallace, M. T. (2013). Multisensory Speech Perception in Children with Autism

Spectrum Disorders. *Journal of Autism and Developmental Disorders*, *43*(12), 2891. https://doi.org/10.1007/S10803-013-1836-5

Yamamoto, H. W., Kawahara, M., & Tanaka, A. (2019). The Development of Eye Gaze Patterns during Audiovisual Perception of Affective and Phonetic Information. *The 15th International Conference on Auditory-Visual Speech Processing*, 27–32. https://doi.org/10.21437/AVSP.2019-6

Yi, L., Fan, Y., Quinn, P. C., Feng, C., Huang, D., Li, J., Mao, G., & Lee, K. (2013). Abnormality in face scanning by children with autism spectrum disorder is limited to the eye region: Evidence from multi-method analyses of eye tracking data. *Journal of Vision*, *13*(10), 5–5. https://doi.org/10.1167/13.10.5

Yip, M. (2002). *Tone*. Cambridge University Press. https://doi.org/10.1017/CBO9781139164559

Zatorre, R. J., & Belin, P. (2001). Spectral and Temporal Processing in Human Auditory Cortex. *Cerebral Cortex*, *11*(10), 946–953. https://doi.org/10.1093/cercor/11.10.946

Zee, E. (1991). Chinese (Hong Kong Cantonese). *Journal of the International Phonetic Association*, *21*(1), 46–48. https://doi.org/10.1017/S0025100300006058

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2011). Cognitive Load During Speech Perception in Noise: The Influence of Age, Hearing Loss, and Cognition on the Pupil Response. *Ear & Hearing*, *32*(4), 498–510. https://doi.org/10.1097/AUD.0b013e31820512bb

Zhang, J., Meng, Y., He, J., Xiang, Y., Wu, C., Wang, S., & Yuan, Z. (2019). McGurk Effect by Individuals with Autism Spectrum Disorder and Typically Developing Controls: A Systematic Review and Meta-analysis. *Journal of Autism and Developmental Disorders*, *49*(1), 34–43. https://doi.org/10.1007/s10803-018-3680-0

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