A Cross-linguistic Examination on the McGurk Effect in Different Developmental States

Zhou Yang

(Student Number: 6701159)

Research Master's Thesis in Linguistics

Utrecht University

Supervisor: Prof. dr. René Kager

Second reader: dr. Anne-France Pinget

Table of Contents

Abstract

McGurk and MacDonald (1976) found the visual illusion in audiovisual speech perception, called “the McGurk effect”, where a face articulating /ga/ was presented synchronously with an auditory /ba/ syllable and subjects reported that they had perceived it as the syllable /da/, which compromised the conflicting visual and auditory cues. Although the visual influence in audiovisual speech perception was robustly demonstrated in several aspects during the last four decades, the consensus regarding the McGurk effect has not been reached in two aspects: (1) why does the magnitude of the McGurk effect differ across languages; (2) why does the development of audiovisual speech perception represent a U- shaped pattern, i.e., the visual influence is strong in infants, adolescents and adults while weak in children especially 4- to 6-year-old preschool children. The current thesis, therefore, focusing on these two aspects, provides two hypotheses to tackle each aspect respectively.

The first hypothesis, called the visual speech intelligibility hypothesis, suggests that speakers native to languages with a higher degree of visual intelligibility adopt more visual information in their audiovisual speech perception and therefore demonstrate a stronger McGurk effect. The second hypothesis suggests that early literacy training, by accelerating the development of phonological awareness, facilitates children to adopt phonetic visual information in audiovisual speech perception and this acceleration is related to the orthography. The two hypotheses, with the first targeting on cross-linguistic differences and the second on differences across developmental states, collectively provides a holistic framework for the cross-linguistic differences on the developmental trajectories of audiovisual speech perception from childhood to adulthood.

Acknowledgement

I wish to express my greatest gratitude to my supervisor René Kager for his attentive guidance and detailed feedbacks, for his inspiring questions and insightful comments, and for his tolerance and patience to my naiveness. During my two-year study at Utrecht University, René's dedication, passion and energy towards linguistics research always encouraged, inspired and motivated me. I cannot express how grateful and fortunate I am to meet René and have him as my supervisor.

I want to thank Anne-France Pinget for accepting the invitation to review my thesis as the second reader. I would also like to thank her for teaching the phonetics class in the foundational course, which sparked my interest in speech perception and the McGurk effect.

I want to thank all brilliant linguists at Utrecht University, who generously took me on their spaceships to travel in the vast universe of linguistics. It was my great honor to attend their classes and learn from their sharp minds, innovative thoughts, and consistent diligence. They have been and will always be the brightest stars shining through my way of linguistics exploration.

I would like to thank my peer group, "Masters of Linguistics", for always being supportive to each other. Thank you, Agnė and Elena, for our invaluable sisterhood without boundaries of languages, races, nationalities, and cultures. Being friend with Agnė and Elena is the most memorable experience of my life. "Proost" to our legendary team AEZ!

My last but not least thank goes to my parents, who always unconditionally support my study, even though it is an elusive subject to them. Their strong confidence in me encourages me all the time.

Chapter 1. Introduction

Speech is not merely acoustic sounds. The process of formulating a speech involves the cooperative movements of various articulators or speech organs, such as jaws, teeth, and lips. In face-to-face multimodal communication, the speaker's auditory speech is heard and at the same time the speaker's articulatory movements are seen to facilitate the perceiver's language comprehension. Thus, the integration of the auditory and visual stimuli became important in the optimal understanding of speech (Brancazio, Miller, & Paré, 2003). Studies found that participants' responses to audiovisual stimuli are significantly faster than auditory-only counterparts (Holle et al., 2008; Kelly et al., 2010; Nagels et al., 2015; Wu & Coulson, 2015). These studies showed that the processing of audio-visual language stimuli is bimodal, and the fact that the visual modality facilitates speech perception is called the "audiovisual benefit effect" (e.g., Bernstein et al., 2004; Jesse & Janse, 2012; MacLeod & Summerfield, 1987; Sumby & Pollack, 1954). The process to synthesize different modalities simultaneously and in synchrony is called 'multisensory integration'. The multisensory integration is involved in the audiovisual speech perception, where the auditory and visual information is integrated simultaneously. The integration is also an involuntary process happening automatically. The audiovisual benefit effect happens when the auditory information and visual information are synchronized, i.e., the visual cue is congruent with the auditory counterpart. What if the visual information is inconsistent with the auditory correspondence? Do perceivers still integrate information from two modalities, or do they disregard the deceptive visual information and still perceive the audiovisual conflicting stimuli same as its auditory-only counterparts? Besides the audiovisual benefit effect, the other type of audiovisual effects is called 'the McGurk effect' (McGurk & MacDonald, 1976), where a face articulating [ga] was presented together with an auditory /ba/ syllable and participants reported that they had perceived it as the syllable /da/, which compromised the two conflicting cues. The McGurk effect and the audiovisual benefit effect altogether make up two main strands of research on

audiovisual speech perception. Although the current thesis provides some discussion about audiovisual benefit effect, the focus of the study is on the McGurk effect.

A large body of research on audiovisual speech perception including both the McGurk effect and audiovisual benefit effect suggests that the magnitude of visual influences on audiovisual speech perception varies not only across languages but also across developmental states (Burnham & Lau, 1998; Chen & Hazan, 2009; Chen & Massaro, 2004; de Gelder & Vroomen, 1992; de Gelder et al., 1995; Dupont et al., 2005; Hayashi & Sekiyama, 1998; Lalonde & Werner, 2019; Lalonde & McCreery, 2020; Magnotti et al., 2015; Massaro et al., 1995; Sekiyama, 1997; Sekiyama & Burnham, 2004, 2008; Sekiyama & Tohkura, 1991, 1993; Tremblay et al., 2007). However, the reasons accounting for cross-linguistic differences and developmental differences in audiovisual speech perception are not clear yet. Therefore, the current thesis aims to provide underlying reasons responsible for different degrees of visual influences in audiovisual speech perception across developmental states and across languages. Two main research questions are addressed:

RQ1: Why does the magnitude of the McGurk effect differ across languages?

RQ2: Why does the developmental trajectory of the McGurk effect demonstrate a U-shaped pattern, i.e., the degree of visual impacts is smallest in children while salient in infants and adults?

The present thesis consists of six chapters to account for the two research questions. Chapter 2 provides a review of studies on audiovisual speech perception and discusses insufficiencies in the experimental paradigm. It starts with the cross-linguistic studies of the McGurk effect in adults and different aspects of the research design in adult studies are discussed in each subsection. The second part of Chapter 2 examines previous research on audiovisual speech perception in infants and children, and furthermore discusses previous accounts for the U-shaped developmental trajectory. This chapter aims to summarize earlier findings of audiovisual speech perception across the life span and inspect hypotheses proposed in previous studies. With those findings,

the present thesis in subsequent chapters pinpoints drawbacks of previous experiments and supplies a new framework to account for the audiovisual speech perception across languages and across developmental states.

Chapter 3 introduces the two hypotheses suggested in the current thesis. The first, the visual speech intelligibility hypothesis, is to elucidate empirical results of the magnitude of the McGurk effect in different languages in adults. The purpose of the second hypothesis is resolving the U-shaped pattern in the development of audiovisual speech perception. The U-shaped pattern, however, is not universal across languages due to the binding of the visual speech intelligibility hypothesis and therefore the two hypotheses altogether account for the different degree of influences of visual information in audiovisual speech perception in different languages and in different developmental states. In Chapter 4 and Chapter 5, two experiments are proposed respectively to justify the two hypotheses elaborated in Chapter 3.

Chapter 4 proposes a design of an adult multimodal speech perception experiment with speakers of six language backgrounds as participants. The six languages, including American English, Mandarin Chinese, Dutch, Cantonese, Japanese, and Russian, were selected mainly because of their various degrees of visual intelligibility. Different aspects of flaws in the research methods as discussed in Chapter 2 will be carefully avoided. The main purpose of this chapter is to justify the first hypothesis of the current thesis, and two peripheral research aims that can be achieved are: (1) comparing results of American English, Mandarin Chinese, Dutch, Cantonese and Japanese with previous studies since those five languages were most frequently studied; (2) providing McGurk experimental results in Russian, a language that has barely been examined in audiovisual speech perception research.

Chapter 5 aims to resolve the following: how is early literacy training at school important for first-grade children to integrate visual cues in speech perception? The hypothesis is that phonemic awareness is crucial for adopting visual phonetic cues in

audiovisual speech perception. Various languages demonstrating different developmental trajectory of phonemic awareness should be included to justify the hypothesis. Dutch and Cantonese children are included in the experiment because the development of phonemic awareness is fast in Dutch children and slow in Cantonese children. More importantly, Dutch and Cantonese have the same degree of visual intelligibility, which can control the possible factor of the first hypothesis as suggested in Chapter 3 and Chapter 4.

Chapter 6 contains a general discussion of the current thesis. It evaluates whether the hypotheses and experiments proposed in this study effectively address the two research questions raised above. Furthermore, limitations in the current thesis are suggested, which provide implicatures for future studies.

Chapter 2. Literature review

McGurk and MacDonald (1976) discovered 'the McGurk effect', where a face articulating [ga] was presented together with an auditory /ba/ syllable and participants reported that they had perceived it as the syllable /da/, which compromised the two conflicting cues. The response of /da/ to visual [ga] and auditory /ba/ is called the fused perception, which is mostly studied in previous research (Massaro, 1987). Besides the fused perception, the McGurk effect contains the other perception called the combined perception (Green & Gerdeman, 1995). The combined perception happens when the stimulus is visual /ba/ with auditory /ga/ and the response is /g͡ba/.

Studies have found that the McGurk effect is robust in several aspects. The McGurk effect is demonstrated not only on consonants, but also vowels (Traunmüller & Öhrström, 2007; Valkenier et al., 2012). Traunmüller and Öhrström (2007) found that Swedish high unrounded front vowel /e/ was perceived as the high rounded front vowel /ø/ when the auditory syllable /geg/ was paired concurrently with visual /gyg/. Valkenier et al. (2012) found the McGurk effect when audiovisual incongruent vowels in Dutch were presented to Dutch speakers, e.g., the unrounded high vowel /I/ was perceived when the auditory rounded mid-high vowel /Y/ was synchronized with visual unrounded mid-high vowel /e/. Besides segmental features, studies (Han et al., 2020; Wang, 2018) found the McGurk effect in lexical tones, which are suprasegmental, although the effect of visual information is marginal. Wang (2018) created stimuli where the tonal auditory and tonal visual cue are incongruent. Wang (2018) found that lexical tone perception benefits from adding visual information of corresponding articulatory movement and the duration perception of lexical tones is changed by incongruent visual information, by using both the response time paradigm and event-evoked potential (ERP) methodology. Another demonstration of the robustness of the McGurk effect is that it is an automatic and involuntary process, which is not affected even when participants were told that visual and auditory input are incongruent (Wang, 2018). The robustness of the McGurk effect has been shown in different conditions as

well, e.g., with modifications on the auditory or visual cues (Campbell, 2008; Massaro & Jesse, 2007; Summerfield, 1987). Although the McGurk effect has been found robust in those aspects mentioned above, the consensus has not been reached in another two aspects, i.e., whether the McGurk effect is robust (1) across developmental stages and (2) across languages. Targeting on these two aspects, two methods were suggested by Sekiyama and Burnham (2004) to study the development of the audio-visual speech processing, one is the 'differential experience method' and the other is the 'ontogenetic method'.

The 'differential experience method' is to examine the McGurk effect across languages and the 'ontogenetic method' to across developmental stages. The differential experience method investigates "individuals of the same age brough up in functionally different environments on a common task in order to investigate the effect that the type of exposure or experience may have on the development" and the ontogenetic methods involves "comparing the abilities of individuals of different ages brought up in functionally identical environments on a common task in order to investigate the amount of experience or maturation on development" (Sekiyama & Burnham, 2004, p. 1). The current thesis proposes two experiments in Chapter 4 and 5, one adopting the differential experience method and the other the ontogenetic method. Combining the two methods, this thesis aims to explore the cross-linguistic differences in the developmental trajectory of audio-visual speech perception. The following two subsections first discuss earlier cross-linguistic studies of the McGurk effect in adults and then offer a review on infant and child studies of the McGurk effect.

2.1 Cross-linguistic studies of the McGurk effect in adults

Since the McGurk effect was firstly reported in English (McGurk & MacDonald, 1976), it was reduplicated in abundant studies and the robust McGurk effect was consistently demonstrated in English speakers (for a review, see Summerfield, 1987). Besides English, the McGurk effect has also been examined in other languages, such

as Mandarin Chinese (Chen & Hazan, 2009; Chen & Massaro, 2004; Hayashi & Sekiyama, 1998; Magnotti et al., 2015; Sekiyama, 1997), Japanese (Sekiyama & Burnham, 2008; Sekiyama & Tohkura, 1991, 1993), Cantonese (Burnham & Lau, 1998), and Dutch (de Gelder & Vroomen, 1992; de Gelder et al., 1995; Massaro et al., 1995). A series of Sekiyama's studies (Sekiyama & Burnham, 2008; Sekiyama & Tohkura, 1991, 1993) consistently found that the magnitude of the McGurk effect is smaller in Japanese speakers than in American English speakers. Only the study by Sekiyama (1997) demonstrated that Chinese Mandarin native speakers present a significantly weaker magnitude of the McGurk effect than American English subjects. In contrast, studies developed by Chen and Hazan (2009) and Magnotti et al. (2015) showed that American English speakers and Mandarin Chinese speakers demonstrated no difference on the degree of the McGurk effect. Burnham and Lau (1998) found that Cantonese speakers demonstrated a significantly weaker McGurk effect than Australian English speakers. However, when comparing Cantonese speakers with Dutch speakers, de Gelder et al. (1995) found that Cantonese participants were more influenced by vision (larger McGurk effect) than the Dutch. In summary, abundant studies examined the McGurk effect in different languages with inconsistent results across studies. Furthermore, three hypotheses were proposed to account for the different magnitude of the McGurk effect across languages, but each hypothesis has its limitation as elaborated below. The subsequent texts first compare research methodologies among earlier studies, whose differences may result in the inconsistent findings across studies. After that, the three hypotheses suggested in previous research will be discussed.

2.1.1 Research paradigms

To our knowledge, three main types of behavioral paradigms were adopted in cross-linguistic studies on the McGurk effect. The first is the written-down task, where participants were instructed to write down in Roman letters what they thought they had perceived (e.g., Sekiyama & Tohkura, 1991). Another paradigm is the verbal production task, that is, subjects repeated verbally aloud each time they were presented with one

stimulus (de Gelder et al., 1995; Magnotti et al., 2015). Last, the forced-choice procedure was adopted in previous research, in which participants needed to press one of the response buttons (Burham & Lau, 1998) or click one of the options on the computer (Chen & Hazan, 2009). Furthermore, the study by Mallick et al. (2015, as cited in Marques et al., 2016) compared different methodologies and suggested that the forced-choice paradigm is better to register responses, where the options are restricted and "the response is oriented specifically to one of the options, rather than being open to any possible response" (p. 1116) as in open-choice procedures. The forced-choice paradigm increased by 27 percent of the visually biased responses compared to the open-choice paradigm (Mallick et al., 2015).

2.1.2 Stimuli in previous research

Besides research paradigms, the stimuli used in previous studies are diversified as well. In the original McGurk effect study, McGurk and MacDonald (1976) designed stimuli from syllables of English /ba/, /ga/, /pa/ and /ka/. In a series of Sekiyama's studies (Sekiyama, 1997; Sekiyama & Tohkura, 1991, 1993), 10 syllables were adopted including /ba/, /da/, /ga/, /ka/, /ma/, /na/, /pa/, /ra/, /ta/, and /wa/ in English and in Japanese to create 100 stimuli (10 audio x 10 visual) in each language, and both 100 English stimuli and 100 Japanese stimuli were presented to both Japanese and English participants. In the study by de Gelder et al. (1995), the audio-visual stimuli were composed of six different audio-visual combinations: auditory /apa, aba, ata, ada, ama, ana/ were combined with visual /ata, ada, apa, aba, ana, ama/ respectively to test Dutch and Cantonese subjects. However, de Gelder et al. (1995) did not report which language was used in the auditory stimuli and the native language of the performer in the stimulus video. Burnham and Lau (1998) adopted /ba/ and /ga/ with different tones in Cantonese and Thai as stimuli to test Cantonese native participants and Australian-English native participants. In Chen and Hazan (2009), there were three incongruent audiovisual stimuli: (1) auditory-ba/visual-ga, (2) auditory-da/visual-ba, and (3) auditory-ga/visual-ba, produced by two Chinese and two English native speakers (one male and one female

in each language), and participants consisted of both Chinese and English speakers. Magnotti et al. (2015) included English /ba/, /ga/, /pa/ and /ka/, like in McGurk and MacDonald (1976). In their study (Magnotti et al., 2015), all the nine performers in the stimuli were Americans and the faces were non-Asian (e.g., Caucasian), whilst participants were one group of Chinese and one group of Americans. In summary, syllables consisting of the open vowel /a/ and one of bilabial, alveolar, and velar stops are most frequently adopted as stimuli in previous McGurk studies. In addition, the influence of faces of different ethnic groups was not considered in previous research. However, that is important in audiovisual speech perception since recognizing faces from the own ethnic group is easier than faces from other racial groups, known as the other-race effect (for a review, see Meissner & Brigham, 2001). Moreover, a mismatch may exist between the native language in the stimuli and the native language of participants, which results in the "non-native speaker" effect, as illustrated in the next subsection.

2.1.3 The non-native speaker effect

The non-native speaker effect is another factor that has been investigated in cross-linguistic studies on the McGurk effect. The hypothesis is that people use more visual information (as shown by a greater visual effect) when the stimuli are non-native (cited from Chen & Hazan, 2009, p. 864). It suggests that when listening to non-native speech, it is more difficult for respondents to process the auditory speech compared to their native language and therefore they borrow more information from the visual modality when perceiving a foreign language. In Sekiyama and Tohkura (1993) and Sekiyama (1994), they tested Japanese subjects with Japanese syllables (spoken by Japanese performers), another group of Japanese subjects with 10 corresponding English syllables (same as Japanese syllables but spoken by American English performers), American English subjects with Japanese syllables, and another group of American English subjects with corresponding English syllables. They found that Japanese subjects were much less prone to visual biasing effects than were Americans for the

same Japanese syllables, but for English syllables the difference between Japanese and Americans was not significant (Sekiyama & Tohkura, 1993). Sekiyama and Tohkura (1993) suggested that American English speakers automatically integrate visual and auditory cues in their native language and the integration is strengthened in a non-native language while Japanese speakers incorporate visual cues much less in their native language. Sekiyama and Tohkura (1993) referred to the Japanese style of audiovisual speech processing as "vision-independent processing" and "this vision-independent processing breaks down for non-native syllables, where they were subject to visual biasing effects" (p. 441). Chen and Hazan (2009) found that English perceivers used significantly more visual information when the stimuli were Chinese, but Chinese native perceivers did not show such a difference between the stimuli in English and in Chinese. They suggested a possible explanation: while none of the English participants had learned Chinese, all the Chinese participants had knowledge of English through the classes they had taken for at least 6 years in school (Chen & Hazan, 2009). Previous studies suggested that the (non-)nativeness is important in audiovisual speech perception. However, they did not disentangle whether the different degree of visual influences is resulted from the non-native auditory speech information (the non-native speaker effect) or from the non-native visual cues (the other-race effect).

2.1.4 The number and age of participants

Apart from methodologies, different syllables of the stimuli and the language background of the performers in the stimuli, the number of participants in previous studies was also different from each other. Magnotti et al. (2015) criticized that the sample sizes in Sekiyama's studies (Sekiyama, 1997; Sekiyama & Tohkura, 1991, 1993) were small, which were 10-20 participants in one language group. In contrast, Magnotti et al. (2015) recruited 162 Mandarin speakers native to China and 145 English speakers native to the USA. Most studies of cross-linguistic differences regarding the McGurk effect have recruited 18 – 24 participants in one language group (e.g., Bunham & Lau, 1998; Chen & Hazan, 2009; de Gelder et al., 1995).

Furthermore, the influence of the age of adult participants was not sufficiently discussed. In Magnotti et al. (2015), the age range of the Chinese group is 14 – 23 while in English it is 18 – 26. The Chinese participants are more teenager-like, but the English participants are all in adulthood. In Chen and Hazan (2009), the participants had a great age range, which varied between 20 to 54. Although the developmental course of the audiovisual speech perception is not fully discovered, Pearl et al. (2009) found that adolescents performed significantly more efficiently in the integration process than adults. Furthermore, Marques et al. (2016) suggested that the capacity to integrate auditory and visual modality is decreasing with aging during adulthood. Therefore, it is crucial to control the variabilities of subjects' age in cross-linguistic studies in adults where the dependent variable is the magnitude of the McGurk effect and the independent variable is the language. In other words, it is important not to have a large range of age in participants.

The following table 1 summarizes previous investigations on the different magnitude of the McGurk effect across languages. Only cross-linguistic studies on the McGurk effect that included discussion about at least two languages were included in this table. Other studies that investigated only one language (e.g., McGurk & MacDonald, 1976) are not listed.

	Paradigm	Syllables in the stimuli	Languages in the stimuli (produced by the performer)	Performers' nationality and faces	Number & Native language of participants	Age of participants
Sekiyama & Tohkura, 1993	written-down task	[ba] [da] [ga] [ka] [ma] [na] [pa] [ra] [ta] [wa]	Japanese, American English	Asian, Caucasian	20 Japanese, 20 American English	20 – 30 (All)
Sekiyama & Tohkura, 1994	written-down task	[ba] [da] [ga] [ka] [ma] [na] [pa] [ra] [ta] [wa]	Japanese, American English	Asian, Caucasian	24 Japanese, 20 American English	20 – 30 (All)
de Gelder et al., 1995	verbal production task	[aba] [ada] [ata] [apa] [ama] [ana]	NA	NA	18 Dutch, 18 Cantonese	NA
Sekiyama, 1997	written-down task	[ba] [da] [ga] [ka] [ma] [na] [pa] [ra] [ta] [wa]	Japanese, American English	Asian, Caucasian	14 Mandarin Chinese	19 – 30
Burnham & Lau, 1998	forced-choice procedure	[ba] [ga]	Thai, Cantonese	Asian	24 Cantonese, 24 Australian English	NA
Chen & Hazan, 2009	forced-choice procedure	[ba] [da] [ga]	Chinese, British English	Asian, Caucasian	22 Mandarin Chinese, 18 British English	20 – 54 (All)
Magnotti et al., 2015	verbal production task	[ba] [ga] [bab] [gag] [pa] [ka]	American English	Caucasian	162 Mandarin Chinese, 145 American English	14 – 23 (Chinese) 18 – 26 (American)

Table 1: Review of previous cross-linguistic studies on the McGurk effect
NA: Not Applicable

2.1.5 Hypotheses in previous research

Three hypotheses were suggested to account for the different size of the McGurk effect across languages. The first one is the tonal hypothesis (Sekiyama et al., 2003): the more tonal the language is, the greater reliance speakers with this language background have on the auditory information, and thus the less McGurk effect the speakers demonstrate. Sekiyama et al. (2003) claimed that in tonal languages, since lexical meanings are modulated by different lexical tones and the tonal difference is not saliently demonstrated on speaker's face, tone-language speakers rely less on the visual information to differentiate lexical meanings with different tones (for a review, see Sekiyama & Burnham, 2004). However, Wang (2018) found the McGurk effect regarding Mandarin lexical tones using the response time paradigm and event-evoked potential (ERP) methodology. Also, although the effect of visual information for tone-language speakers in the tone identification tasks is marginal (Han et al., 2020), the marginal McGurk effect of lexical tones does not necessarily imply that the McGurk effect of consonants is also marginal in tonal languages. In other words, the marginal McGurk effect of lexical tones as found in (Han et al., 2020) does not deny the possibility that there might be strong McGurk effect in consonants and vowels in tonal languages since tone languages may also have wide varieties of consonants and vowels such as Mandarin Chinese.

The second hypothesis (Sekiyama, 1997) to account for the different strength of the McGurk effect between American English and Chinese Mandarin is the face-avoidance tradition, which is a sociocultural convention in Asian countries such as China and Japan, i.e., the listener has to avoid eye contact and prevent seeing speaker's face to show respects especially when the speaker's status is social-economically higher than the listener such as between the teacher and the student. However, this convention is increasingly abandoned by young generations under the influence of globalization (Isei-Jaakkola, 2006). Previous studies that found different degrees of the McGurk

effect among American, Japanese and Mandarin speakers were conducted more than 20 years ago (Burnham & Lau, 1998; Sekiyama, 1997; Sekiyama & Tohkura, 1991, 1993). However, recent studies (Chen & Hazan, 2009; Magnotti et al., 2015) found that Mandarin Chinese speakers have the same degree of the McGurk effect as English speakers, which refuted the face-avoidance hypothesis.

Chen and Hazan (2009) suggested another linguistic account - the richness of the language's phonetic inventory - for the different degrees of the McGurk effect across languages. They (Chen & Hazan, 2009) argued that Chinese Mandarin has 16 vowels, much richer than 5 vowels in Japanese, and thus speechreading is more needed for Chinese Mandarin speakers. However, the suggestion provided by Chen and Hazan (2009) cannot account for the results found in Cantonese and English speakers in the study of Burnham and Lau (1998). Like Chen and Hazan's (2009) study, Burnham and Lau (1998) also included native stimuli with Cantonese tones while their results supported the tonal hypothesis, i.e., English speakers made more visually influenced responses compared to Cantonese speakers.

To summarize, the consensus regarding the McGurk effect across languages has not been achieved, and three accounts were proposed, which are (1) the tonal hypothesis, (2) the face-avoidance convention, and (3) the phonetic inventory. As discussed above, all three accounts have limitations. Therefore, the current study provides a new theoretical framework in chapter 3 to account for empirical findings in previous studies. Before that, studies in infants and children are examined in the remainder of this chapter.

2.2 Audiovisual speech perception in infants and children

Above we have seen McGurk-effect studies in adults in different languages and the magnitude of the McGurk effect differs across languages. However, it still remains unknown at which developmental state speakers native to different languages diverge from each other. Therefore, it is important to examine the ontogenetic development of audiovisual speech perception in a specific language as well as across languages. The

subsequent two subsections discuss previous studies on audiovisual speech perception in infants and children respectively. At the end of this section, based on a summary of findings in infants, children and adults, a discussion will occur of the development of audiovisual speech perception across the lifespan, which represents a U-shaped pattern.

2.2.1 Infant studies

Studies have shown that infants are sensitive to the congruency of the facial articulation and voice at the age of 2.5 months (Dodd, 1979; Burnham & Dodd, 1998). Infants are also able to match the auditory and visual signal as young as two months old. When they were first presented with a silent video displaying two same faces side by side simultaneously, one face with visual articulation of one syllable and the other articulating a different syllable, and later infants were presented with the same video of two faces adding the sound of one of the syllables, infants preferentially look at the side where the visual articulation was matched with the sound (Aldridge et al., 1999; Kuhl & Meltzoff, 1982, 1984; Mackain et al., 1983; Patterson & Werker, 1999, 2002, 2003). Besides the cross-modal matching, Hollich et al. (2005) found the audiovisual benefit effect in 7.5-month-old infants, suggesting that the visual information facilitates infants' auditory speech perception. Studies also found that infants demonstrated the McGurk effect like adults (Burnham & Dodd, 2004; Desjardins & Werker, 1996; Rosenblum et al., 1997). Burnham and Dodd (2004) adopted the habituation-test paradigm, where the experimental group was habituated to a McGurk stimulus (e.g., visual information of a person articulating [ga] dubbed with auditory [ba]), and the control group to an audiovisual congruent stimulus (e.g., audiovisual "ba"). Both groups were later presented with the auditory-only stimuli, [ba], [da], and [ða]. It found that infants in the experimental group in the test phase fixated at the auditory [da] or [ða] significantly longer than auditory [ba], which indicated that participants in the experimental group recognized the McGurk stimulus (auditory [ba] accompanied with visual [ga]) as auditory [da] or [ða] rather than [ba] (Burnham & Dodd, 2004). In short, audiovisual speech perception research on infants found that: (1) infants are sensitive to auditory

and visual synchrony; (2) infants are able to adopt visual cues to facilitate the auditory speech perception; (3) infants demonstrate the McGurk effect.

2.2.2 Child studies

Developmental studies of the audiovisual speech perception consistently found that the visual influence is weaker in children compared to adolescents and adults. Dupont et al. (2005) recruited eight 4 and 5-year-old French Canadian children and ten adults as participants, who were presented with visual-only, auditory-only and audiovisual congruent and incongruent stimuli. Adults were asked to write down the perceived utterances while children were asked to repeat what they thought the speaker had just said (Dupont et al., 2005). Results showed that around 90 percent of children's responses to audiovisual incongruent stimuli were identical with the auditory component of the audiovisual incongruent stimuli, and adults' audio percepts only took up approximately 40 percent of responses, which suggested that adults were much more biased towards the visual information than children (Dupont et al., 2005). Tremblay et al. (2007) recruited participants (French Canadian) with wider range of ages, including three age groups 5-9 years old (11 subjects), 10-14 years old (16 subjects), and 15-19 years old (11 subjects). Participants were asked to repeat the syllable they thought they had heard after being presented with the stimuli, which were (1) unimodal auditory [ba], (2) unimodal auditory [va], (3) bimodal congruent [ba], (4) bimodal congruent [va], and (5) bimodal incongruent auditory /ba/ visual /va/ (Tremblay et al., 2007). Tremblay et al. (2007) found that the 5–9-year-old group had 60 percent of /ba/ responses to the incongruent stimulus auditory /ba/ visual /va/, which was significantly higher than the other two groups (around 20 percent of /ba/ responses), and there was no significant difference between 10-14-year-old group and 15-19-year-old group. The results suggested that the audiovisual speech perception may be adult-like after 10 years old and there may be a change of processing audiovisual speech during the development between 5 to 9 years old. Therefore, more refined age groups should be included in the study to pinpoint the age when children went through the developmental change.

Sekiyama and Burnham (2004, 2008) recruited native speakers of Japanese and native speakers of English as participants, who were divided into four age groups in both language groups, namely 6 years, 8 years, 11 years and adults. Participants were asked to press one of the three buttons for a 'ba', 'da' or 'ga' response accurately and without delay after they were presented with syllables in audiovisual (with congruent or mismatched auditory and visual components), audio-only and video-only trials at various signal-to-noise levels (Sekiyama & Burnham, 2004, 2008). In line with studies of Dupont et al. (2005) and Tremblay et al. (2007), Sekiyama and Burnham (2004, 2008) found that in English speakers, the degree of visual influence on speech perception was low at the age of 6 and increased over the age, and the increase was particularly prominent between 6 and 8 years old. For Japanese participants, the degree of the visual influence in 6-year-old children was the same compared to 6-year-old English children. In contrast, the degree of visual influence remained the same for Japanese speakers over the development (Sekiyama & Burnham, 2004, 2008). In summary, the strength of the McGurk effect was found to be adult-like in children older than 10 years old. English-speaking children and Canadian French-speaking children younger than 10 years old demonstrate a weaker McGurk effect compared to adolescents and adults and the degree of the McGurk effect gradually increases between 4 to 9 years old. This increase might be the most prominent between 6 to 8 years old. However, for Japanese speakers, the strength of the McGurk effect remains intact after 6 years old.

So far, we have seen that the visual influence is strong in audiovisual speech perception in infants and in adults (except Japanese adult speakers) whereas in 6-year-old children regardless of language backgrounds the visual influence is weak. Jerger et al. (2009) suggested a U-shaped pattern for the developmental trajectory of audiovisual speech perception, where the auditory perception in infants, adolescents and adults is strongly influenced by visual information while the influence is much less prominent in children. However, the U-shaped pattern in audiovisual speech perception may not be language universal since it is present in English speakers but not found in Japanese speakers in Sekiyama and Burnham (2004, 2008). Furthermore, the underlying

mechanism remains unclear about why the age of 6 is the critical period for English speakers to change the direction of the trajectory. The next paragraph discusses possible accounts for the U-shaped pattern in the development of audiovisual speech perception.

2.2.3 The U-shaped pattern

One interpretation for the U-shaped pattern is that different methods adopted in previous studies, with different complexity of tasks, require different cognitive burdens between children and infants (Lalonde & Werner, 2021). In infants, experiments measuring looking time indirectly examine participants' sensitivity to the stimuli. However, direct experiments for children explicitly instruct them to give responses that consciously reflect their awareness (e.g., to repeat what they thought they had perceived). Due to the same methodology used between children and adults (e.g., Sekiyama & Burnham, 2008; Tremblay et al., 2007), the degree of the visual influence can be compared across age groups with the percentage of the visually influenced responses. However, the methodology of habituation-test paradigm only measures whether infants demonstrated the McGurk effect or not by comparing the looking time difference, but it cannot determine the percentage of the McGurk responses or the degree of the visual influence on the continuum. To have a consistent methodology among infants, children and adults, studies (Lalonde & Werner, 2019; Lalonde & McCreery, 2020) adapted an audiovisual benefit paradigm in adults and made it compatible with infants and children by using a modified observer-based psychoacoustic procedure in Werner (1995). The test setup and most of the procedural details are the same between infants and adults in three visual conditions (i.e., auditory-only, audiovisual, and onset-offset cue) in Lalonde and Werner (2019). The experiment included three phases in both infant group and adult group: familiarization, training and test (Lalonde & Werner, 2019). Lalonde and Werner (2019) found that in audiovisual speech detection, infants like adults benefited from both the audiovisual and onset-offset cue. By using typical audiovisual speech detection experiments where participants were asked to "repeatedly indicate which of two noise intervals contains

acoustic speech" (Lalonde & Werner, 2021, p. 2), Lalonde and McCreery (2020) observed that 6- to 12-year-old children and adults demonstrate the same degree of audiovisual detection benefit, i.e., both adults and children "detect speech at about a 2 dB lower signal-to-noise ratio in audiovisual conditions than in auditory-only conditions" (Lalonde & Werner, 2021, p. 3). The two studies (Lalonde & Werner, 2019; Lalonde & McCreery, 2020), suggesting that infants, children and adults demonstrated the same degree of visual benefit effect, seem to contradict with the U-shaped trajectories found in other studies. However, these results by Lalonde and colleagues (Lalonde & Werner, 2019; Lalonde & McCreery, 2020) only suggested that the audiovisual benefit in syllable detection tasks is the same across the development but it does not deny that the audiovisual benefit may differ among different age groups in speech recognition or discrimination tasks. Whereas detection requires only basic awareness of the presence of speech, discrimination requires perception of more fine-grained differences between the spectrotemporal properties of speech sounds (Lalonde & Werner, 2019).

Another account for the U-shaped pattern is that the underlying processing mechanism differs between infants and adults, although both groups are strongly influenced by visual cues in audiovisual speech perception. Infants adopt only temporal cues including the onset/offset of speech and amplitude envelope in all types of audiovisual speech tasks whereas adults administer the temporal cues in audiovisual speech detection tasks and use phonetic information in audiovisual speech discrimination tasks (for a review, see Lalonde & Werner, 2021). The use of phonetic cues is a "speech-specific and lexical mechanism" where interlocutors "associate salient visual cues with phonemes, syllables, and words in their native language" (Lalonde & Werner, 2021, p. 6). To determine whether the temporal or phonetic cues are adopted, Baart et al. (2014) used sine-wave replicas of auditory speech (SWS), which preserves the temporal characteristics of speech and the amplitude of the formants that correlate with the visual amplitude envelope (Chandrasekaran et al., 2009; Grant & Seitz, 2000), but phonetic information is removed. The study (Baart et al.,

2014) found that adults matched natural auditory speech very well to visual speech but matched SWS poorly to visual speech, while infants did well in matching both SWS and unprocessed auditory speech to visual speech, suggesting that adults used phonetic cues and infants rely heavily on temporal cues. Hollich et al. (2005) found that when being presented with a female speaker reading a story with infant-directed speech while a competing male reading out an academic paper, 7.5-month-old infants were able to segregate the competing speech stream (male speech) and attend to the target speech stream (female speech) even though the visual speech was replaced by an oscilloscope pattern that kept only the temporal envelope of the speech signal. Lalonde and Werner (2019) designed an audiovisual stimulus where the visual component contained only the visual onset and offset of the auditory syllable to test whether the onset-offset cue is used for detection and discrimination tasks in adults and 6- to 8.5-month-old infants. They found that (1) infants' speech detection and discrimination were both facilitated by the onset-offset cue; (2) adults relied on the onset-offset cue in detection while not in discrimination and the full visual cue benefited adults to a larger extent than infants, and they suggested that 6-month-old infants are mature in using temporal cues but not phonetic cues for audiovisual speech perception and adults relied mostly on phonetic information (Lalonde & Werner, 2019). These studies of audiovisual speech perception in infants imply that the underlying mechanism of the McGurk effect in infants is integrating temporal cues of visual and auditory information (although they are asynchronous) in the McGurk stimuli. More importantly, neurological research suggested that the cortical structures responsible for using visual phonetic/lexical information had limited development during the first year but the brain networks responsible for the synchrony between the auditory and visual speech showed maturation before the age of 6 months (Bushara et al., 2001; Eggermont & Moore, 2012). Some may argue that the aforementioned suggestion that infants are unable to adopt visual phonetic cues contradicts with empirical results (e.g., Teinonen et al., 2008; Weikum et al., 2007) which showed that infants demonstrated visual speech discrimination before 6 months old. However, the audio-visual stimuli used in those

studies (Teinonen et al., 2008; Weikum et al., 2007) are natural speech materials containing both temporal and phonetic information. Therefore, these studies (Teinonen et al., 2008; Weikum et al., 2007) failed to exclude the possibility that infants may benefit only from temporal cues to accomplish visual speech discrimination. In contrast, the finding in (Lalonde & Werner, 2019) that infants' audiovisual benefit effect facilitated by temporal-only cues is the same with that facilitated by temporal and phonetic combined cues suggests that phonetic information may be unimportant in audiovisual speech perception in infants.

In summary, previous studies robustly suggested that the difference of audiovisual speech perception between adults and infants is on the use of phonetic and temporal information. Infants are sensitive to temporal cues but not able to use phonetic information in audiovisual speech perception. The sensitivity to temporal cues may decrease from infancy to adulthood since temporal cues are the only sources for infants to rely on while adults adopt phonetic cues besides temporal cues. Due to the decrease of using temporal cues in audiovisual speech perception from infancy to childhood and at the same time children's ability to use phonetic cues like adults is not matured in children. Therefore, the size of visual influence on audiovisual speech perception is weaker in children compared to infants and adults. However, it is still unknown which factors make children start using visual phonetic information in audiovisual speech perception and whether speakers with different language backgrounds share the same developmental pattern. To unravel those issues, Chapter 3 subsequently suggested two hypotheses, one accounting for the different magnitude of the visual influence across languages in adults, and the other one for the influence of literacy training at school on children's audiovisual speech perception, altogether to provide a holistic account for the cross-linguistic development of audiovisual speech perception. In Chapter 4 and Chapter 5, two experiments are proposed to test the two hypotheses suggested in Chapter 3 respectively.

Chapter 3. Hypotheses in the present study

As discussed in Chapter 2, the magnitude of the McGurk effect differs across languages in adult studies. In ontogenetic studies, the degree of visual influence is also different at individual's different developmental states. However, the underlying reasons are unclear regarding why the magnitude of the McGurk effect differs in those two aspects; some potential accounts in previous studies were rebutted in Chapter 2. Therefore, Chapter 3 aims to provide two hypotheses to account for the two aspects respectively, which together holistically account for the cross-linguistic differences on the developmental trajectory of audiovisual speech perception.

3.1 Hypothesis 1: visual speech intelligibility

The size of the McGurk effect of consonants is different according to the different vowel context the consonant is situated in. Studies (Shigeno, 2000; Hampson et al., 2003) found that the McGurk effect of consonants is the most prominent in the /i/ context, moderate in the /a/ context, and weakest (almost nonexisitent) in the /u/ context. The current thesis postulates the following hypothesis:

> **Visual speech intelligibility hypothesis**
> The quality of vowels influences the visual intelligibility of the mouth's articulatory movements. With larger visual intelligibility influenced by the vowels, more visual information from speech can be adopted by perceivers, and thus more visual influences, as well as a greater size of the McGurk effect, will be demonstrated in audiovisual speech perception.

From the aspect of articulatory movements, the tongue moves forward for the front vowel and moves down for the low vowel (Zsiga, 2013), which both enlarge the visual intelligibility of mouth movements. For the back non-low vowels, such as the round vowel /u/, perceivers have difficulty detecting the articulatory movements of the speaker in such vowel context, in which lip bursting is slight and ambiguous (Shigeno,

2000). Thus, this hypothesis predicts that the McGurk effect is the most prominent in the front low vowel /æ/ context and weakest in the /u/ context among all the vowels since /æ/ has the largest visual intelligibility while /u/ has the least. However, this prediction has no empirical support since no McGurk study has been conducted in the /æ/ context yet. Despite that, the current hypothesis suggests that languages in which the front low vowel /æ/ occurs have larger visual speech intelligibility than languages where the vowel /æ/ is absent.

The production of any one phonetic segment is highly influenced by the production of phonetic segments occurring both before and after the target segment (Green & Gerdeman, 1995), which is named as coarticulation. The coarticulation has been found obvious between consonants and vowels. Liberman et al. (1967) found that vowel context (e.g., /i/ vs. /u/) heavily influenced the production of the preceding consonants (such as /d/). The phenomenon of coarticulation was accounted for from the perspective of articulatory phonology (Hall, 2010). There is a gestural overlap between the consonant and its incoming vowel – that is, the body gestural movements of the consonant and the vowel occur at the same time – for example, when you say a word key [ki], the tongue body gesture of the /k/ and the tongue body gesture of the /i/ start at the same time (for more details, see Hall, 2010). The articulation of the /i/ involves a target position in the front of the mouth and it makes the closure of the /k/ happens in a more front position than usual (Hall, 2010). The articulatory and acoustic realization of the consonant /k/ in the syllable /ki/ is therefore different compared to the /k/ in /ku/. It thus theoretically supports that the audiovisual speech perception of the consonants varies according to different vowel contexts. Furthermore, empirical studies (Shigeno, 2000; Hampson et al., 2003) supported the influence of the vowel on the size of the McGurk effect of the preceding consonant, finding that the McGurk effect on consonants is the largest in the /i/ context, moderate in the /a/ context, and weakest (almost nonexistent) in the /u/ context.

The aforementioned hypothesis assumes that the audiovisual speech perception

of consonants is affected by the visual salience of articulatory movements of vowels and thus the varied vowel inventories across languages determine speakers' different degrees of reliance on the visual information. The current hypothesis further accounts for the difference in the magnitude of the McGurk effect across languages. Each language has its vowel inventory. If the inventory in one language contains a larger proportion of visibly more intelligible vowels, the McGurk effect on the consonants in speakers of this language will be more conspicuous. Based on this proposal, the different magnitude of the McGurk effect can be compared across languages by comparing their vowel inventories. Furthermore, the current thesis postulates that only accounting for monophthongs (excluding diphthongs) is sufficient for measuring visual intelligibility because both the beginning and ending position of a diphthong have corresponding positions in one of the monophthongs in the inventory. To quantify the degree of visual intelligibility from the monophthong inventory, the current thesis proposes to follow the three subsequent rules.

Rule 1: Languages which include the vowel /æ/ have larger visual intelligibility than languages where the vowel /æ/ is absent.

Rule 2: With higher proportion of open and front vowels, the visual intelligibility is higher.

Rule 3: With higher proportion of back non-low vowels, the visual intelligibility is lower.

Rule 1 is proposed since as suggested above the vowel /æ/ has the largest visual intelligibility among all the vowels. Open vowels and front vowels increase the visual intelligibility and back non-low vowels decrease the visual intelligibility and therefore Rule 2 and Rule 3 were proposed. In the subsequent texts, it will measure the visual intelligibility in Mandarin Chinese, American English, Dutch, Cantonese, and Japanese, which are languages frequently investigated in previous research of audiovisual speech perception.

The vowel charts of American English and Mandarin Chinese are demonstrated in table 2 and table 3, summarized by Lin (2007).

Table 2 Chart of American English monophthongs

	front	*central*	*back*
high	i		u
	ɪ		ʊ
mid	e	ʌ	o
	ɛ	ə	
low			ɔ
	æ	a	ɑ

(Adapted from Lin, 2007, p. 64)

Table 3 Chart of Mandarin Chinese monophthongs

	front	*central*	*back*
high	i		u
	y		ɤ
mid	e		o
	ɛ	ə	
low	æ	a	ɑ

(Adapted from Lin, 2007, p. 65)

We see that the number of front vowels and low vowels is similar across the two languages. The low front vowel /æ/, which is predicted to be the most visually prominent, is present in both charts. Following Rule 1, the two languages both have high visual intelligibility. Furthermore, the proportion of front vowels and low vowels is 62% (8/13) in American English and 64% (7/11) in Mandarin Chinese. The back non-low vowels take up 23% (3/13) in American English and 28% (3/11) in Mandarin Chinese. The proportion of open vowels and front vowels in American English is lower

compared to Mandarin Chinese (62% versus 64%), which indicates that Mandarin Chinese may be slightly more visually intelligible than American English regulated by Rule 2. However, the proportion of back non-low vowels is also lower in American English than in Mandarin Chinese (23% versus 28%), which may indicate that the visual intelligibility is lower in Mandarin Chinese bound by Rule 3. Therefore, above all, we predict that the visual intelligibility of speech is high in both American English and Mandarin Chinese, and it is similar between the two languages. This finding predicts that the articulations in both languages have similar degrees of visual intelligibility, and the vowel context in the two languages provides a similar degree of visual influence on the audiovisual speech perception of consonants. Finally, the magnitude of the McGurk effect is predicted to have no difference between Mandarin Chinese and American English.

Compared with Mandarin Chinese and American English, Japanese vowel inventory is observably different, shown in table 4.

Table 4 Japanese monophthongs

	front	*central*	*back*
high	i		u
mid	e		o
low		a	

(Adapted from Tsujimura, 2013)

We can see that there are five vowels in Japanese, including two front vowels /i/ and /e/, and one low vowel /a/. The number and variety of front and low vowels are obviously lower than those in American English or in Mandarin Chinese. Furthermore, the front low vowel /æ/, the most visually salient vowel as suggested above, is absent in Japanese. Although the proportion of front vowels and low vowels in Japanese is 60% (3/5), marginally smaller to that in Mandarin Chinese (64%) and American English (62%), back non-low vowels take up 40% (2/5) in Japanese vowel

inventory which is much bigger than in Mandarin Chinese (28%) or in American English (23%). Therefore, the McGurk effect is weaker in Japanese compared to American English and Mandarin Chinese.

Different from Japanese, the vowel inventory is more complex in Dutch and Cantonese. Table 5 and table 6 respectively depict the chart of Dutch and Cantonese vowels.

Table 5 Chart of Dutch monophthongs

	front	*central*	*back*
high	i y		u
	e I		o
mid	ɛ Y	ø	ɔ
	œ	ə	
low		a	

(Adapted from Gussenhoven, 1992)

Table 6 Chart of Cantonese monophthongs

	front	*central*	*back*
high	i y		u
	e		o
mid	ɛ		ɔ
	œ	ɵ	
low		ɐ	
		a	

(Adapted from Zee, 1991)

The front open vowel /æ/ is absent in both Cantonese and Dutch, which indicates that the visual intelligibility in Cantonese and Dutch may not be as high as in Mandarin

Chinese and American English. In Dutch, 7 front vowels and 1 low vowel altogether account for 62% (8/13) of the inventory, and the proportion of front vowels and low vowels in Cantonese is 64% (7/11). As for non-low back vowels, the percentage is 23% (3/13) in Dutch while 28% (3/11) in Cantonese. Therefore, we predict that Dutch and Cantonese have similar degrees of visual speech intelligibility, which is lower compared to Mandarin Chinese or American English but higher than Japanese. Therefore, the current hypothesis predicts that American English speakers and Mandarin speakers should show the stronger McGurk effect than Dutch and Cantonese speakers, and in turn, the McGurk effect in Dutch and Cantonese is predicted to be larger than that in Japanese. The metric of measuring the visual speech intelligibility is given in the following table 7.

Table 7 The metric of measuring the visual speech intelligibility

	Presence of /æ/	Proportion of front vowels and open vowels	Proportion of back non-low vowels
American English	YES	62%	23%
Mandarin Chinese	YES	64%	28%
Cantonese	NO	64%	28%
Dutch	NO	62%	23%
Japanese	NO	60%	40%

In summary, the current hypothesis estimates that the McGurk effect is strongest in American English and Mandarin Chinese, intermediate in Dutch and Cantonese, and weakest in Japanese.

Furthermore, the current visual speech intelligibility hypothesis, seems to be compatible with results of previous cross-linguistic McGurk studies to the utmost compared to aforementioned hypotheses, e.g., the tonal hypothesis. The prediction that American English speakers and Mandarin Chinese speakers demonstrate no difference on the magnitude of the McGurk effect supports findings in Chen and Hazan (2009)

and Magnotti et al. (2015), although contradict results in Sekiyama (1997). However, the tonal hypothesis and the face-avoidance account cannot account for results in Chen and Hazan (2009) or Magnotti et al. (2015). The visual speech intelligibility hypothesis predicts that Cantonese speakers have a smaller size of the McGurk effect than English speakers, which is also consistent with the finding in Burnham and Lau (1998). The study by de Gelder et al. (1995), suggesting that Cantonese participants were more influenced by vision than the Dutch, seems to contradict with the visual speech intelligibility hypothesis. Two sets of stimuli were designed in their study: (1) an auditory bilabial was paired with a visual lingual (e.g., auditory /ba/ with visual /da/), (2) a visual bilabial was paired with an auditory lingual (e.g., ba-visual and da-auditory) (de Gelder et al., 1995). In the first set of stimuli, Cantonese participants reported 75 percent of visually biased responses while Dutch speakers reported 65 percent, and no significant difference was found in the second set (de Gelder et al., 1995). The difference between Dutch and Cantonese speakers is slight (65% vs. 75% in the first set of stimuli and no significant difference in the second set). Furthermore, it is unclear whether de Gelder et al. (1995) controlled the other-race effect and the non-native speaker effect since the native language and ethnicity of the speaker in the stimuli were not specified, which may cause the slight difference between Cantonese and Dutch speakers in their study. Therefore, to evaluate the visual speech intelligibility hypothesis, it is important to conduct a study across these languages with a scientific and consistent methodology. Such an experimental design is thus proposed in Chapter 4. Before that, the next subsection proposes the second hypothesis of the current thesis, to account for the difference of the magnitude of the McGurk effect between children and adults.

3.2 Hypothesis 2: phonemic awareness hypothesis - the influence of orthography and early literacy training on audiovisual speech perception

Sekiyama and Burnham (2004) found that the increase of the visual influence in Australian English children is the most prominent between 6 to 8 years and they

suggested that what is learned at school influences audiovisual speech perception because 6 years old is the onset of schooling. However, for Japanese children, the degree of the visual influence remains intact from 6 years old to adulthood. Therefore, it is possible that different schooling between Japanese and Australian English children have different degrees of influence on each group of children. Jerger et al. (2009) suggested that visual phonetic recognition is related to phonological reorganization. One of the most important periods for phonological reorganization is the early literacy training during the age between 6 to 9 years old (Anthony & Francis, 2005; Morais et al., 1986). Burnham (2003) found that children's reading ability is associated with their performance in language speech perception. Studies of last four decades have undoubtedly demonstrated that the phonological awareness in children is strongly associated with their reading ability at least in languages with alphabetic orthographies e.g., English (for reviews, see Deacon & Kirby, 2004; National Reading Panel, 2000; Snow et al., 1998). Phonological awareness refers to the insight that spoken words consist of smaller units of sound including syllable awareness, onset and rime awareness, phonemic awareness and tone awareness (Chen et al., 2004; Treiman & Zukowski, 1991). Above it was noted that the phonetic cue in visual information is used to facilitate their speech perception for adults but not for infants. As aforementioned, we postulated that children are developing their ability of using the phonetic information in audiovisual speech perception between 6 to 8 years old. Therefore, we hypothesize that:

Phonemic awareness hypothesis

The phonological awareness especially the phonemic awareness is involved in McGurk tasks, which require participants to differentiate consonants such as 'b', 'd' and 'g'. The training of early literacy and reading, which is crucial for the development of phonological awareness, is facilitating children to adopt visual phonetic information in audiovisual speech perception. However, this facilitation may be with different degrees due to different orthographies across languages that demonstrate different pace of the phonological awareness development.

The reading ability facilitates phonological awareness and most importantly phonemic awareness. Phonemic awareness allows individuals to be sensitive to the phonetic information in not only auditory speech but also visual speech. After being equipped with phonemic awareness, children are able to use phonetic information in audiovisual speech perception. Therefore, we see the nadir of the U-shaped pattern at around 6 years old, a period when children are developing their phonological awareness through literacy training. Furthermore, the developmental trajectory of audiovisual speech perception is different across languages, as we discussed in English and Japanese. Noteworthily, due to the difference of the orthography and the transparency of the grapheme-phoneme or letter-sound correspondence (for more details, see Goswami, 2008), the development of phonological awareness also differs across languages.

Defior (2004) suggested that in alphabetic languages, orthographies in languages such as German, Dutch, Turkish, Italian are considered as "transparent" while in languages like English as "opaque" depending on the application of grapheme-phoneme correspondence rules (for more details, see Frost & Katz, 1992). Previous research on alphabetic languages showed that children speaking a language with a higher transparency of orthography i.e., higher regular grapheme-phoneme correspondences, developed phonemic awareness earlier and with a speedier pace (Cossu et al., 1988; Demont & Gombert, 1996; Durgunoğlu & Öney, 1999; Høien et al., 1995; Liberman et al., 1974; Wimmer et al., 1991). For example, Cossu et al. (1988) compared the development of phonemic awareness between Italian and English preschool (4 and 5 years old) and first-grade (6 years old) children by using the tapping task and found that correct responses for phonemic segmentation are 13%, 27% and 97% respectively in 4-, 5- and 6-year-old Italian children, 0%, 17% and 70% accordingly in American English-speaking children. However, non-alphabetic languages where logograms are used, such as Chinese and Japanese, may be considered as languages with an extremely low degree of grapheme-phoneme correspondences. Mann (1986) found that Japanese first-graders performed 10% correct responses in phoneme-based

tasks, compared to 70% in English first-graders, and the performance of Japanese children in third, fourth, fifth and sixth grade was 56%, 73%, 81% and 75% respectively. It suggested that Japanese children develop phonemic awareness with a much slower pace than English children, and despite of the low transparency in Japanese orthography, Japanese children have relatively high degrees of phonemic awareness when they are older (e.g., 81% in the fifth grade). The gradual increase of phonemic awareness in Japanese children could be because Japanese orthography is not only composed of logograms but also kana characters which represent mora. Unlike Japanese, Chinese characters are 100% logograms, which makes it relevant to study the relationship between logographic languages and phonemic awareness. However, the education of romanization systems for logograms in school may significantly facilitate phonemic awareness, e.g., pin-yin system adopted in China and zhu-yin-fu-hao in Taiwan. Read et al. (1986) compared Chinese adult readers who are literate only in logograms with those who are also educated with the alphabetic pinyin system. The former group demonstrated weaker phoneme skills (20% correct responses) compared to the latter (80% correct responses) in Read et al. (1986). McBride-Chang et al. (2004) included Cantonese-speaking children in Hong Kong, Mandarin Chinese children in Xian (China), and English-speaking children in Toronto. Pinyin training at school was given only to Mandarin subjects in Xian but not to Cantonese children in Hongkong. The phonemic awareness was scored 0 for Hong Kong subjects whereas it was fairly high for children in Xian and Toronto (McBride-Chang et al., 2004). In summary, children speaking an orthography-transparent language are more advantageous in acquiring phonemic awareness than children speaking an orthography-opaque language or a logographic language. However, in logographic languages, education of the romanization system at school compensates the deficiency.

Above it was noted that orthographies are important in the development of phonemic awareness, which may be significant for the use of phonetic information in audiovisual speech perception. Previous studies in audiovisual speech perception including the audiovisual benefit effect and the McGurk effect that suggested the U-

shaped pattern may target on the perception at the phonemic. Stimuli in experiments using the McGurk effect paradigm (Dupont et al., 2005; Tremblay et al., 2007; Sekiyama & Burnham, 2004, 2008) are syllables such as 'ba', 'da', and 'ga', and we hypothesize that in those the tasks where participants are required to differentiate consonants such as 'b', 'd' and 'g', the phonemic awareness of participants is involved. In experiments for the audiovisual benefit effect such as in Lalonde and Werner (2019), syllables such as 'mu', 'gu' and 'lu' were used, which also requires phonemic awareness and the use of phonetic information. Since those tasks demand ability of audiovisual speech perception on phonemes while children's phonemic awareness is not fully developed, we see poor performance in young children and thus U-shaped pattern in previous research. Furthermore, the development of phonemic awareness varies across languages due to different orthographies (e.g., Japanese, English and Italian as seen above) and the school education across cultures (e.g., the pinyin system is taught in China but not in Hong Kong), and therefore the U-shaped pattern is not universal or identical across languages. It is noteworthy that Japanese children still developed phonemic awareness but with a much slower pace compared to alphabetic-language speakers such as English and Italian (Mann, 1986). However, Japanese adolescents and adults showed the same low degree of the visual influences as in Japanese children in Sekiyama and Burnham's (2004, 2008) McGurk studies even though Japanese adolescents and adults possess phonemic awareness. To account for this, we hypothesize that the visual speech intelligibility hypothesis as elaborated above should play a role. The development of phonemic awareness allows individuals to be able to segregate and recognize phonemes in visual speech. However, the maximum degree of the visual influence at the endpoint of the development depends on the degree of visual intelligibility of the language. Therefore, in Japanese, even though the phonemic awareness grows, the degree of the visual influence in audiovisual speech perception intactly keeps low from 6 years old to adulthood due to the limitation set by the low degree of visual intelligibility in Japanese vowels. As for English, the high degree of visual intelligibility allows the visual influence increase gradually when the phonemic

awareness grows with age in children. Therefore, we see that in English speakers, the degree of visual influence drops from infancy to preschool childhood and it rebounds after children receive literacy training at school, which altogether represent a U-shaped developmental trajectory.

In summary, the current thesis suggests two hypotheses to account for the developmental trajectory regarding the degree of the visual influence in audiovisual speech perception across languages: 1) the visual speech intelligibility hypothesis binds the maximum of the visual influence that can be reached until adulthood; 2) phonemic awareness is correlated with the degree of visual influences, i.e., with higher phonemic awareness, individuals demonstrate a higher degree of visual influences. Following the visual intelligibility hypothesis and the phonemic awareness hypothesis, the current thesis makes three predictions: first, at the endpoint of the development (in adults), the magnitude of the McGurk effect differs across languages depending on the degree of visual speech intelligibility; second, the acceleration of development between preschool children to adults is different across languages because of the different pace of phonemic awareness development due to different orthographies; last, the U-shaped pattern is absent in some languages such as Japanese in Sekiyama and Burnham (2008). In Chapter 4 and Chapter 5, two experiments are designed to examine the validity of the two hypotheses.

Chapter 4. Experiment 1

4.1 Introduction

As suggested above, the three hypotheses in previous studies, namely the tonal hypothesis, the face-avoidance hypothesis, and the phonetic inventory hypothesis are all incompatible with previous empirical studies. The first hypothesis proposed in the current thesis, the visual speech intelligibility, is compatible with results of previous studies to the utmost compared to other hypotheses, as discussed in the previous chapter. The main purpose of Chapter 4 is to design an experiment to examine the validity of the visual speech intelligibility hypothesis.

Six languages will be examined, consisting of American English, Mandarin Chinese, Cantonese, Dutch, Japanese and Russian. The first five languages, English, Chinese, Cantonese, Dutch and Japanese were all investigated in previous studies, as summarized in table 1, but they were never included in one single study at the same time. Reduplicating investigations on these five languages will provide more discussions on cross-linguistic McGurk effect and comparisons with previous studies. Russian is included in the current study due to its same vowel inventory with Japanese, as seen in Table 8 below, which can exemplify whether the same degree of the McGurk effect can be found in two languages with same degree of visual intelligibility. The six languages form three levels of visual intelligibility as suggested in the previous chapter, with highest in English and Mandarin Chinese, middle in Cantonese and Dutch, and lowest in Japanese and Russian. Following the hypothesis of the visual intelligibility, the order of the magnitude of the McGurk effect across these six languages is: (Mandarin Chinese ≈ American English) > (Dutch ≈ Cantonese) > (Japanese ≈ Russian). In addition, the inclusion of Russian provides new language data for McGurk-effect experiments since no such study has been conducted in Russian.

Table 8 Russian monophthongs

	front	*central*	*back*
high	i		u
mid	e		o
low		a	

(Adapted from Jones & Ward, 1969)

This study plans to examine the fused perception (the response of the /da/ to the stimulus of auditory /ba/ and visual /ga/) of the McGurk effect and exclude investigations of the combined response (/g͡ba/ to the stimulus of visual /ba/ and auditory /ga/). Sekiyama and Tohkura (1993, 1994) found that Japanese speakers had few combination responses, and they suggested that it was because the Japanese phonological system does not allow any phonological consonant clusters, which may cause Japanese speakers more biased to the single consonantal articulation (e.g., /ba/ or /ga/) rather than double articulation (e.g., /g͡ba/) compared to speakers of other languages which allow consonant clusters. Like Japanese, consonant clusters are also not allowed in Mandarin Chinese. Therefore, to avoid the impact of the presence of consonant clusters, only the fused McGurk effect will be studied in the current proposal.

Regarding the syllables, only /pa/, /ta/ and /ka/ will be included, which are all devoiced stops. Although the most frequently investigated syllables in previous studies were voiced stops /ba/, /da/, and /ga/, there are two reasons to adopt devoiced stops. First, there is no voiced stop in Mandarin Chinese, while the devoicing aspirated stop and its unaspirated counterpart (such as /pʰ/ and /p/) are separate phonemes. However, in Chinese pinyin system, the unaspirated stops are written as devoiced counterparts (e.g., /p/ is written as “b”). It is possible that Chinese subjects produce unaspirated devoiced /p/ even though they are presented with voiced /b/. Another reason is that when Dutch speakers see the letter “g”, they pronounce it as the velar fricative /x/ or /ɣ/, instead of the velar stop /g/ as in English. In contrast, the three devoiced stops /pa/,

/ta/, and /ka/ are legitimate syllables in all the six languages that are being investigated, and therefore they are chosen in the current proposal.

Magnotti et al. (2015) criticized that it is problematic for cross-linguistic studies to use stimuli produced by only two talkers since different speakers may evoke different degrees of the McGurk effect due to their great variabilities. Therefore, the current proposal will recruit 8 native speakers with gender counterbalanced in each language to create stimuli. The face of the speaker will be in line with the language she/he speaks. In other words, the one who speaks Chinese, Japanese, or Cantonese has an Asian face, and the one who speaks English, Dutch, or Russian has a Caucasian face.

As discussed in the literature review, it is crucial to control the age of subjects since there may be age effects in audiovisual speech perception even after adulthood. Tse & Herrup (2017) suggested that the cognitive performance in lifespan peaks at 20 – 30 years. To be more cautious, the age range of the participants in each language group will be 23 – 26 years in the current proposal.

Unlike studies by Sekiyama and Tohkura (1993) or Chen and Hazan (2009), the current proposal decides not to examine the non-native speaker effect. The current study aims to investigate whether the magnitude of the McGurk effect differs across languages when perceivers process their native languages. Thus, English stimuli will only be presented to American subjects and the same procedure will be applied to the other five languages. In other words, each participant will be presented with stimuli of their native language.

4.2 Method

4.2.1 Participants

Native monolingual speakers in each language of Mandarin Chinese, American English, Cantonese, Dutch, Japanese and Russian will be recruited in this study. The

age of all participants will fall at the range of 23 to 26 years old. In each language group, there will be 30 participants and the gender will be counterbalanced. All participants should have a normal (or corrected-to-normal) vision and no language or hearing deficits.

4.2.2 Stimuli

The stimuli will be created by three syllables /pa/, /ta/, and /ka/ uttered by 8 different speakers in each of the aforementioned six languages, and therefore there will be 48 speakers to be recorded, with 4 males and 4 females in each language. Videos will be recorded in a sound-attenuated room with high-end equipment for all speakers. Since speakers may communicate with the experimenter only in English, which may not be their native language, they may stay in the mode of speaking a foreign language before their speech was collected. Therefore, they will first be instructed to read verbally some newspaper texts in their native language, which is designed to allow performers to attune to their native speech. Later, they will be asked to articulate 10 repetitions of each syllable. Both the face and the speech will be recorded for each speaker. one of the repetitions with the highest quality for each syllable will be chosen for each speaker.

Three types of stimuli will be designed based on the collected materials: visual-only (V), auditory-only (A), and audio-visual (AV) stimuli. The AV stimuli will be composed of AV congruent stimuli and AV incongruent stimuli. The AV incongruent stimuli will be the McGurk stimuli, where the audio will be /pa/ and the visual information will be /ka/. Furthermore, the AV incongruent stimuli will be the experimental items, which will take up 1/4 of the whole AV stimuli. However, the AV congruent stimuli will be fillers, consisting of audiovisual /pa/, audiovisual /ta/, and audiovisual /ka/, which will account for 3/4 of the AV stimuli. As in Chen and Hazan (2009), the audio track will be cut out to create the V stimuli. In the A stimuli, the visual information will be substituted by a black screen. In either A or V stimuli, there will be

24 trials (8 speakers x 3 syllables). In AV stimuli, there will be 32 trials (8 speakers x (3 congruent AV + 1 incongruent AV)). Thus, there will be 80 trials (24 + 24 + 32) in total.

4.2.3 Procedure

The forced-choice paradigm will be adopted in the experiment. After each trial of the stimuli, three buttons of characters representing sound “pa”, “ta” and “ka” will appear on the screen in orthographies of each language. For example, in Dutch and English, they will be Roman letters, and Chinese logograms that correspond to the same syllable as in English /pa/, /ta/, and /ka/ will be used to test Mandarin Chinese subjects. Two reasons of doing that are: (1) participants may have no education on the pronunciation of Roman letters; (2) participants may switch to the mode of speaking a foreign language once they see foreign characters while the current study aims to find responses to their native stimuli. After being shown each stimulus, participants will be asked to click one of the three buttons they think they have perceived.

The stimuli will be presented to participants in three blocks in the order of AV, A and V, as in Chen and Hazan (2009). The V condition was given last because of its difficulty and the A condition was interposed to avoid a transfer of visual effect from AV to V block (Chen & Hazan, 2009). The presentation of the stimuli will be randomized in each block. There will be a two-second interval between each stimulus. After each block, a one-minute break will be given. The experiment is expected to last around 20 minutes in total for each participant.

4.3 Hypothesized results

For each of the six language groups, we will collect results from 30 participants, with each of them giving 80 responses. In sum, the data will consist of 2400 (30x80) responses for each language. Like in Magnotti et al. (2015), responses to McGurk effect stimuli (experimental stimuli/incongruent AV stimuli) will be categorized in three

different ways. The responses “ta” will be categorized as McGurk fusion responses. The responses “pa” will be categorized as auditory responses and “ka” as visual responses.

In general, across the six languages, we hypothesize that the visual responses will be rare and the fused and auditory responses will be more frequent. The current study also hypothesizes that the proportion of the three types of responses differs across languages. The fused responses and the visual responses should take up a higher proportion in a language with greater articulatory visual intelligibility, predicted by the visual speech intelligibility hypothesis. However, in a language with weaker visual intelligibility, the auditory responses will be more frequent since it has less clear visual information.

To analyze the data, a repeated-measure ANOVA will be performed with participant languages as factors and the percentage of McGurk fusion responses as the dependent measure. The current study hypothesizes that the percentage of McGurk fusion responses in English and Mandarin will be significantly higher than in Dutch and Cantonese, in which the percentage of McGurk fusion responses will in turn be significantly higher than in Japanese and Russian. In addition, we hypothesize that no significant difference will be found between English and Mandarin, between Dutch and Cantonese, or between Japanese and Russian.

4.4 Discussion

The current experiment aims to provide a clear understanding of the influence of language background on the audiovisual speech perception, more specifically, the McGurk effect, since the consensus has not been reached regarding the various magnitude of the McGurk effect across languages. It first examines whether the McGurk effect is less saliently seen in Mandarin speakers compared to English speakers like in Sekiyama (1997), or there is no significant difference between Mandarin and English speakers as found in Chen and Hazan (2009). We hypothesize that the result

will be consistent with Chen and Hazan (2009) and Mandarin speakers demonstrate the same magnitude of the McGurk effect with English speakers. Second, we predict that the result that Japanese speakers showed a much weaker McGurk effect than English speakers found in Sekiyama and Tohkura (1991; 1993) will be replicated in the current proposal. Furthermore, as mentioned before, previous findings (Burnham & Lau, 1998; de Gelder et al., 1995) in English, Dutch and Cantonese speakers suggested that Cantonese speakers demonstrated a weaker McGurk effect than English speakers but stronger than Dutch speakers. However, the difference between Dutch and Cantonese speakers in de Gelder et al. (1995) was slight and might be due to the problematic methodology, as suggested earlier. Therefore, we aim to provide more discussion on the McGurk effect in Dutch and Cantonese speakers and we predict there will be no significant difference of that in the two groups of speakers. Last, we include one another language, Russian, that has never been observed regarding the McGurk phenomenon. Another reason for choosing this language is that it has almost the same vowel inventory as Japanese, and we predict that the magnitude of the McGurk effect in Russian is not different from Japanese but weaker than the other four languages.

Three hypotheses were suggested to account for the different degrees of the McGurk effect across languages: (1) the tonal hypothesis, (2) face-avoidance cultural convention, (3) the phonetic inventory (for more details, see Sekiyama & Burnham, 2004). The tonal hypothesis and the face-avoidance convention will be countered by the theoretical framework proposed in Chapter 3, the visual speech intelligibility hypothesis, which is developed from the phonetic inventory account. However, the phonetic inventory takes the number of phonemes into account while the current proposal focuses only on the vowels, which are suggested to be crucial for the McGurk effect of consonants.

The tonal hypothesis suggests that if more tonal information is adopted in one language, speakers are prone to use less visual information in this language since auditory cues are more informative in tone perception than visual cues (Chen & Hazan,

2009). The face-avoidance convention means, in Asian cultures such as in China and Japan, the cultural habit to avoid eye contact during conversation especially with the one who has a higher socio-economic status. Under the two hypotheses, the McGurk effect should be strongest in English speakers, intermediate in Japanese speakers, and weakest in Chinese Mandarin speakers, in line with Sekiyama's studies (Sekiyama & Tohkura 1991, 1993; Sekiyama, 1994, 1997). In contrast, the current proposal predicts that there is no difference in the magnitude of the McGurk effect between in Mandarin speakers and in English speakers, and the McGurk effect is weaker in Japanese speakers compared to the previous two groups.

To further examine the visual speech intelligibility hypothesis, six languages were selected with different vowel inventories as illustrated in the current chapter, which are Mandarin Chinese, American English, Dutch, Cantonese, Japanese and Russian. The degree of visual intelligibility is highest in Mandarin Chinese and American English, intermediate in Dutch and Cantonese, and smallest in Japanese and Russian. According to the hypothesis, the articulatory movement is the most visually intelligible in Mandarin Chinese and American English, mediocre in Dutch and Cantonese, and the least salient in Japanese and Russian. Therefore, the reliance on the visual information and the McGurk effect are accordingly the most prominent in Mandarin Chinese and American English speakers, secondary in Dutch and Cantonese speakers, and least in Japanese and Russian speakers.

Chapter 5. Experiment 2

5.1 Introduction

Results in Sekiyama and Burnham (2004, 2008) are in line with the current two hypotheses i.e., visual speech intelligibility hypothesis and phonemic awareness hypotheses, finding that the degree of visual influences increased after 6 years old in English participants but stayed intact in Japanese participants. Chen and Hazan (2009) using the McGurk effect paradigm found that Mandarin Chinese perceivers and English perceivers had the identical developmental pattern, with smaller visual influences in 8- to 9-year-old children and larger in adults in both language groups, and there is no difference between the two language backgrounds. The study by Chen and Hazan (2009) also supports the two hypotheses. Above it was noted that English (an alphabetic language) and Mandarin Chinese (with education of pinyin system) children both develop their phonemic awareness relatively fast after schooling, e.g., in McBride-Chang et al. (2004). Furthermore, English and Chinese have the same and high degree of visual intelligibility.

To further justify the two hypotheses, studies are needed to investigate children with other language backgrounds and control either the transparency of the orthography (impacts of hypothesis 2) or the degree of visual intelligibility (impacts of hypothesis 1). Dutch and Cantonese are ideal in several aspects. Dutch and Cantonese have the same degree of the visual intelligibility and are predicted to be identical in the magnitude of the McGurk effect in adults, as suggested in Chapter 3 and 4. However, Dutch has relatively high grapheme-phoneme transparency (Frost & Katz, 1992; Borleffs et al., 2017) whereas Cantonese-speaking children in Hong Kong learn Chinese logograms with extremely low transparency without education of the pinyin system. Although no study directly compared phonemic awareness between Cantonese and Dutch children, Patel et al. (2004) found that Dutch children were faster and more accurate in tests of phonemic awareness than English children of the same age and

McBride-Chang et al. (2004) found that Cantonese children in Hong Kong were much poorer in the performance of phonemic awareness compared to same-aged English children. The two studies (McBride-Chang et al. 2004; Patel et al., 2004) collectively suggested that Dutch speakers develop full phonemic awareness faster than Cantonese speakers growing up in Hong Kong.

The current study aims to investigate the magnitude of the McGurk effect across ages in Dutch and Hong Kong Cantonese speakers. Although adult McGurk studies have been conducted in Cantonese and Dutch as reviewed in Chapter two, no study has explored the developmental trajectory of audiovisual speech perception in Dutch and Cantonese children. Also, results from the current developmental study in Dutch and Cantonese can be compared with developmental research in other languages such as English and Japanese in Sekiyama and Burnham (2004, 2008), and English and Mandarin Chinese in Chen and Hazan (2009). The development of audiovisual speech perception in infancy is not included in the current study due to two reasons: (1) a large body of infant studies in audiovisual speech perception was already conducted with consistent findings; (2) it is beyond the scope of the current thesis. Therefore, the current study focuses on the development from preschool children of 6 years old to adulthood.

5.2 Method

5.2.1 Participants

Four age groups respectively in Dutch and Cantonese monolinguals will participate in the current experiment, including 6-year-old preschoolers, 7-year-old first graders, 8-year-old second graders, and 9-year-old third graders. Each age group in each language will consist of 30 participants, same as the number in the adult study proposed in chapter two, which will lead to 240 participants in total. Dutch participants will be recruited from kindergartens and elementary schools in the Netherlands, and Cantonese participants in Hong Kong. Experimenters will ensure that all Cantonese participants

have not received any education on pinyin system or zhu-yin-fu-hao at home, school or any training center by requesting a survey from each caregiver. The gender will be counterbalanced in each group. All participants should have a normal (or corrected-to-normal) vision and no language or hearing deficits.

5.2.2 Stimuli and procedure

The stimuli will be the same with those in the adult study in the previous chapter. The research method, however, will be different in children. Like in Tremblay et al. (2007), participants will be asked to repeat the syllable they think they have heard after being presented with the stimuli since preschoolers may not recognize characters. It takes around 20 minutes for each adult participant to complete the experiment. For children, however, intermission especially for 6-year-old children will be necessary as suggested in Sekiyama and Burnham (2008), and therefore it takes around 45 minutes for each child participant to finish the experiment.

5.3 Hypothesized results

To calculate the size of visual effects, the current study adopts the method in Sekiyama and Burnham (2008). Combining both the positive and negative visual effects, the total degree of visual influence can be taken as the difference between the percent of auditorily correct responses to audiovisual congruent stimuli and the percent of auditorily correct responses to McGurk (audiovisual incongruent) stimuli (Sekiyama & Burnham, 2008).

As suggested in the adult study in Chapter 4, Dutch and Hong Kong Cantonese speakers in adulthood are predicted to present the same magnitude of the McGurk effect and demonstrate the same size of visual influence. However, due to the transparency in Dutch orthography and the logographic feature in Chinese orthography, Dutch speakers achieve the adultlike degree of visual influence at a faster pace than Cantonese speakers.

Combining the hypothesized results in both child and adult studies, the following figure 1 depicts the hypothesized difference of the developmental trajectory between Dutch and Cantonese speakers.

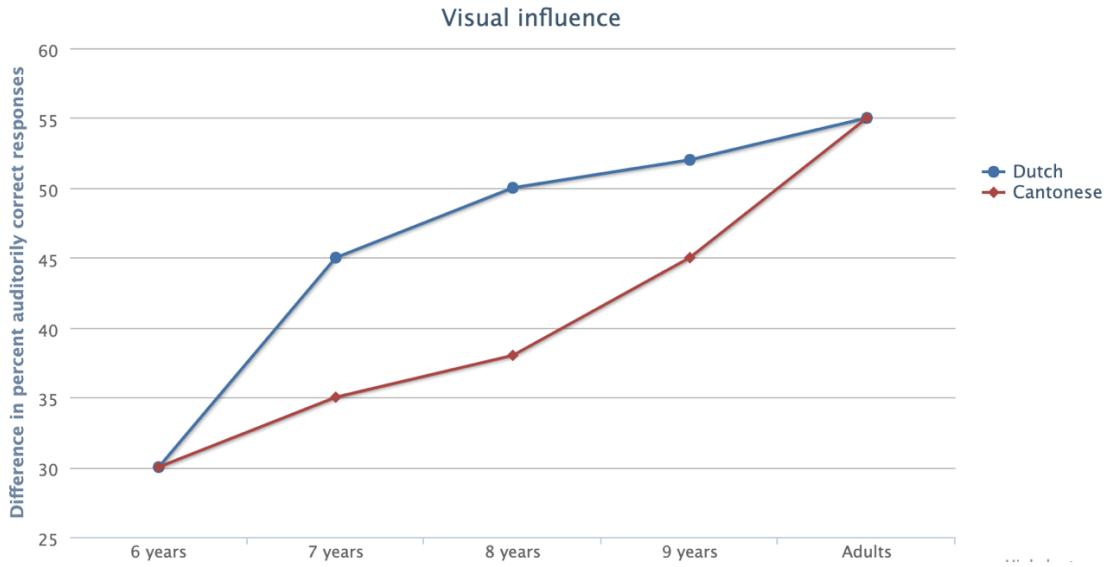

Figure 1 Visual influence across ages between Dutch and Cantonese

Before schooling, Dutch children and Cantonese children should have the same degree of the visual influence at 6 years old since the literacy training at school has not started yet. After learning to read, the visual influence gradually increases in both language groups with a faster pace in Dutch children. Finally Dutch and Cantonese adults demonstrate the identical visual influence due to the same degree of visual speech intelligibility.

5.4 Discussion

The current proposed experiment aims to examine the development of audiovisual speech perception in Dutch and Cantonese speakers, specifically in children of 6 to 9 years old. In Tremblay et al. (2007), the size of visual influence is adultlike after 9 years old in English speakers (found in 10-14-year-old group and 15-19-year-old group). In addition, previous research on the relationship between reading and development of phonemic awareness suggested that the most important period for

phonological reorganization is the early literacy training during the age between 6 to 9 years old (Anthony & Francis, 2005; Morais et al., 1986). Furthermore, Sekiyama and Burnham (2008) suggested that the increase of visual influence between 6 years old and 8 years old is highly prominent in English speakers. Therefore, the present study focuses on the development from 6 to 9 years old.

As suggested above, the current thesis considered two factors as significant, namely the literacy training and the visual speech intelligibility, to influence the developmental trajectory of audiovisual speech perception across languages. Results in English and Japanese speakers in Sekiyama and Burnham (2004, 2008) and in English and Mandarin Chinese speakers in Chen and Hazan (2009) supported the hypotheses, as previously discussed. The present study investigates Dutch and Cantonese, which have distinct types of literacy training and visual speech intelligibility compared to Japanese, English and Mandarin Chinese, to further testify the hypotheses.

More importantly, results in the current study can be compared with studies in Sekiyama and Burnham (2008) and Chen and Hazan (2009). Although different age groups were included across studies, i.e., only 8-year-old children and adults in Chen and Hazan (2009) and 6-, 8-, 11-year-old children and adults in Sekiyama and Burnham (2009), we hypothesized the developmental trajectories in English, Mandarin Chinese, and Japanese as shown below in figure 2.

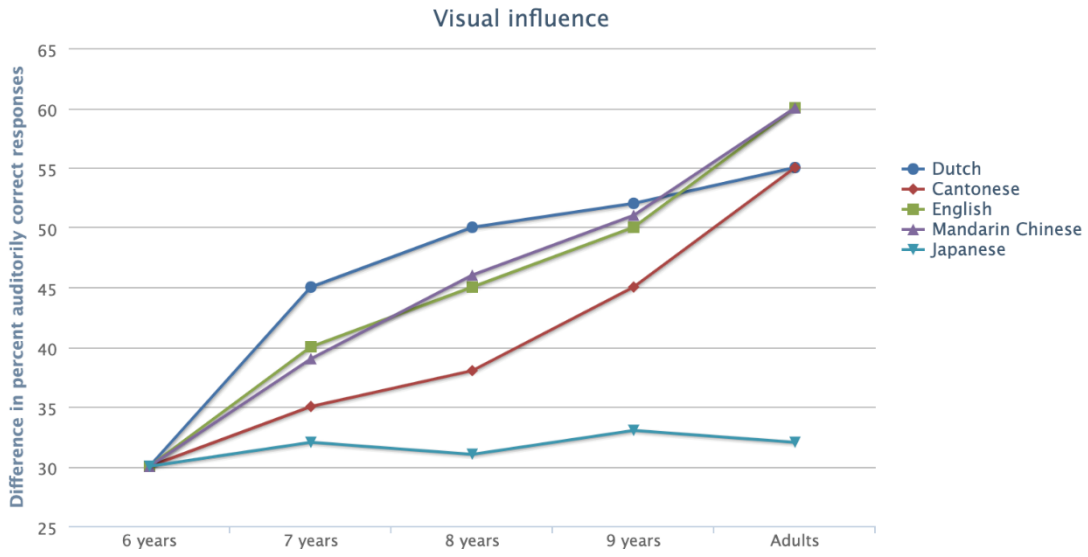

Figure 2 Visual influence across ages in Dutch, Cantonese, English, Mandarin Chinese and Japanese

As seen from figure 2, due to the highest transparency of orthography in Dutch, we hypothesize that Dutch children acquire phonemic awareness at the fastest speed and therefore demonstrate the fastest increase in the visual influence. The developmental trajectory in Mandarin Chinese is identical with in English, in line with results in Chen and Hazan (2009). The trajectories in English and Japanese in figure 2 are also consistent with findings in Sekiyama and Burnham (2008). At the end of each trajectory, the magnitude of visual influence is largest in English and Mandarin Chinese adults, intermediate in Dutch and Cantonese, lowest in Japanese, as suggested in Chapter 4. In summary, the visual speech intelligibility, results in the cross-linguistic difference in the magnitude of visual effects in speech perception in adults; the transparency of the orthography and literacy education determine the relative speed of the increase in the visual influence, with fastest in Dutch, intermediate in English and Mandarin, lowest in Cantonese.

Chapter 6 General discussion

This thesis investigated why the magnitude of the McGurk effect differs in two aspects, first in different languages and second in different ages, since in previous research inconsistent results have been found and no consensus has been agreed on. Two main research questions were thus raised in the current thesis: (1) why does the magnitude of the McGurk effect differ across languages? (2) why does the developmental trajectory of the McGurk effect demonstrate a U-shaped pattern?

To answer the first research question, this thesis proposed the visual speech intelligibility hypothesis. This hypothesis suggested that in the vowel inventory of a language, if vowels have wider and more drastic mouth movements, the articulation of consonants in this language will present more salient and intelligible visual information. Furthermore, the stronger McGurk effect is shown in languages with higher visual intelligibility in consonants. As mentioned above, the hypothesis is supported by articulatory phonetics and coarticulation. The perception of a consonant is closely related to the surrounding vowels. Also, the McGurk effect of consonants is more easily demonstrated if the vowel is front or open (e.g., /a/ and /i/) compared to back non-low vowels such as /u/ (Shigeno, 2000; Hampson et al., 2003). With non-low back vowels, less visual information can be adopted in audiovisual speech perception while visual cues of consonants are more salient if the following vowel is front and/or open. Therefore, the visual influence is more prominent in a vowel inventory with more front and open vowels. To validate this hypothesis, this thesis proposed a cross-linguistic experiment in adults in Chapter 4. It suggested that the McGurk effect is strongest in American English and Mandarin Chinese speakers, intermediate in Dutch and Cantonese speakers, and lowest in Japanese and Russian speakers. This prediction was made after qualitatively analyzing the vowel inventory of each language. Following the visual speech intelligibility hypothesis, we suggested that American English and Mandarin Chinese are the least rounded languages while Japanese and Russian have the highest degree of roundedness and Dutch and Cantonese are in between. Previous

cross-linguistic studies in the McGurk effect consisted of only two languages and the hypotheses from those studies failed to account for other languages. For example, the tonal hypothesis and the face-avoidance hypothesis are plausible to account for results in Japanese and American English speakers as in Sekiyama and Tohkura (1991, 1993) but the results in Mandarin Chinese and American English speakers countered the two hypotheses as in Chen and Hazan (2009) and Magnotti et al. (2015). The visual speech intelligibility hypothesis, which is possible to be applied to every language, complements this shortcoming present in previous studies. However, this hypothesis also has limitations. In this hypothesis, the way to quantify the visual intelligibility of different languages is by spotting front and open vowels in each vowel inventory. It suggested that with a higher proportion of front and open vowels, the visual speech is more articulatorily salient. However, a higher proportion of front and open vowels in the inventory does not necessarily infer that the front and open vowels are more frequently present in this language. To resolve this, future corpus studies may be conducted to calculate the frequency of open and front vowels across languages and examine its correlation with the degree of the McGurk effect. Another limitation of the visual speech intelligibility hypothesis is that even for the same front and/or open vowel in different languages, the frontness and/or openness may be different. For example, although the open vowel /a/ is present in both Mandarin Chinese and Japanese, this vowel is opener in Mandarin Chinese compared to Japanese, judged from experiences of the author as a Mandarin Chinese native speaker and second language learner of Japanese. To more accurately quantify the degree of vowel openness and frontness across languages, the electromagnetic articulography could be adopted in the future experiment to measure the mouth openness in speech of different languages.

The second hypothesis in the current thesis, the phonemic awareness hypothesis, was suggested to account for the second research question, i.e., why does the audiovisual speech perception development demonstrate a U-shaped pattern. As suggested earlier, the U-shaped pattern in the development of audiovisual speech perception means that the visual information of speech impacts infants, adolescents and

adults to a larger extent compared to children especially preschoolers between 4 to 6 years old (Dupont et al., 2005; Ross et al., 2011; Sekiyama & Burnham, 2008; Tremblay et al., 2007; Wightman et al., 2006; for a review, see Lalonde & Werner, 2021). The phonemic awareness hypothesis suggests that due to the underdevelopment of phonemic awareness in preschool children, they are unable to adopt the phonetic information of visual speech and therefore they rely mostly on auditory information for phoneme distinguishing tasks. After children enter school and receive literacy training, which greatly helps developing their phonemic awareness, the visual influences in audiovisual speech perception increase gradually with age. Therefore, the U-shaped pattern is demonstrated in audiovisual speech perception development. Although results in previous developmental studies in Mandarin Chinese and English children (Chen & Hazan, 2009; Sekiyama & Burnham, 2008) already supported the phonemic awareness hypothesis, this thesis in Chapter 5 proposed to investigate Dutch and Cantonese, two languages that are identical in the degree of visual speech intelligibility whereas different in the development of phonemic awareness, as illustrated before. We predicted that the developmental trajectory is also U-shaped in Dutch and Cantonese. However, the increase is speedier in Dutch because of the faster development of phonemic awareness in Dutch children compared to Cantonese children educated in Hong Kong. Therefore, although the endpoint of the two languages is identical as hypothesized in the adult study in Chapter 4, the U-shaped trajectory is different between Dutch and Cantonese.

The two main research questions, one targeting on the cross-linguistic difference in adults, the other on the difference throughout the development, altogether address one central question: why is the development of audiovisual speech perception different across languages? The two hypotheses, the visual speech intelligibility and the phonemic awareness hypothesis, collectively answers the question. Although the U-shaped pattern was found in English and Canadian French speakers (Dupont et al., 2005; Ross et al., 2011; Sekiyama & Burnham, 2008; Tremblay et al., 2007; Wightman et al., 2006; for a review, see Lalonde & Werner, 2021) and we suggested that it should also

be illustrated in Dutch and Cantonese, the U-shaped pattern is not identical or universal across languages in three regards. First, the endpoint differs across languages as estimated by the visual speech intelligibility hypothesis; second, the acceleration of development between preschool children to adults is different across languages because of the different pace of phonemic awareness development due to different orthographies; last, the U-shaped pattern is absent in some languages such as Japanese in Sekiyama and Burnham (2008).

Experiments proposed in the current thesis adopt behavioral research paradigm. Behavioral methods reflecting the conscious response of the subjects, however, may not be as sensitive as neurological methods in audiovisual speech integration, which is an involuntary process that happens in several milliseconds. Therefore, future neurolinguistic research will be significant to further assess the two hypotheses suggested in the current thesis. An fMRI study (Sinozaki et al., 2016) found that different neural networks were adopted between Japanese and English monolingual adults in audiovisual speech processing. More specifically, connectivities of Thalamus-Calcarine, Thalamus-Heschl, and Thalamus-MT were present in English speakers but not in Japanese speakers, and Japanese speakers merged visual and auditory information only at the Superior Temporal Sulcus (STS), through connectivities of Calcarine/MT-STS and Heschl-STS (Sinozaki et al., 2016). However, it remains unclear since when the brain network for audiovisual speech perception in Japanese and English speakers diverges from each other. It may thus be significant to conduct fMRI studies of audiovisual speech perception in Japanese and English children. If the neural network for audiovisual speech perception in preschoolers is different from that in literate children with phonemic awareness and if the different brain regions are associated with the development of phonemic awareness, it will significantly acknowledge the influence of reading and early literacy training on audiovisual speech perception and thus support the phonemic awareness hypothesis.

References

Aldridge, M. A., Braga, E. S., Walton, G. E., & Bower, T. G. R. (1999). The intermodal representation of speech in newborns. *Developmental Science*, *2*(1), 42-46. https://doi.org/10.1111/1467-7687.00052

Anthony, J. L., & Francis, D. J. (2005). Development of phonological awareness. *Current Directions in Psychological Science*, *14*(5), 255-259. https://doi.org/10.1111/j.0963-7214.2005.00376.x

Baart, M., Vroomen, J., Shaw, K., & Bortfeld, H. (2014). Degrading phonetic information affects matching of audiovisual speech in adults, but not in infants. *Cognition*, *130*(1), 31-43. https://doi.org/10.1016/j.cognition.2013.09.006

Bernstein, L. E., Auer, E. T., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication*, *44*(1), 5-18. https://doi.org/10.1016/j.specom.2004.10.011

Borleffs, E., Maassen, B. A., Lyytinen, H., & Zwarts, F. (2017). Measuring orthographic transparency and morphological-syllabic complexity in alphabetic orthographies: a narrative review. *Reading and Writing*, *30*(8), 1617-1638. https://doi.org/10.1007/s11145-017-9741-5

Brancazio, L., Miller, J. L., & Paré, M. A. (2003). Visual influences on the internal

structure of phonetic categories. *Perception & Psychophysics*, *65*(4), 591–601. https://doi.org/10.3758/BF03194585

Burnham, D., & Dodd, B. (1998). Familiarity and novelty in infant cross-language studies: factors, problems, and a possible solution. *Advances in Infancy Research*, *12*, 170-187.

Burnham, D., & Dodd, B. (2004). Auditory–visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology*, *45*(4), 204-220. https://doi.org/10.1002/dev.20032

Burnham, D., & Lau, S. (1998). The effect of tonal information on auditory reliance in the McGurk effect. In *AVSP'98 International Conference on Auditory-Visual Speech Processing*.

Bushara, K. O., Grafman, J., & Hallett, M. (2001). Neural correlates of auditory–visual stimulus onset asynchrony detection. *Journal of Neuroscience*, *21*(1), 300-304. https://doi.org/10.1523/JNEUROSCI.21-01-00300.2001

Campbell, R. (2008). The processing of audio-visual speech: empirical and neural bases. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 1001-1010. https://doi.org/10.1098/rstb.2007.2155

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

Chen, T. H., & Massaro, D. W. (2004). Mandarin speech perception by ear and eye follows a universal principle. *Perception & Psychophysics*, *66*(5), 820-836. https://doi.org/10.3758/BF03194976

Chen, X., Anderson, R. C., Li, W., Hao, M., Wu, X., & Shu, H. (2004). Phonological awareness of bilingual and monolingual Chinese children. *Journal of Educational Psychology*, *96*(1), 142-151. https://doi.org/10.1037/0022-0663.96.1.142

Chen, Y., & Hazan, V. (2009). Developmental factors and the non-native speaker effect in auditory-visual speech perception. *The Journal of the Acoustical Society of America*, *126*(2), 858-865. https://doi.org/10.1121/1.3158823

Cossu, G., Shankweiler, D., Liberman, I. Y., Katz, L., & Tola, G. (1988). Awareness of phonological segments and reading ability in Italian children. *Applied Psycholinguistics*, *9*(1), 1-16. https://doi.org/10.1017/S0142716400000424

de Gelder, B., & Vroomen, J. (1992). Auditory and visual speech perception in alphabetic and non-alphabetic Chinese-Dutch bilinguals. In R. J. Harris (Ed.), *Cognitive processing in bilinguals*, (pp. 413-426). (Advances in psychology). North-Holland Publishing Company.

de Gelder, B., Bertelson, P., Vroomen, J., & Chen, H. C. (1995). Inter-language differences in the McGurk effects for Dutch and Cantonese listeners. In *Eurospeech 1995: Proceedings of the Fourth European Conference on Speech Communication and Technology*, *Madrid*, *Spain*, *September 18-21*, *1995* (pp. 1699-1702). International Speech Communication Association (ISCA).

Deacon, S. H., & Kirby, J. R. (2004). Morphological awareness: Just" more phonological"? The roles of morphological and phonological awareness in reading development. *Applied Psycholinguistics*, *25*(2), 223-238. https://doi.org/10.1017/S0142716404001110

Demont, E., & Gombert, J. E. (1996). Phonological awareness as a predictor of recoding skills and syntactic awareness as a predictor of comprehension skills. *British Journal of Educational Psychology*, *66*(3), 315-332. https://doi.org/10.1111/j.2044-8279.1996.tb01200.x

Desjardins, R. N., & Werker, J. F. (1996). 4-month-old female infants influenced by visible speech. *Infant Behavior and Development*, (19), 421. https://doi.org/10.1016/S0163-6383(96)90475-0

Dodd, B. (1979). Lip reading in infants: Attention to speech presented in-and out-of-synchrony. *Cognitive Psychology*, *11*(4), 478-484. https://doi.org/10.1016/0010-0285(79)90021-5

Dupont, S., Aubin, J., & Ménard, L. (2005). A 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

Durgunoğlu, A. Y., & Öney, B. (1999). A cross-linguistic comparison of phonological awareness and word recognition. *Reading and Writing*, *11*(4), 281-299. https://doi.org/10.1023/A:1008093232622

Eggermont, J. J., & Moore, J. K. (2012). Morphological and functional development of the auditory nervous system. In Werner L., Fay R., Popper A. (Eds.) *Human auditory development* (pp. 61-105). Springer, New York, NY. https://doi.org/10.1007/978-1-4614-1421-6_3

Frost, R., & Katz, M. (Eds.). (1992). *Orthography, phonology, morphology and meaning*. Amsterdam: Elsevier Science Publishers.

Goswami, U. (2008). The development of reading across languages. *Annals of the New York Academy of Sciences*, *1145*(1), 1-12.

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, *108*(3), 1197-1208. https://doi.org/10.1196/annals.1416.018

Green, K. P., & Gerdeman, A. (1995). Cross-Modal discrepancies in coarticulation and the integration of speech information: The McGurk effect with mismatched vowels. *Journal of Experiment Psychology: Human Perception and*

Performance, *21*(6), 1409–1426. https://doi.org/10.1037/0096-1523.21.6.1409

Gussenhoven, C. (1992). Dutch. *Journal of the International Phonetic Association*, *22*(1-2), 45-47. https://doi.org/10.1017/S002510030000459X

Hall, N. (2010). Articulatory phonology. *Language and Linguistics Compass*, *4*(9), 818-830. https://doi.org/10.1111/j.1749-818X.2010.00236.x

Hampson M., Guenther F., Cohen M., Nieto-Castanon A. (2003). Changes in the McGurk effect across phonetic contexts. Technical Report CAS/CNS 03-006, Boston University, MA, USA. Retrieved from https://pdfs.semanticscholar.org/dc98/5ea3a57d1d1175b0a4ab595e6649de409e9a.pdf.

Han, Y., Goudbeek, M., Mos, M., & Swerts, M. (2020). Relative Contribution of Auditory and Visual Information to Mandarin Chinese Tone Identification by Native and Tone-naïve Listeners. *Language and Speech*, *63*(4), 856-876. https://doi.org/10.1177/0023830919889995

Hayashi, Y., & Sekiyama, K. (1998). Native-foreign language effect in the Mcgurk effect: A test with Chinese and Japanese. In *AVSP'98 International Conference on Auditory-Visual Speech Processing*, 61-66.

Hockley, N. S., & Polka, L. (1994). A developmental study of audiovisual speech

perception using the McGurk paradigm. *The Journal of the Acoustical Society of America*, *96*(5), 3309. https://doi.org/10.1121/1.410782

Høien, T., Lundberg, I., Stanovich, K. E., & Bjaalid, I. K. (1995). Components of phonological awareness. *Reading and Writing*, *7*(2), 171-188. https://doi.org/10.1007/BF01027184

Holle, H., Gunter, T. C., Rüschemeyer, S. A., Hennenlotter, A., & Iacoboni, M. (2008). Neural correlates of the processing of co-speech gestures. *Neuroimage*, *39*(4), 2010-2024. https://doi.org/10.1016/j.neuroimage.2007.10.055

Hollich, G., Newman, R. S., & Jusczyk, P. W. (2005). Infants' use of synchronized visual information to separate streams of speech. *Child Development*, *76*(3), 598-613. https://doi.org/10.1111/j.1467-8624.2005.00866.x

Isei-Jaakkola, T. (2006). Cognition and physio-acoustic correlates - audio and audio-visual effects of a short English emotional statement: On JL2, FL2 and EL1. In Salakoski T., Ginter F., Pyysalo S., Pahikkala T. (Eds.), *Advances in Natural Language Processing* (pp. 161–173). *FinTAL 2006.* Springer, Berlin, Heidelberg. https://doi.org/10.1007/11816508_18

Jerger, S., Damian, M. F., Spence, M. J., Tye-Murray, N., & Abdi, H. (2009). Developmental shifts in children's sensitivity to visual speech: A new multimodal picture–word task. *Journal of Experimental Child Psychology*, *102*(1), 40-59. https://doi.org/10.1016/j.jecp.2008.08.002

Jesse, A., & Janse, E. (2012). Audiovisual benefit for recognition of speech presented with single-talker noise in older listeners. *Language and Cognitive Processes*, *27*(7-8), 1167-1191. https://doi.org/10.1080/01690965.2011.620335

Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, *21*(2), 260-267. https://doi.org/10.1177/0956797609357327

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science*, *218*(4577), 1138-1141. https://doi.org/10.1126/science.7146899

Kuhl, P. K., & Meltzoff, A. N. (1984). The intermodal representation of speech in infants. *Infant Behavior and Development*, *7*(3), 361-381. https://doi.org/10.1016/S0163-6383(84)80050-8

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), 705-719. https://doi.org/10.1097/AUD.0000000000000830

Lalonde, K., & Werner, L. A. (2019). Infants and adults use visual cues to improve detection and discrimination of speech in noise. *Journal of Speech, Language, and Hearing Research*, *62*(10), 3860-3875. https://doi.org/10.1044/2019_JSLHR-H-19-0106

Lalonde, K., & Werner, L. A. (2021). Development of the Mechanisms Underlying Audiovisual Speech Perception Benefit. *Brain Sciences*, *11*(1), 49-67. https://doi.org/10.3390/brainsci11010049

Liberman, A., Cooper, P., Shankweiler, D., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*, 431-461. https://doi.org/10.1037/h0020279

Liberman, I. Y., Shankweiler, D., Fischer, F. W., & Carter, B. (1974). Explicit syllable and phoneme segmentation in the young child. *Journal of Experimental Child Psychology*, *18*(2), 201-212. https://doi.org/10.1016/0022-0965(74)90101-5

Lin, Y.-H. (2007). *The sounds of Chinese*. Cambridge University Press.

MacKain, K., Studdert-Kennedy, M., Spieker, S., & Stern, D. (1983). Infant intermodal speech perception is a left-hemisphere function. *Science*, *219*(4590), 1347-1349. https://doi.org/10.1126/science.6828865

MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, *21*(2), 131-141. https://doi.org/10.3109/03005368709077786

Magnotti, J. F., Mallick, D. B., Feng, G., Zhou, B., Zhou, W., & Beauchamp, M. S. (2015). Similar frequency of the McGurk effect in large samples of native Mandarin Chinese and American English speakers. *Experimental Brain*

Research, *233*(9), 2581-2586. https://doi.org/10.1007/s00221-015-4324-7

Mallick, D. B., Magnotti, J. F., & Beauchamp, M. S. (2015). Variability and stability in the McGurk effect: Contributions of participants, stimuli, time, and response type. *Psychonomic Bulletin & Review*, 1–9. https://doi.org/10.3758/s13423-015-0817-4

Mann, V. A. (1986). Phonological awareness: The role of reading experience. *Cognition*, *24*(1-2), 65-92.

Marques, L. M., Lapenta, O. M., Costa, T. L., & Boggio, P. S. (2016). Multisensory integration processes underlying speech perception as revealed by the McGurk illusion. *Language, Cognition and Neuroscience*, *31*(9), 1115-1129. https://doi.org/10.1016/0010-0277(86)90005-3

Massaro, D. W. (1987). Psychophysics versus specialized processes in speech perception: An alternative perspective. *The Psychophysics of Speech Perception*, *39*, 46–65. https://doi.org/10. 1007/978-94-009-3629-4_3

Massaro, D. W., Cohen, M. M., & Smeele, P. M. (1995). Cross-linguistic comparisons in the integration of visual and auditory speech. *Memory & Cognition*, *23*(1), 113-131. https://doi.org/10.3758/BF03210561

Massaro, D. W., & Jesse, A. (2007). Audiovisual speech perception and word recognition. In M. G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics*

(pp. 19-35). Oxford: Oxford University Press.

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

McBride-Chang, C., Bialystok, E., Chong, K. K., & 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

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

Meissner, C.A. & Brigham, J.C. (2001). Thirty years of investigating the own-race bias memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, 7(1), 3-35. https://doi.org/10.1037/1076-8971.7.1.3

Morais, J., Bertelson, P., Cary, L., & Alegria, J. (1986). Literacy training and speech segmentation. *Cognition*, *24*(1-2), 45-64. https://doi.org/10.1016/0010-0277(86)90004-1

Nagels, A., Kircher, T., Steines, M., & Straube, B. (2015). Feeling addressed! The role of body orientation and co-speech gesture in social communication. *Human Brain Mapping*, *36*(5), 1925-1936. https://doi.org/10.1002/hbm.22746

National Reading Panel (US), National Institute of Child Health, & Human Development (US). (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*. National Institute of Child Health and Human Development, National Institutes of Health.

Patel, T. K., Snowling, M. J., & de Jong, P. F. (2004). A cross-linguistic comparison of children learning to read in English and Dutch. *Journal of Educational Psychology*, *96*(4), 785-797. https://doi.org/10.1037/0022-0663.96.4.785

Patterson, M. L., & Werker, J. F. (1999). Matching phonetic information in lips and voice is robust in 4.5-month-old infants. *Infant Behavior and Development*, *22*(2), 237-247. https://doi.org/10.1016/S0163-6383(99)00003-X

Patterson, M. L., & Werker, J. F. (2002). Infants' ability to match dynamic phonetic and gender information in the face and voice. *Journal of Experimental Child Psychology*, *81*(1), 93-115. https://doi.org/10.1006/jecp.2001.2644

Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science*, *6*(2), 191-196. https://doi.org/10.1111/1467-7687.00271

Pearl, D., Yodashkin-Porat, D., Katz, N., Valevski, A., Aizenberg, D., Sigler, M., ... & Kikinzon, L. (2009). Differences in audiovisual integration, as measured by

McGurk phenomenon, among adult and adolescent patients with schizophrenia and age-matched healthy control groups. *Comprehensive Psychiatry*, *50*(2), 186-192. https://doi.org/10.1016/j.comppsych.2008.06.004

Read, C., Yun-Fei, Z., Hong-Yin, N., & Bao-Qing, D. (1986). The ability to manipulate speech sounds depends on knowing alphabetic writing. *Cognition*, *24*(1-2), 31-44. https://doi.org/10.1016/0010-0277(86)90003-X

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics*, *59*(3), 347-357. https://doi.org/10.3758/BF03211902

Ross, L. A., Molholm, S., Blanco, D., Gomez-Ramirez, M., Saint-Amour, D., & Foxe, J. J. (2011). The development of multisensory speech perception continues into the late childhood years. *European Journal of Neuroscience*, *33*(12), 2329-2337. https://doi.org/10.1111/j.1460-9568.2011.07685.x

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. *Perception & Psychophysics*, *59*(1), 73-80. https://doi.org/10.3758/BF03206849

Sekiyama, K., & Burnham, D. (2004). Issues in the development of auditory-visual speech perception: Adults, infants, and children. In *Eighth International Conference on Spoken Language Processing INTERSPEECH-2004,* 1137-1140.

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., & Tohkura, Y. I. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *The Journal of the Acoustical Society of America*, *90*(4), 1797-1805. https://doi.org/10.1121/1.401660

Sekiyama, K., & Tohkura, Y. I. (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

Sekiyama, K., Burnham, D., Tam, H., and Erdener, D. (2003). Auditory-visual speech perception development in Japanese and English speakers. *Proceedings of the International Conference on Auditory-Visual Speech Processing*, St. Jorioz, France, pp. 61–66.

Shigeno, S. (2000). Influence of vowel context on the audio-visual speech perception of voiced stop consonants. *Japanese Psychological Research*, *42*(3), 155-167.

https://doi.org/10.1111/1468-5884.00141

Shinozaki, J., Hiroe, N., Sato, M. A., Nagamine, T., & Sekiyama, K. (2016). Impact of language on functional connectivity for audiovisual speech integration. *Scientific Reports*, *6*(1), 1-13. https://doi.org/10.1038/srep31388

Snow, C. E., Burns, M. S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children.* Washington, DC: National Academy Press.

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *Journal of the Acoustical Society of America*, *26*(2), 212. https://doi.org/10.1121/1.1907309

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: psychology of lipreading* (pp. 3-51). Hillsdale, NJ: Erlbaum.

Teinonen, T., Aslin, R. N., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition*, *108*(3), 850-855. https://doi.org/10.1016/j.cognition.2008.05.009

Traunmüller, H., & Öhrström, N. (2007). Audiovisual perception of openness and lip rounding in front vowels. *Journal of Phonetics*, *35*(2), 244-258. doi: https://doi.org/10.1016/j.wocn.2006.03.002

Treiman, R., & Zukowski, A. (1991). Levels of phonological awareness. In S. A. Brady & D. P. Shankweiler (Eds.), *Phonological processes in literacy: A tribute to Isabelle Y. Liberman* (pp. 67-83). Hillsdale, NJ: Erlbaum.

Tremblay, C., Champoux, F., Voss, P., Bacon, B. A., Lepore, F., & Théoret, H. (2007). Speech and non-speech audio-visual illusions: a developmental study. *PLOS ONE*, *2*(8), e742. https://doi.org/10.1371/journal.pone.0000742

Tse, K. H., & Herrup, K. (2017). Re-imagining Alzheimer's disease–the diminishing importance of amyloid and a glimpse of what lies ahead. *Journal of Neurochemistry*, *143*(4), 432-444. https://doi.org/10.1111/jnc.14079

Tsujimura, N. (2013). *An introduction to Japanese linguistics*. John Wiley & Sons, Blackwell.

Valkenier, B., Duyne, J. Y., Andringa, T. C., & Baskent, D. (2012). Audiovisual perception of congruent and incongruent Dutch front vowels. *Journal of Speech, Language, and Hearing Research*, *55*(6), 1788-1801. https://doi.org/10.1044/1092-4388(2012/11-0227)

Wang, R. (2018). *Audiovisual perception of Mandarin lexical tones* (Doctoral dissertation, Bournemouth University).

Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastián-Gallés, N., & Werker, J. F. (2007). Visual language discrimination in infancy. *Science*,

316(5828), 1159-1159. https://doi.org/10.1126/science.1137686

Wightman, F., Kistler, D., & Brungart, D. (2006). Informational masking of speech in children: Auditory-visual integration. *The Journal of the Acoustical Society of America*, *119*(6), 3940-3949. https://doi.org/10.1121/1.2195121

Wimmer, H., Landerl, K., Linortner, R., & Hummer, P. (1991). The relationship of phonemic awareness to reading acquisition: More consequence than precondition but still important. *Cognition*, *40*(3), 219-249. https://doi.org/10.1177/0956797615597671

Wu, Y. C., & Coulson, S. (2015). Iconic gestures facilitate discourse comprehension in individuals with superior immediate memory for body configurations. *Psychological Science*, *26*(11), 1717-1727.

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

Zsiga, E. C. (2013). *The sounds of language: an introduction to phonetics and phonology*. Wiley-Blackwell.