Acoust. Sci. & Tech. **41**, 1 (2020)

INVITED REVIEW

Influence of language backgrounds on audiovisual speech perception across the lifespan

Kaoru Sekiyama*

Graduate School of Advanced Integrated Studies in Human Survivability, Kyoto University, Higashi Ichijo Bldg., 1 Yoshida Naka-adachi-cho, Sakyo-ku, Kyoto, 606–8306 Japan

Abstract: Speech perception is often audiovisual, as demonstrated in the McGurk effect: Auditory and visual speech cues are integrated even when they are incongruent. Although this illusion suggests a universal process of audiovisual integration, the process has been shown to be modulated by language backgrounds. This paper reviews studies investigating inter-language differences in audiovisual speech perception. In these examinations with behavioral and neural data, it is shown that native speakers of English use visual speech cues more than those of Japanese, with different neural underpinnings for the two language groups.

Keywords: McGurk effect, Language background, Japanese, Development, Aging, Event-related potential, Functional magnetic resonance imaging

PACS number: 43.71m [doi:10.1250/ast.41.37]

1. INTRODUCTION

Speech perception is often audiovisual. A good demonstration of audiovisual speech perception is the McGurk effect, a perceptual fusion of incongruent visual and auditory speech [1]. For example, a video of auditory /ba/ dubbed onto a /ga/-mouth is often heard as "da," illustrating a strong tendency to integrate concurrent auditory and visual speech irrespective of their congruity.

Although such an illusion suggests a universal process of audiovisual integration, studies conducted in our laboratory have revealed that it is in fact modulated by the perceiver's language background. Here, I will review these studies.

2. BEHAVIORAL FINDINGS

2.1. Modulation by Native Language

We started by testing native Japanese speakers (JSs) to replicate the McGurk effect. However, using highly intelligible speech stimuli articulated by JSs, the occurrence of the McGurk effect was highly limited in a noise-free condition. In contrast, the effect was clearly replicated in a noise-added condition [2]. Subsequently, we compared native English speakers (ESs) and JSs by using both native and non-native speech stimuli and found that JSs are less susceptible to the McGurk effect than are ESs [3]. Thus, our cross-linguistic study revealed that the visual influence on speech perception is relatively small for JSs compared to ESs.

It is difficult to determine the reason for interlanguage differences, but there are some possible explanations. Linguistically, compared to English, Japanese visual speech is less informative, and the phonemic space is less crowded. Thus, Japanese phonemes are easier to identify by the auditory modality alone. From a cultural perspective, Asians tend to look at the eyes, whereas Westerners at the mouth. There is a possibility that these factors are intertwined, as language and culture influence each other.

2.2. Different Developmental Courses

We examined the onset of interlanguage differences in a cross-linguistic developmental study [4]. By testing four age groups (6-, 8-, and 11-year-olds, and university students) of JSs and ESs, we found that the onset of differences is between 6 and 8 years of age. The degree of visual influence was low and equivalent for 6-year-old JSs and ESs. This increased with age for ESs, but remained the same for JSs. These results suggest that the English language environment features characteristics to promote the use of visual speech cues as the speaker become a more skilled language user. The crowded phoneme space and more informative visual speech (with more visemes and salient mouth movements) might explain this. Consistent with this idea of language-specific development, we also found that adult and older children ESs processed visual speech for lipreading relatively faster than auditory speech; no such inter-modal differences were found in JSs' reaction times.

*e-mail: sekiyama.kaoru.8a@kyoto-u.ac.jp

In our current project, we are investigating the development of audiovisual speech perception in Japanese infants and toddlers [5]. Using eye-tracking during hearing and seeing continuous natural speech, we are finding a developmental course that is different from that of English-learning infants [6].

2.3. Aging-related Increase in the Use of Visual Speech

Hearing ability decreases with age. It is likely that this leads to greater dependence on visual speech in face-to-face communication; however, investigating this presents significant challenges. One problem is that native ESs may already be too visually dependent by early adulthood to show further aging-related increases. Thus, testing older adult JSs is one solution. We investigated aging-related increases in the use of visual speech within JSs. By controlling age-related differences in the hearing threshold, we found an enhanced McGurk effect in adults in their 60s relative to young adults [7].

3. NEURAL CORRELATES

We also investigated the neural correlates of interlanguage differences in the use of visual speech. To accomplish this, we employed both event-related potential (ERP) and functional magnetic resonance imaging (fMRI). In these studies, we used congruent auditory-visual stimuli rather than the McGurk stimuli because congruency maximizes neural responses (especially in latency) associated with multisensory integration [8].

3.1. Visual Speech Speeds Up Auditory Speech Processing for ESs But Not JSs

We used ERPs to investigate the temporal characteristics of auditory-visual speech processing in JSs and ESs. We examined several early ERP peaks (N1, P2 within 100–200 ms after auditory speech onset) in response to auditory-only and auditory-visual speech (congruent /ba/ or /ga/) during a passive perception task [9]. The ERP P2 amplitude indicated that ESs process multisensory speech more efficiently than auditory-only speech; JSs exhibited the opposite pattern. These results were consistent with behavioral results for the same stimuli. During a syllable identification task, congruent mouth movement shortened ESs' reaction time; the opposite was observed for JSs. In addition, eye-tracking data during audiovisual speech perception revealed a gaze bias to the mouth for ESs but not for JSs, especially before audio onset. These results indicate that ESs pay more attention to the mouth before auditory speech commences. Thus, visual speech has a greater influence on subsequent auditory speech for ESs than it does for JSs.

3.2. Functional Connectivity Reveals Earlier Visual Input to ESs' Auditory Cortex

To explore which regions of the brain are related to interlanguage differences in the use of visual speech, we used fMRI to compare JSs and ESs during speech perception [10]. We analyzed the functional connectivity of four regions of interest (ROI-based analysis): the primary auditory cortex (A1), primary visual cortex, motion-sensitive middle temporal visual area, and superior temporal sulcus (STS). The STS is the region traditionally considered the site of multisensory integration for speech. The results revealed very clear group differences: ESs showed A1-centered connectivity, whereas JSs showed STS-centered connectivity. The results suggest that the convergence of visual and auditory information is as early as A1 for ESs, but only at STS for JSs. Such an early multisensory convergence in ESs might cause the strong tendency to integrate visual and auditory speech.

4. CONCLUSION

We now have converging evidence indicating that ESs use visual speech more than JSs. These findings signal the profound impact that language backgrounds have on our perceptual system. Future research should study other languages to help deepen our understanding of this topic.

REFERENCES

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

[2] K. Sekiyama and Y. Tohkura, "McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility," *J. Acoust. Soc. Am.*, **90**, 1797–1805 (1991).

[3] K. Sekiyama and Y. Tohkura, "Inter-language differences in the influence of visual cues in speech perception," *J. Phon.*, **21**, 427–444 (1993).

[4] K. Sekiyama and D. Burnham, "Impact of language on development of auditory-visual speech perception," *Dev. Sci.*, **11**, 303–317 (2008).

[5] S. Hisanaga, R. Mugitani and K. Sekiyama, "Selective attention to the mouth of a talking face in Japanese-learning infants and toddlers," Paper presented at ICIS, Philadelphia, June (2018).

[6] D. J. Lewkowicz and A. Hansen-Tift, "Infants deploy selective attention to the mouth of a talking face when learning speech," *Proc. Natl. Acad. Sci. USA*, **109**, 1431–1436 (2012).

[7] K. Sekiyama, T. Soshi and S. Sakamoto, "Enhanced audiovisual integration with aging in speech perception: A heightened McGurk effect in older adults," *Front. Psychol.*, **5**, 322 (2014).

[8] V. van Wassenhove, K. W. Grant and D. Poeppel, "Visual speech speeds up the neural processing of auditory speech," *Proc. Natl. Acad. Sci. USA*, **102**, 1181–1186 (2005).

[9] S. Hisanaga, Sekiyama, T. Igasaki and N. Murayama, "Culture/language modulates brain and gaze processes in audiovisual speech perception," *Sci. Rep.*, **6**, 35265 (2016).

[10] J. Shinozaki, N. Hiroe, M.-A. Sato, T. Nagamine and K. Sekiyama, "Impact of language on functional connectivity for audiovisual speech integration," *Sci. Rep.*, **6**, 31388 (2016).