Audiovisual speech perception in Japanese and English: Inter-language differences examined by event-related potentials

Satoko Hisanaga [1], *Kaoru Sekiyama* [2], *Tomohiko Igasaki* [3], *and Nobuki Murayama* [3]

[1]Graduate School of Social and Cultural Sciences, Kumamoto University, Japan
[2]Faculty of Letters, Kumamoto University, Japan
[3]Graduate School of Science and Technology, Kumamoto University, Japan
[1]satoko.hisanaga@gmail.com, [2]sekiyama@kumamoto-u.ac.jp

Abstract

By using event-related potentials (ERPs, Experiment 2) and reaction times (RTs, Experiment 1), the present study examined interlanguage differences between Japanese and English in audiovisual speech perception [1, 2, 3, 4]. There were an auditory-only (AO) and a congruent auditory-visual (AV) conditions. In Experiment 1, RTs showed opposite tendencies between English-language (EL) and Japanese-language (JL) groups for the AO-AV relationship: the additional congruent visual information speeded up the speech perception processes for the EL group, but it slowed down the processes for the JL group. Thus, the visual influence was promoting for the EL but disturbing for the JL group. In Experiment 2, different ERP patterns were found between the EL and JL groups: Whereas the visual influence was sustained (maintained from N1 to P2) in the EL group, the influence was transient (limited only to N1) in the JL group. The ERPs and RTs data were both consistent with the reported interlanguage differences that the JL perceivers use visual information to the less extent than the EL perceivers do.
Index Terms: native language, audiovisual speech perception, event-related potentials (ERPs)

1. Introduction

In face-to-face communication, what we hear is influenced by visual information of articulatory movements, as demonstrated in the McGurk effect [5], in which discrepant auditory and visual information results in fused percept. For example, audio /ba/ and video /ga/ will generate fused "da" percept.

The size of the McGurk effect is known to depend on age, native language and some other environmental factors. Interlanguage differences have been found between JL and EL adults: the size of the McGurk effect is smaller for the JL adults than for the EL adults [1, 2, 3, 4]. In the EL perceivers, developmental changes are reported in the size of the McGurk effect: the EL young children are less influenced by visual information than adults are [5, 6]. However, there is no such developmental change in the visual influence in the JL perceivers. Interlanguage differences between the JL and EL in the development of auditory-visual speech perception were investigated by [4]. They found that the JL group did not show a developmental increase of the visual influence from 6 years to adulthood whereas a developmental increase was observed in the EL counterparts.

These authors suggested that the JL adults' smaller visual influence over the EL adults' might be due to differences in the speed of unimodal processing. RTs for the JL adults were about the same in the AO and VO (visual-only) conditions, whereas RTs for the EL adults showed faster VO processing over AO processing. The present study examined neural processes of such interlanguage differences by ERPs.

ERPs during auditory-visual speech perception in English were previously examined [e.g. 7]. This study found the visual influence in relatively early peaks of ERPs such as the N1 (negative 100-millisecond peak) and P2 (positive 200-millisecond peak): the N1/P2 amplitude and latency reduced in the AV condition compared to the AO condition. Similar tendencies were also reported in other studies [8, 9]. If such a reduction in the N1/P2 amplitude and latency represents the visual influence, we may capture the interlanguage differences between Japanese and English by looking at these ERP measures. Therefore, we examined the N1 and P2 during auditory-visual speech perception in the JL and EL adults. Experiment 1 examined RTs and experiment 2 examined ERPs in the AO and AV conditions.

2. Experiment 1: Behavioral study

2.1. Method

2.1.1. Participants

Sixteen JL (mean age, 21.4 years; range, 20-26 years; 6 males and 10 females) and five EL adults (mean age, 20.4 years; range, 19-22 years; 3 males and 2 females) participated. They were students at Kumamoto University (JL participants) or international students visiting Kumamoto University (EL participants). All were right handed. They all had normal hearing and normal or corrected-to-normal vision.

2.1.2. Stimuli

Stimuli were prepared by using two female talkers (a native Japanese and a native English talker) who uttered syllables /ba/ and /ga/. The movie clips were edited for two conditions (AO and AV conditions) by Adobe Premiere. Both talkers were included in a given block for each condition. An AO stimulus consisted of auditory speech of /ba/ or /ga/ and a visual fixation point. The AV stimuli consisted of matching auditory and visual speech. Video was digitized by 29.97 frames per s in 640 × 480-pixel. Sound was digitized by 16-bit 44k Hz resolution, and was stored in stereo (speech signals in one channel and trigger tone signals in the other channel, see details in Experiment 2).

2.1.3. Procedure

The auditory stimuli were presented through a loudspeaker (AIWA SC-B10) at 65dB with a band noise (300-12000 Hz). The SN ratio was +13dB. The noise was added to mask machine noise. The visual stimuli were presented at the center of a 15-inch SONY SDM-S51 monitor and the loudspeaker was placed above the monitor. Participants sat approximately 90cm from the monitor. One experimental block consisted of 40 trials (/ba/ and /ga/ presented 20 times per block), and there were two conditions (AO and AV conditions), resulting in a total of 80 trials. The presentation order was counterbalanced across participants. In both conditions, there were two response alternatives "ba" and "ga". The participants were instructed to look at the display, listen to the sound, make a decision whether it is "ba" or "ga", and press one of two buttons. RTs were measured as the time from the audio onset to the button press (see Figure 1). The onset of the next stimulus was 1500-ms after the button press. The experiment took about 10 minutes per participant.

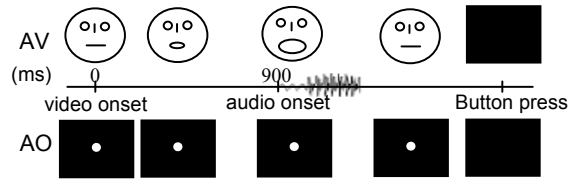

Figure 1: A trial flow of movie
The talker's face appeared 900-ms before the audio onset.

2.2. Results

The RT data were analyzed by ANOVA [Between subjects: language group (2) × Within subjects: condition (2) × talker (2) × syllable (2)]. There was a significant interaction between language group and condition [F (1, 19) =15.37, p<.001] (Figure 2). RTs of the JL group significantly increased in the AV condition compared with the AO condition [F (1, 19) =6.63, p<.05] (Figure 2-a). In contrast, RTs of the EL group were significantly shortened in the AV condition compared with the AO condition [F (1, 19) =8.81, p<.01] (Figure 2-b). In addition, RTs of the AV condition in the EL group were shorter than in the JL group [F (1, 38) =4.28, p<.05]. (*= p<.05, **= p<.01)

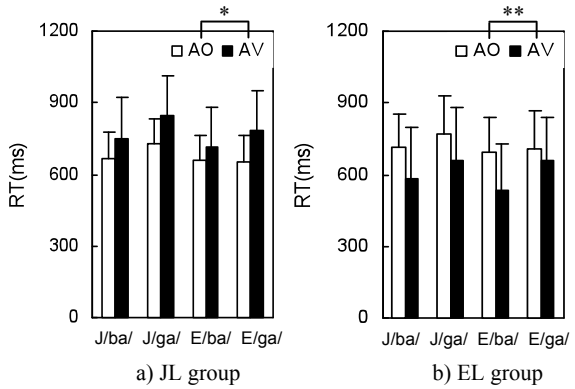

Figure 2: RTs for two language groups
RTs for the JL group are shown in the left panel (a) and RTs for the EL group in the right panel (b). The X axis indicates the type of stimulus. The error bars show standard deviation.
Audio onset=0ms

2.3. Discussion

Experiment 1 examined RTs in the AO and AV conditions. The results showed opposite tendencies between the JL and EL groups for the AO-AV relationship. The additional visual information speeded up the speech perception processes for the EL group, but it slowed down the processes for the JL group. These results indicate language-specific weightings of auditory and visual information, that is, the visual weight is greater than auditory weight in the EL perceivers' weighting, and the weighting is opposite in the JL perceivers. This finding is consistent with a previous result by [4] that the EL group was faster in the VO condition than in the AO condition but the JL group did not show such a visual advantage. These language-specific weightings of different sensory cues may be related to the weaker McGurk effect in the JL group than in the EL group [1, 2, 3].

3. Experiment 2: ERPs study

3.1. Method

3.1.1. Participants

From the participants of Experiment 1, eight JL (mean age, 22.0 years; range, 21-26 years; 3 males and 5 females) and three EL adults (mean age, 19.7 years; range, 19-22 years; 2 males and 1 female) also participated in Experiment 2. They all had normal hearing and normal or corrected-to-normal vision.

3.1.2. Stimuli

Stimuli were identical to those used in Experiment 1.

3.1.3. Procedure

The procedure was similar to that of Experiment 1, but no responses were requested in Experiment 2 in order to rule out motor-related potentials during ERP measurement. A trial consisted of a 2000-ms stimulus period and 1500-ms blank interval. The participants were instructed to look at the display, listen to the sound, and make a decision whether it is "ba" or "ga" without any overt responses. Electroencephalographic (EEG) recordings were conducted during each experimental block. Experiment 2 consisted of ten AO blocks and ten AV blocks. Each block had 40 trials (each of /ba/ and /ga/ was presented 20 times per block). In total, there were 800 trials. The AO and AV blocks were alternated, and the presentation order was counterbalanced across participants. The experiment took about 270 minutes per participant, and EEG recordings were divided into three days.

3.1.4. EEG Recording

Neurofax EEG-1100 (Nihon Kohden, Tokyo) and an electro-cap (International, Inc. Eaton, Ohio USA: 10/20 system) were used. EEG data were acquired (sampling rate,500 Hz) from 19 channels of the electro-cap (Fp1, Fp2, F7, F3, Fz, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1, O2). Recording electrodes were referred to the linked earlobes (A1+A2) and ground electrodes were placed on Fpz and nasion. Trigger signals (30-ms) had been inserted on an audio track so that each of them was synchronized with the onset of the speech stimuli (Figure 3). The trigger signals were recorded on the 20th channel of the EEG, and the speech signals on the 21st channel to ensure synchronization.

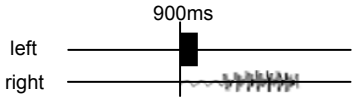

Figure 3: The audio signals

3.1.5. *Analysis*

Average ERPs were processed with the trigger of audio onset. Baseline was corrected on 200ms before an audio onset per participant. ERP components (N1 and P2) at Cz (Figure 4) were analyzed by ANOVA.

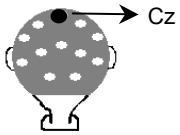

Figure 4: The vertex potential is Cz.

3.2. Results

ERP patterns were different between the native and non-native stimuli. Therefore, ERP data were separated for native and non-native stimuli and analyzed by ANOVA [Between subjects: language group (2) × Within subjects: condition (2) × syllable (2)].

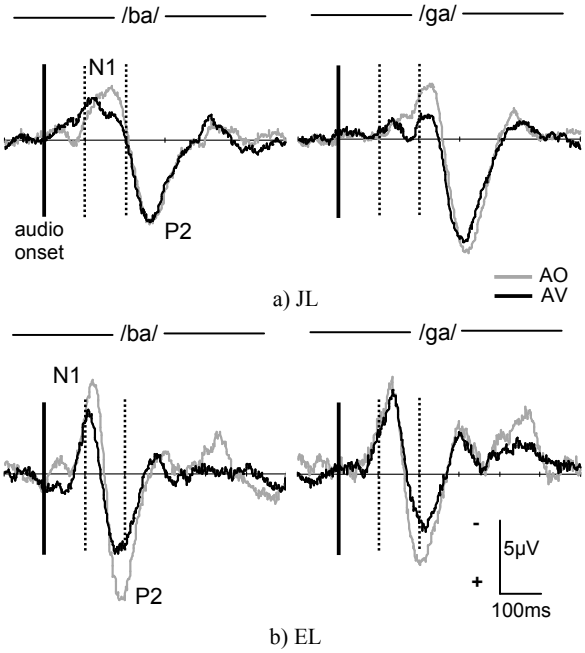

Figure 5: Averaged ERPs at Cz for native stimuli.
(a) JL group perceiving Japanese stimuli, (b) EL group perceiving English stimuli.

< The native stimuli >
N1 amplitude was reduced in the AV condition compared to the AO condition in both language groups (Figure 5, 6a). The ANOVA on N1 amplitude confirmed this: The main effect of the condition was significant [F (1, 9) =11.62, p<.01] and the "condition × language group" interaction was not significant. Similarly, N1 latency was significantly shorter in the AV condition compared to the AO condition [F (1, 9) =22.30, p<.005] (Figure 6b).

Whereas N1 parameters were similar for JL and EL groups, there were interlanguage differences in P2. In the ANOVA on P2 amplitude, in addition to the main effect of the condition [F (1, 9) =11.16, p<.01], there was a significant interaction between language group and condition [F (1, 9) =5.26, p<.05] (Figure 5, 6c). P2 amplitude of the EL group was significantly reduced in the AV condition compared to the AO condition [F (1, 9) =15.88, p<.005]. On the other hand, P2 amplitudes of the JL group did not differ between the AO and AV conditions. In both language groups, P2 latencies were not significantly different between the AO and AV conditions (Figure 5).

(*= p<.05 ; **= p<.01 ; ***= p<.005)

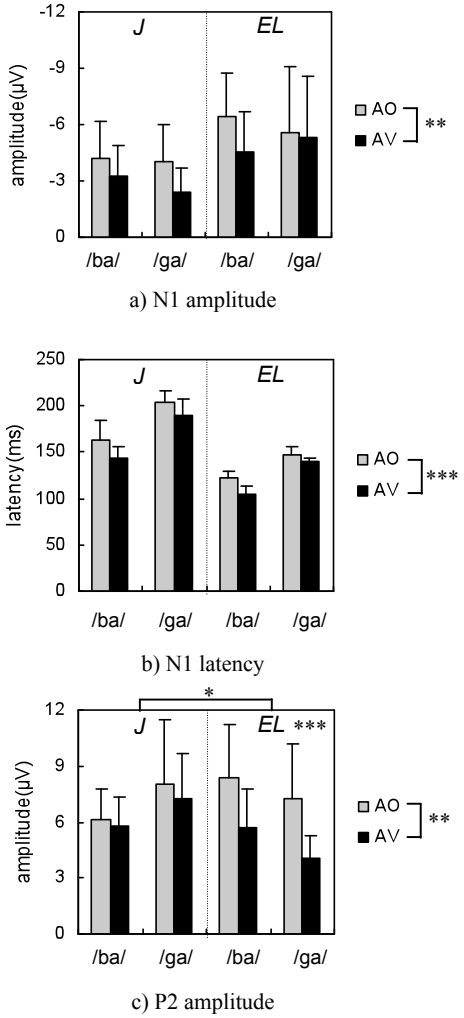

Figure 6: Averages of ERP components at Cz in JL (left) and EL (right) groups for native stimuli.
(a) N1 amplitude, (b) N1 latency, and (c) P2 amplitude. The X axis indicates the type of stimulus. The error bars show standard deviation. Audio onset=0msec

< The non-native stimuli >
There were no significant main effects or no interactions in any of amplitudes and latencies of N1 and P2 (Figure 7).

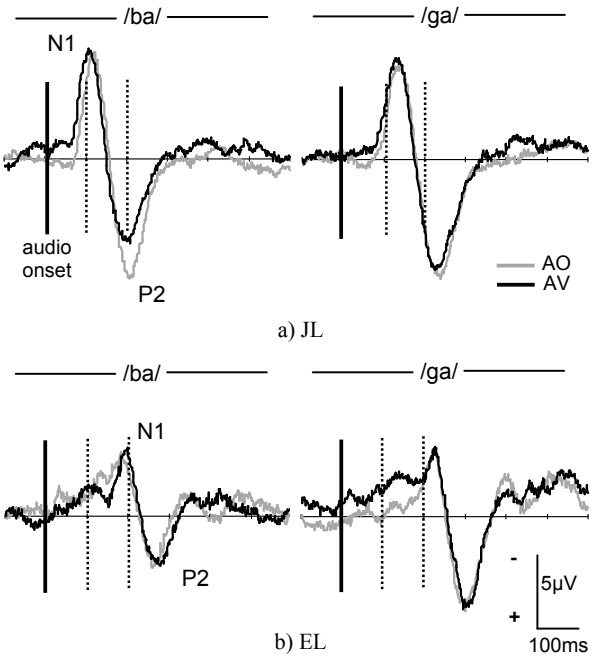

Figure 7: Averaged ERPs at Cz for non-native stimuli.

3.3. Discussion

Experiment 2 examined ERPs in the AO and AV conditions. Our results showed that ERP modulations due to additional visual information were evident only in native stimuli for both EL and JL participants. In the native stimuli, the results for the EL group were consistent with the findings by [7], in terms of the reduction in the N1 amplitude, N1 latency, and P2 amplitude. The JL group also showed the reduction in the N1 amplitude and latency, but no reduction was observed in the P2 amplitude. Thus, this study revealed interlanguage differences between JL and EL adults in the P2. These results indicate that the visual influence is sustained (maintained from N1 to P2) in the EL group whereas the influence is transient (limited only to N1) in the JL group.

4. General discussion

In this study, Experiment 1 examined RTs and Experiment 2 examined ERPs in the AO and congruent AV conditions. In Experiment 1, RTs showed that the additional visual information speeded up the speech perception processes for the EL group, but it slowed down the processes for the JL group. Thus, the visual influence was promoting for the EL but disturbing for the JL group. This result is similar to the previous finding on RTs that the EL adults are faster in the VO than in AO condition whereas the JL adults are equally fast in the AO and VO conditions [4]. These results consistently indicate that the visible speech plays a role as a priming cue for the EL adults' speech perception, but not for the JL adults' perception. It is reasonable that such a visual priming effect results in the greater McGurk effect in the EL adults as reported previously [2, 3, 4].

In Experiment 2, ERPs showed that the visual influence was maintained from N1 to P2 in the EL group but it was limited only to N1 in the JL group. It is plausible that the EL participants' more maintained visual processing is a cause of their greater McGurk effect reported previously. The maintained visual processing in the EL group may be also related to the promoting visual effect observed in RTs for the EL group in Experiment 1.

As compared with a previous study on ERPs during audiovisual speech perception in English [7], our ERP results for the equivalent condition (EL group perceiving the native stimuli) were similar in terms of the reduction in N1 amplitude, N1 latency, and P2 amplitude. One difference was a lack of the reduction in P2 latency in our EL participants. Currently, the number of the EL participants is smaller than that of the JL participants and additional ERP recordings with EL participants are under way. With more data added from EL participants, we will possibly replicate the reduction in P2 latency as well.

For the non-native stimuli, our ERP results showed no amplitude or latency modulation for N1 or P2, irrespective of the participants' language. The reason for this sharp difference between the native and non-native stimuli is not clear, but it indicates that we process visible speech more automatically for the native speech stimuli than for non-native ones, to the extent that the visual information modulates early components of neural processes, at least for congruent AV speech.

For the native stimuli, the N1 modulations were found in both language groups. It was in the P2 that the present study demonstrated interlanguage differences. The results indicate that the EL participants are still visually influenced at 200 ms from the audio onset with the amplitude reduction of P2 in the AV condition compared with the AO condition, but the JL participants are already free from the visual influence by then.

Although the functional significance of the AV-AO amplitude reduction of the early ERP components is not clear, one interpretation is that a portion of cognitive resource is spared for non-auditory processing in the AV condition. If so, our ERP data demonstrate that the EL adults spare the cognitive resource for visual speech processing more continuously than the JL adults do in AV speech perception.

Another interpretation of the amplitude reduction, which does not rule out the above, may be the proportion of time-locked activation. Although the ERP waveform for an auditory stimulus is sharply time-locked to the audio onset, the waveform for a VO speech stimulus usually shows no time-locked nature [7, 8; also in our preliminary study]. It may be due to the fact that participants pay attention to various regions of the talker's face at various timing in visual speechreading. Therefore, the more the participant processes visible speech, the more the averaged ERP waveform loses the time-locked nature, resulting in the more reduction in amplitude.

In this study, the data analyses were based on only data from the Cz electrode, following a previous study [7]. In contrast, there is a recent ERP study using the data from all the electrodes for current density reconstruction (CDR) analyses [10] that provide spatial information of neural processes. It was found that the congruent AV condition results in temporally more maintained and spatially more spread activation compared with the AO condition. It is of interest to see if JL participants' activation, in response to the additional visual information, shows the same amount of temporal and spatial extension as the EL participants.

5. Acknowledgements

We thank to Hironori Kikuchi for his technical support in data analyses, and the volunteers who participated in this study. This study was supported by a Grant-in-Aid for Scientific Research from Japan Society for the Promotion of Science (21243040) to KS.

6. References

[1] K. Sekiyama, and Y. Tohkura, "McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility," *Journal of the Acoustical, Society of America*, vol. 90, pp. 1797-1805, 1991.

[2] K. Sekiyama, and Y. Tohkura, "Inter-language differences in the influence of visual cues in speech perception," *Journal of Phonetics*, vol. 21, pp. 427-444, 1993.

[3] P. K. Kuhl, M. Tuzaki, Y. Tohkura, and A. N. Meltzoff, "Human processing of auditory-visual information: Potential for multimodal human-machine interfaces," in *Proceedings of International Conference on Spoken Language'94*, pp. 539-542, 1994.

[4] K. Sekiyama, and D. Burnham, "Impact of language on development of auditory-visual speech perception," *Developmental Science*, vol. 11, pp. 306-320, 2008.

[5] H. McGurk, and J. MacDonald, "Hearing lips seeing voices," *Nature*, vol. 264, pp. 746-748, 1976.

[6] D. W. Massaro, L. A. Thompson, B. Barron, and E. Laren, "Developmental changes in visual and auditory contributions to speech perception," *Journal of Experimental Child Psychology*, vol. 41, pp. 93-113, 1986.

[7] V. van Wassenhove, K. W. Grant, and D. Poeppel, "Visual speech speeds up the neural processing of auditory speech," *PNAS, 25*, vol. 102, pp. 1181-1186, 2005.

[8] R. A. Reale, G. A. Calvert, T. Thesen, R. L. Jenison, H. Kawasaki, H. Oya, M. A. Howard, and J. F. Brugge, "Auditory-visual processing represented in the human superior temporal gyrus," *Neuroscience*, vol. 145, pp. 162-184, 2007.

[9] C. Davis, D. Kislyuk, J. Kim, and M. Sams, "The effect of viewing speech on auditory speech processing is different in the left and right hemispheres," *Brain Research*, vol. 1242, pp. 151-161, 2008.

[10] L. E. Bernstein, E. T. Auer Jr., M. Wagner, and C. W. Ponton, "Spatiotemporal dynamics of audiovisual speech processing," *NeuroImage*, vol. 39, pp. 423-435, 2008.