ISCA Archive
http://www.isca-speech.org/archive

AVSP 2010 – International Conference
on Audio-Visual Speech Processing
Hakone, Kanagawa, Japan
September 30–October 3, 2010

Impact of language on audiovisual speech perception examined by fMRI

Jun Shinozaki[1], *Kaoru Sekiyama*[2],*Nobuo Hiroe*[3], *Taku Yoshioka*[3], *Masa-aki Sato*[3]

[1] Department of System Neuroscience, School of Medicine, Sapporo Medical University, JAPAN
[2] Division of Cognitive Psychology, Faculty of Letters, Kumamoto University, Kumamoto, JAPAN
[3]ATR Neural Information Analysis Laboratories, Seika-cho, JAPAN
jshino@sapmed.ac.jp

Abstract

Both auditory and visual information plays an important role for audiovisual speech perception during face-to-face communication. Several behavioral studies have shown that native English speakers and native Japanese speakers behaved differently in audiovisual speech perception. We hypothesized that there would be differences in neural processing between native English speakers and native Japanese speakers. Twenty-two English language speakers and 22 Japanese speakers watched talker's face and listened to talker's speaking during functional magnetic resonance imaging (fMRI) scanning. The lateral occipitotemporal gyrus was associated with visual domain of audiovisual speech perception in native Japanese speakers, but not in native English speakers, suggesting that language environment affects neural processes for audiovisual speech perception.
Index Terms: speech perception, language, fMRI

1. Introduction

During face-to-face communication, both auditory and visual domains play an important role in speech perception. One of striking example is the McGurk effect that an auditory syllable (phoneme) is perceived differently depending on whether mouth movements of a speaker pronouncing the same syllable or a different, incongruent syllable (McGurk & MacDonald, 1976).

Previous behavioral studies have shown that perceiver's language background affects audiovisual speech perception: Native speakers of English were more sensitive to visual influence in the McGurk effect than native speakers of Japanese (Massaro, Tsuzaki, Cohen, Gesi, & Heredia, 1993; Sekiyama & Burnham, 2008; Sekiyama & Tohkura, 1991). These behavioral results suggest that there would be differences between native English speakers and native Japanese speakers in the neural processing for the audiovisual speech perception.

To clarify the neural correlates of impact of language on audiovisual speech perception, we examined neural activity in the native English speakers and native Japanese speakers using fMRI.

Several electrophysiological and imaging studies have indicated that the superior temporal sulcus (STS), superior temporal gyrus (STG) (Beauchamp, 2005; Calvert, Campbell, & Brammer, 2000; Sekiyama, Kanno, Miura, & Sugita, 2003), and middle temporal gyrus (MTG) (Callan et al., 2004) are related to audiovisual speech perception. Transcranial magnetic stimulation (TMS) study has demonstrated that TMS of the STS reduced the likelihood of the McGurk effect (Beauchamp, Nath, & Pasalar, 2010).

Together, we hypothesized that there would be differences in neural activities and/or neural pathways in audiovisual speech perception between native English speakers and native Japanese speakers, especially in the STS, STG and MTG., as well as visual and auditory areas.

2. Materials and Methods

2.1. Subjects

Twenty-two native English speakers (average age: 23.0 years, 12 males and 10 females) and 22 native Japanese speakers (average age: 24.2 years, 12 males and 10 females) participated in this study. All subjects were right handed and had normal hearing, and normal or corrected to normal vision. All individuals gave written informed consent to participate in the study, and the protocol was approved by the local ethics committee.

2.2. Stimuli and Tasks

Stimuli were created from "ba" and "ga" uttered monosyllabically by a native English male talker and a native Japanese male talker. The utterances were recorded, digitized, and edited for audio-only (A), visual-only (V), and audiovisual (AV) stimuli. Video digitized at 29.93 frames/s in 640 × 480 pixels, and audio digitized at 44.1kHz in 16 bit. Movie duration was in average 1.72s. The position of the acoustic signals was adjusted so that the onset of acoustic energy was at 900ms from the onset of the movie file.

The A stimuli (/ba/ and /ga/ of the two talkers) consisted of only the auditory component of speech, but were combined with the talker's still face with mouth neutrally closed (i.e., no visual information about the utterance). The V stimuli ([ba] and [ga] of the two talkers) consisted of only the visual component of speech, but there was no auditory component (i.e., silent talking face). The AV stimuli consisted of the synchronized auditory and visual components.

The stimuli were presented in a blocked design by alternating three stimulus conditions and one rest condition in an AV-A-V-rest pattern. Each of the 4 stimuli ("ba" and "ga" of the two talkers) was presented twice in each block with jittered stimulus onset asynchrony (SOA; average 4.2s) in order to increase vigilance. The duration of each block was in average 32s. One functional run was composed of four AV-A-V-rest sequences. In total, three functional runs were repeated.

The subject's task was syllable identification. The subjects were instructed to watch the talker's face and listen to the talker's speaking, and were asked to report what they perceived by pressing button ("ba" or "ga") during fMRI scanning.

2.3. Procedure

Each subject lay supine on a scanner bed, with a button response device held in the left hand. Sound was delivered via MR-compatible headphones. Auditory stimuli were presented with sufficiently large sound compared to MR scanner noise. The subjects viewed visual stimuli that were back-projected

onto a screen through a built-in mirror. Foam pads were used to minimize head motion.

2.4. Image acquisition and analysis

Functional MRI experiments were conducted on a 3-Tesla whole-body scanner equipped with a 12ch phased array coil (Siemens Tim Trio, Erlangen, Germany). Functional images were obtained in a T2*-weighted gradient-echo echo-planar imaging sequence. The image acquisition parameters were as follows: repetition time (TR) = 3.0 s; echo time (TE) = 30 ms; flip angle (FA) = 80°; field of view (FOV) = 192 mm; matrix = 64 × 64; 50 interleaved axial slices with 3-mm thickness without gaps (3-mm cubic voxels). The first six images were not saved to allow for signal stabilization. For anatomic images, T1-weighted three-dimensional structural images were also obtained using a magnetization-prepared rapid-gradient echo sequence.

The fMRI data were analyzed with SPM5 (Wellcome Department of Imaging Neuroscience, University College London) implemented on MATLAB2009b (MathWorks, Natick, MA), using the principles of the general linear model. The functional images were corrected for differences in slice-acquisition timing, and were then spatially realigned to the first image of the initial run to adjust for residual head movements. The realigned images were spatially normalized to fit to a Montreal Neurological Institute (MNI) template (Evans et al., 1993) based on the standard stereotaxic coordinate system (Talairach & Tournoux, 1988). Subsequently, all images were smoothed with an isotropic Gaussian kernel of 8-mm full-width at half-maximum (FWHM). Each of the three stimulus conditions (AV, A, V) and 6 head motion parameters was separately modeled as a regressor for the first-level multi-regression analysis. This analysis was performed for each subject, to test the correlation between the MRI signals and boxcar functions convolved with the canonical hemodynamic response function. Global signal normalization was performed only between runs. Low-frequency noise was removed using a high-pass filter with a cut-off of 128 s, and serial correlation was adjusted using an AR(1) model. By applying the appropriate linear contrast to the parameter estimates, mean effect images reflecting the magnitude of correlation between the signals and the model of interest were computed. These were used for the subsequent second-level random-effect model analysis. Group-level statistical parametric maps were produced using the one-sample t-test. Two-sample t-test was calculated to clarify group differences between native English speakers and native Japanese speakers.

The results are shown at a height threshold of $p < 0.001$ (uncorrected) with an extent threshold of 6 voxels, and reported based on the coordinates of the MNI template.

3. Results

3.1. Neural representation of audiovisual speech perception in native English speakers and native Japanese speakers

Figure 1 and 2 show activated areas by AV stimuli in native English speakers and native Japanese speakers, respectively. The AV stimuli activated the bilateral STG including the primary auditory cortex, and the occipital cortex including the primary and higher order visual cortex in both native English speakers and native Japanese speakers.

Neural activity in the right precentral gyrus was due to the task using left hand.

3.2. Neural basis of visual influences on audiovisual speech perception in native English speakers and native Japanese speakers

To investigate how the visual factor affects audiovisual integration, AV condition was compared with A condition. The bilateral primary auditory cortices and visual areas were activated in AV condition than A condition significantly in both native English speakers and native Japanese speakers (Figure 3, 4).

Table 1. *Activated brain area related to Audiovisual speech perception. BA; Brodmann's Area. *; subpeak*

Region	*BA*	MNI			z-value
		x	y	z	
AV (Eng)					
*Superior Temporal Gyrus	*41*	51	–33	9	5.41
*Superior Temporal Gyrus	*22*	–66	–39	6	5.13
Cuneus		–6	–102	–6	6.58
Precentral Gyrus	*6*	45	–21	66	5.62
Medial Frontal Gyrus	*6*	–6	3	57	4.87
Cerebellum		–12	–63	–45	4.53
AV (Jpn)					
*Superior Temporal Gyrus	*41*	–42	–36	9	4.97
*Superior Temporal Gyrus	*41*	54	–24	6	5.52
Superior Temporal Gyrus	*22*	66	–36	15	6.06
Cuneus	*18*	–15	–102	3	7.10
Precentral Gyrus	*4*	33	–27	57	5.97
Medial Frontal Gyrus	*6*	9	–6	60	5.06
AV – A (Eng)					
Superior Temporal Gyrus	*41*	45	–33	9	5.12
Superior Temporal Gyrus	*13*	–42	–24	6	4.14
Middle Occipital Gyrus	*18*	–21	–96	2	5.34
Posterior Cingulate	*29*	9	–48	12	4.98
Thalamus		–21	–27	0	4.63
AV – A (Jpn)					
Superior Temporal Gyrus	*41*	–42	–33	12	5.42
Superior Temporal Gyrus	*41*	54	–24	9	5.07
Middle Occipital Gyrus	*18*	24	–93	9	5.59
Posterior Cingulate	*29*	12	–48	18	4.33
Anterior Cingulate	*32*	0	39	–3	4.08
Thalamus		–24	–27	–3	3.77
AV – A (Eng – Jpn)					
No region					
AV – A (Jpn – Eng)					
Lateral Occipitotemporal Gyrus	*37*	–54	–66	0	3.27

3.3. Impact of vision in audiovisual speech perception; differences in neural basis between native English speakers and native Japanese speakers

To identify brain regions associated with impact of language on visual influence on audiovisual speech perception between native English speakers and native Japanese speakers, we compared (AV – V) condition in native English speakers and (AV – V) condition in native Japanese speakers. The left

lateral occipitotemporal gyrus was activated stronger in Japanese speakers than English speakers, while no brain area was found by English speakers vs. Japanese speakers (Figure 5). The region of interest analysis showed that only AV raised neural activity in the left lateral occipitotemporal gyrus in native Japanese speakers, but not in native English speakers (Figure 6).

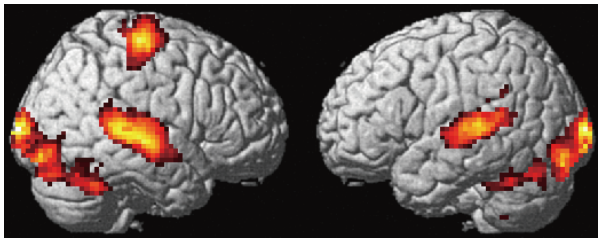

Figure 1: *Brain areas activated by audiovisual stimuli in native English speakers (cluster level p < 0.05, corrected).*

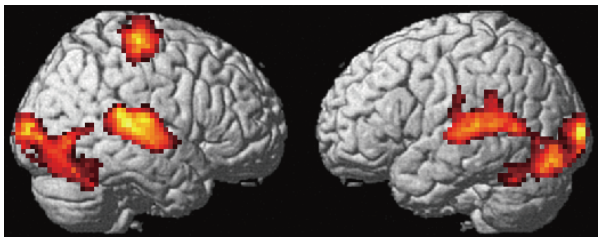

Figure 2: *Brain areas activated by audiovisual stimuli in native Japanese speakers (cluster level p < 0.05, corrected).*

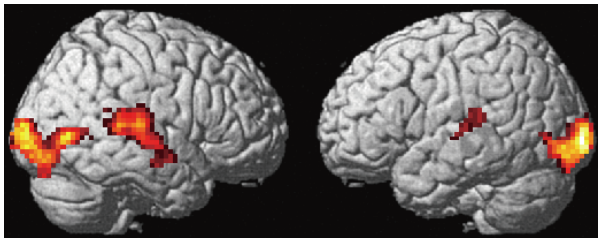

Figure 3: *The bilateral primary auditory cortex and visual area showed grater activation by AV condition than A condition in native English speakers (cluster level p < 0.05, corrected).*

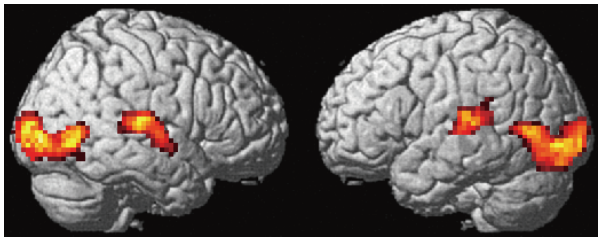

Figure 4: *The bilateral primary auditory cortex and visual area showed grater activation by AV condition than A condition in native Japanese speakers (cluster level p < 0.05, corrected).*

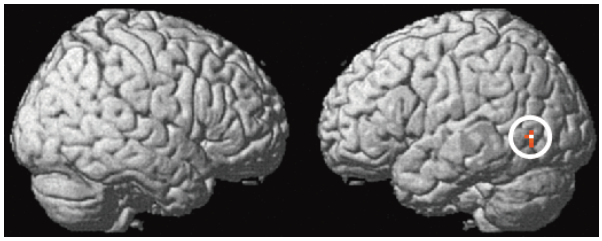

Figure5: *The influence of vision on audiovisual speech perception activated stronger in the left lateral occipitotemporal gyrus in native Japanese speakers than native English speakers (p < 0.001, uncorrected).*

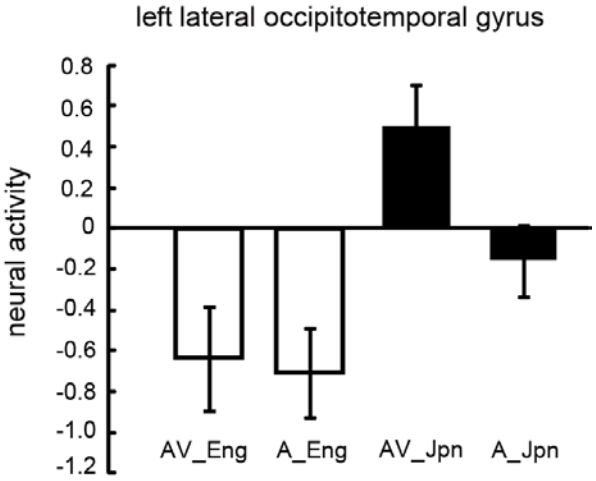

Figure6: *Neural activity in the left lateral occipitotemporal gyrus (mean plusminus S.E.M.). AV_Eng; AV in native English speakers, A_Eng; A in native English speakers, AV_Jpn; AV in native Japanese speakers, A_Jpn; A in native Japanese speakers*

4. Discussion

To clarify the neural basis related to the impact of language on audiovisual speech perception, we conducted fMRI experiment in native English speakers and native Japanese speakers. Presentation of audiovisual speech produced strong activation in the bilateral STG and visual areas both in native English speakers and native Japanese speakers. The bilateral primary auditory cortices were associated with the influence of vision on auditory perception in both groups. The impact of vision on audiovisual speech perception induced stronger activity in the left lateral occipitotemporal gyrus in the native Japanese speakers than the native English speakers.

4.1. Audiovisual speech perception

For the AV stimuli, activation in the native English speakers and native Japanese speakers was observed in the bilateral STG and visual areas. The activation in the STG including the primary auditory cortex is involved in the perception of speech sounds (Belin, Zatorre, Lafaille, Ahad, & Pike, 2000; Binder et al., 2000; Sekiyama et al., 2003). The visual areas, especially the MT area, have been implicated in visual motion processing and lip-reading (Campbell et al., 2001; Sekiyama et al., 2003; Zeki, 1993). Thus, these activities in the STG and visual areas are reasonable for audiovisual speech perception.

4.2. Influence of vision on audiovisual speech perception

To investigate the influence of vision on audiovisual speech perception, we compared AV condition with A condition in each group. In both group, visual factor increased neural activities in the bilateral primary auditory cortex and visual areas.

An effective connectivity analysis study has shown that a visual factor of audiovisual stimuli increased the stimulus salience by increasing the auditory response in the primary auditory cortex (Werner & Noppeney, 2010). Therefore, the activity in the primary auditory cortex in the present study may be associated with stimulus salience evoked by visual component in the audiovisual speech processing.

4.3. Impact of language on audiovisual speech perception

In the present study, the visual component corresponded to movement of the lip. The lateral occipitotemporal gyrus has been shown to be implicated in lip-reading in imaging studies (Bernstein et al., 2002; Calvert et al., 1997; Campbell et al., 2001), human lesion study (Campbell, Zihl, Massaro, Munhall, & Cohen, 1997), and human intracranial recordings (Besle et al., 2008). Sekiyama et al. have shown that the lateral occipitotemporal gyrus was one of the major areas related to lip-reading in native Japanese speakers (Sekiyama et al., 2003). These studies suggest that the lateral occipitotemporal gyrus is more or less related to lip-reading in both native English speakers and native Japanese speakers. In the present study, the visual influence on audiovisual speech perception was stronger for Japanese participants than for English speakers in the left lateral occipitotemporal gyrus, suggesting that the left lateral occipitotemporal gyrus is more associated with lip-reading in native Japanese speakers than in native English speakers. In the latter group, the left lateral occipitotemporal gyrus may not be necessarily involved in an easy task as in the present study (identification of /ba/ and /ga/).

The visual information about mouth movements occurs earlier than auditory information in natural speech production. Behavioral studies have shown that Japanese language speakers took more time to process the visual information than English language speakers (Sekiyama & Burnham, 2008). Sekiyama and Burnham have raised the possibility that the auditory and visual information would be accessible at a similar point in time for native Japanese speakers, while the visual information would be accessible at an earlier point than auditory information for native English speakers (Sekiyama & Burnham, 2008). In the present study, visual information enhanced the neural activity in the left lateral occipitotemporal gyrus in the native Japanese speakers, but not in the native English speakers, suggesting that the left lateral occipitotemporal gyrus might play more important role in native Japanese compared with English speakers for the deliberate neural processing of the visual component in audiovisual speech perception.

5. Conclusions

In audiovisual speech perception, activation in the primary auditory cortex is enhanced by additional visual information regardless of language background, while the neural processing in the lateral occipitotemporal gyrus for visual information depended on language. Language environment may change the neural process and/or neural pathway related to audiovisual speech perception. Future studies using magnetoencephalography (MEG) or effective connectivity analysis are needed to clarify the neural pathway and time course of audiovisual speech perception.

6. Acknowledgements

This research was supported by a contract with the National Institute of Information and Communications Technology entitled, ‘Multimodal integration for brain imaging measurements’ and a Grand-in-Aid for Scientific Research (21243040) from Ministry of Education, Culture, Sports, Science, and Technology, Japan.

7. References

[1] Beauchamp, M. S. (2005). See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex. *Curr Opin Neurobiol, 15*(2), 145-153.

[2] Beauchamp, M. S., Nath, A. R., & Pasalar, S. (2010). fMRI-Guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. *J Neurosci, 30*(7), 2414-2417.

[3] Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature, 403*(6767), 309-312.

[4] Bernstein, L. E., Auer, E. T., Jr., Moore, J. K., Ponton, C. W., Don, M., & Singh, M. (2002). Visual speech perception without primary auditory cortex activation. *Neuroreport, 13*(3), 311-315.

[5] Besle, J., Fischer, C., Bidet-Caulet, A., Lecaignard, F., Bertrand, O., & Giard, M. H. (2008). Visual activation and audiovisual interactions in the auditory cortex during speech perception: intracranial recordings in humans. *J Neurosci, 28*(52), 14301-14310.

[6] Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S., Springer, J. A., Kaufman, J. N., et al. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cereb Cortex, 10*(5), 512-528.

[7] Callan, D. E., Jones, J. A., Munhall, K., Kroos, C., Callan, A. M., & Vatikiotis-Bateson, E. (2004). Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. *J Cogn Neurosci, 16*(5), 805-816.

[8] Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., et al. (1997). Activation of auditory cortex during silent lipreading. *Science, 276*(5312), 593-596.

[9] Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Curr Biol, 10*(11), 649-657.

[10] Campbell, R., MacSweeney, M., Surguladze, S., Calvert, G., McGuire, P., Suckling, J., et al. (2001). Cortical substrates for the perception of face actions: an fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). *Brain Res Cogn Brain Res, 12*(2), 233-243.

[11] Campbell, R., Zihl, J., Massaro, D., Munhall, K., & Cohen, M. M. (1997). Speechreading in the akinetopsic patient, L.M. *Brain, 120 (Pt 10)*, 1793-1803.

[12] Evans, A. C., Collins, D. L., Mills, S. R., Brown, E. D., Kelly, R. L., & Peters, T. M. (1993, October). *3D statistical neuroanatomical models from 305 MRI volumes.* Paper presented at the IEEE-Nuclear Science Symposium and Medical Imaging Conference, San Francisco, CA, USA.

[13] Massaro, D. W., Tsuzaki, M., Cohen, M. M., Gesi, A., & Heredia, R. (1993). Bimodal speech perception: an examination across languages. *Journal of Phonetics, 21*, 445-478.

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

[15] Sekiyama, K., & Burnham, D. (2008). Impact of language on development of auditory-visual speech perception. *Dev Sci, 11*(2), 306-320.

[16] Sekiyama, K., Kanno, I., Miura, S., & Sugita, Y. (2003). Auditory-visual speech perception examined by fMRI and PET. *Neurosci Res*, *47*(3), 277-287.

[17] Sekiyama, K., & Tohkura, Y. (1991). McGurk effect in non-English listeners: few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *J Acoust Soc Am*, *90*(4 Pt 1), 1797-1805.

[18] Talairach, J., & Tournoux, P. (1988). *Co-Planar Stereotaxic Atlas of the Human Brain*. New York: Thieme Medical Publishers.

[19] Werner, S., & Noppeney, U. (2010). Distinct functional contributions of primary sensory and association areas to audiovisual integration in object categorization. *J Neurosci*, *30*(7), 2662-2675.

[20] Zeki, S. (1993). *A Vision of the Brain*. Oxford: Blackwell Scientific Publications.