Available online at www.sciencedirect.com
ScienceDirect
Neuroscience Letters 434 (2008) 71–76
Neuroscience Letters
www.elsevier.com/locate/neulet

Cross-modal integration during vowel identification in audiovisual speech: A functional magnetic resonance imaging study

Mika Murase [a,b], Daisuke N. Saito [b,c], Takanori Kochiyama [d], Hiroki C. Tanabe [a,b], Satoshi Tanaka [b], Tokiko Harada [b], Yu Aramaki [b,c], Manabu Honda [b,e,f], Norihiro Sadato [a,b,c,g,*]

[a] *Department of Physiological Sciences, The Graduate University for Advanced Studies (Sokendai), Kanagawa, Japan*
[b] *National Institute for Physiological Sciences, Okazaki, Japan*
[c] *Japan Science and Technology Corporation (JST)/Research Institute of Science and Technology for Society (RISTEX), Kawaguchi, Japan*
[d] *Department of Engineering, Kagawa University, Takamatsu, Japan*
[e] *Department of Cortical Function Disorders, National Institute of Neuroscience, National Center of Neurology and Psychiatry, Kodaira, Japan*
[f] *JST/Solution Oriented Research for Science and Technology (SORST), Kawaguchi, Japan*
[g] *Department of Functional Neuroimaging, Faculty of Medical Sciences, University of Fukui, Fukui, Japan*

Received 9 October 2007; received in revised form 10 January 2008; accepted 15 January 2008

Abstract

To investigate the neural substrates of the perception of audiovisual speech, we conducted a functional magnetic resonance imaging study with 28 normal volunteers. We hypothesized that the constraint provided by visually-presented articulatory speech (mouth movements) would lessen the workload for speech identification if the two were concordant, but would increase the workload if the two were discordant. In auditory attention sessions, subjects were required to identify vowels based on auditory speech. Auditory vowel stimuli were presented with concordant or discordant visible articulation movements, unrelated lip movements, and without visual input. In visual attention sessions, subjects were required to identify vowels based on the visually-presented vowel articulation movements. The movements were presented with concordant or discordant uttered vowels and noise, and without sound. Irrespective of the attended modality, concordant conditions significantly shortened the reaction time, whereas discordant conditions lengthened the reaction time. Within the neural substrates that were commonly activated by auditory and visual tasks, the mid superior temporal sulcus showed greater activity for discordant stimuli than concordant stimuli. These findings suggest that the mid superior temporal sulcus plays an important role in the auditory–visual integration process underlying vowel identification.

Keywords: Cross-modal integration; Human voice; Perception; Lip reading; Spoken language; Superior temporal sulcus

Spoken language perception is enhanced by combining audible speech with corresponding visible articulation movements in what is known as audiovisual (AV) speech [16]. The neural substrates of AV speech have been investigated using functional neuroimaging [7]. The brain regions reported to be involved in AV integration are the superior temporal sulcus (STS), intraparietal sulcus (IPS), inferior frontal gyrus (IFG), primary auditory cortex, claustrum, and superior colliculus [2–6,10,11]. Among these, the STS is the core substrate for AV speech. However, the mechanism underlying this activity remains largely unknown. The purpose of the present study was to elucidate the functional role of the STS in cross-modal integration during AV speech. We conducted event-related functional magnetic resonance imaging (fMRI) with mixed concordant and discordant AV speech, in conjunction with a task in which vowels were identified based on either auditory or visual speech. We adopted the logic of Raij et al. [13], who suggested that brain areas participating in AV integration should show signs of convergence (that is, both auditory and visual stimuli should activate the same region) and of interaction (that is, the activation evoked by AV stimulation should differ depending on the workload required for AV integration). We predicted that the reaction time (RT) required to identify the vowel would be shortened if AV speech was concordant, and elongated if AV speech was discordant (concordance

* Corresponding author at: Division of Cerebral Integration, Department of Cerebral Research, National Institute for Physiological Sciences, Myodaiji, Okazaki, Aichi 444-8585, Japan. Tel.: +81 564 55 7841; fax: +81 564 55 7786.
E-mail address: sadato@nips.ac.jp (N. Sadato).

0304-3940/$ – see front matter © 2008 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.neulet.2008.01.044

effect). Correspondingly, the evoked neural activity representing the audio-visual integration process should be larger during the discordant condition than the concordant condition, because of an increase in signal uncertainty during the former condition [19].

In total, 28 (14 male and 14 female) native Japanese speakers participated in this experiment. Their ages ranged from 25 to 38 years (mean = 27.9 years; standard deviation = 3.3 years). Of these, 26 subjects were classed as right-handed and two male subjects were classed as left-handed according to the Edinburgh handedness inventory [9]. None of the subjects had a history of neurological or psychiatric illness. The protocol was approved by the Ethical Committee of the National Institute for Physiological Sciences, Japan, and all subjects gave their written informed consent for participation.

Echo-planar imaging (EPI) images were captured (repetition time [TR] = 2 s; echo time [TE] = 30 ms; flip angle [FA] = 75°; field of view [FOV] = 19.2 cm; 64 × 64 pixels; 32 slices of 3.5 mm thickness and 0.5 mm gap to cover the entire cerebral and cerebellar cortices) with an Allegra 3.0 Tesla MR imager (Siemens, Erlangen, Germany).

The auditory and visual stimuli were constructed by editing a digitally recorded female voice and face pronouncing five syllables (/a/, /i/, /u/, /e/, and /o/) using a video-recorder (Sony, Tokyo, Japan). Natural speech was recorded with a matrix size of 640 × 480 pixels, a digitization rate of 30.0 frames/s (1 frame = 33.3 ms), and stereo soundtracks at an 11.025-kHz sampling rate with 16-bit resolution. Video clips (24 frames; 33 ms/frame) containing a syllable were edited using Adobe Premiere software (Adobe, San Jose CA, USA) so that the 12th frame contained the onset of the speech; this format was chosen because the average duration of the speech was 12 frames, and the onset of lip movement preceded that of the sound by seven frames. Five auditory noise streams were generated by randomly shuffling the frames of the auditory streams of five syllables (12 frames per syllable; a total of 60 frames). The onset of the noise was adjusted to correspond with the 12th frame. Additionally, closed lip movements were recorded and the video clips were adjusted so that the onset of the lip movement occurred in the fourth frame. Each video clip was composed of a soundtrack and a movie stream, which were stored in separate files. Hence, the video clips for each condition were created by dubbing the soundtracks onto the movie streams of different clips.

The auditory vowel identification task consisted of vowel identification based on speech with or without concomitant lip movements. The maximum intensity (90 dB at the ear), frequency range, and duration of each stimulus were adjusted, and the stimuli were presented via headphones using Presentation software (Neurobehavioral Systems, Albany, CA, USA). The visual stimuli were presented at a visual angle of 7.6 × 9.6°. We used an event-related design in order to minimize habituation and learning effects. Each subject placed their right hand over a box with five response buttons. Throughout the session, the subjects were asked to fixate a small cross-hair at the center of the screen. We explicitly instructed the subjects not to close their eyes during the session, except when blinking.

The design consisted of five types of event condition. The first event condition was auditory vowel identification without visual stimuli (A), in which a single voice pronouncing a vowel (for example, /a/) was presented. Subjects were instructed to press the button corresponding to the vowel as follows: /a/ = thumb; /i/ = index finger; /u/ = middle finger; /e/ = ring finger; and /o/ = little finger. All subjects were required to respond as quickly as possible and within 1600 ms (Fig. 1, supplementary Table 1). The second event condition was auditory vowel identification with concordant lip movement (AVcon), in which a single voice pronouncing a vowel was presented, accompanied by the lip movement for the same vowel. The timing of the speech was naturally synchronized with the lip movement. The subjects were instructed to press the button corresponding to the spoken

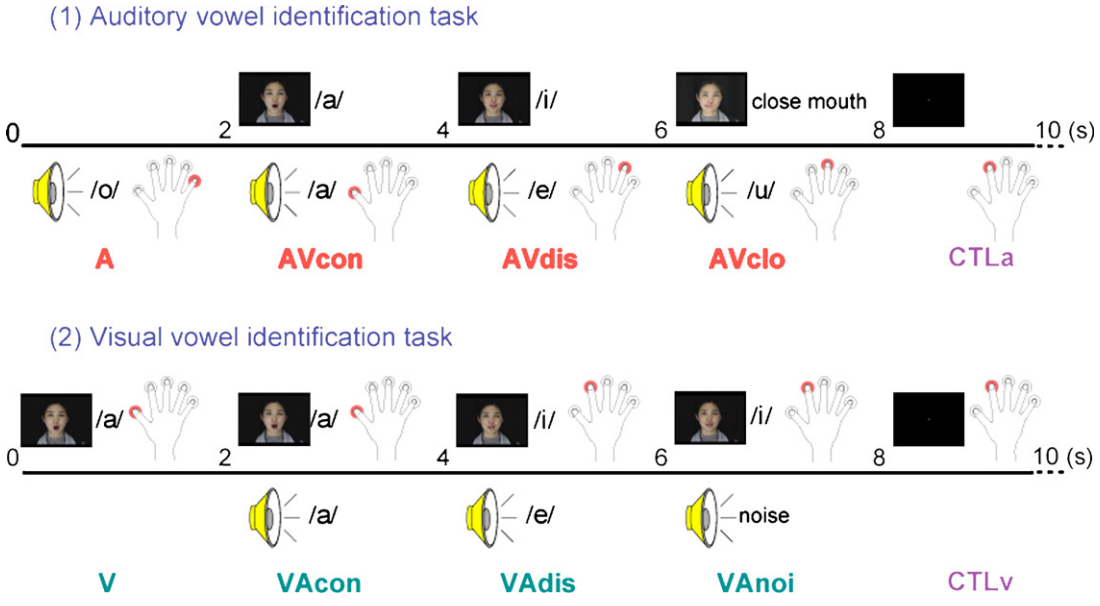

Fig. 1. Task design. The rapid event-related auditory (Top) and visual (Bottom) vowel identification tasks were conducted as separate sessions. The ITI was 2 s.

vowel as quickly as possible. The third event condition was auditory vowel identification with discordant lip movement (AVdis). The procedure was similar to that described for AVcon, with the exception that the lip movement differed from the spoken vowel. The fourth event condition was auditory vowel identification with closed lip movement (AVclo). The procedure was similar to that described for AVcon, with the exception that the lips twitched but remained closed, thereby resulting in no visual phonetics. The fifth event condition was the motor control condition (CTLa), which included no visual or auditory stimuli except for the cross-hair fixation point. The subjects were instructed to press a button when the color of the cross-hair changed. Nine of the subjects were instructed to push the button with their thumb, 11 with their index finger, and the remaining eight with their middle finger.

The visual vowel identification task was performed with an identical setup to that described for the auditory vowel identification session. The design consisted of five types of event condition (Fig. 1). The first event condition was visual vowel identification without auditory stimuli (V), in which a face pronouncing a vowel was presented without speech. As in the auditory vowel identification task the subjects were instructed to press the button. The second event condition was visual vowel identification with concordant speech (VAcon), in which a face pronouncing a vowel was presented along with the sound of the same vowel. The subjects were instructed to press the button corresponding to the lip movement as quickly as possible. The third event condition was visual vowel identification with discordant speech (VAdis). The procedure was similar to that described for VAcon, with the exception that the spoken vowel differed from the lip movement. The fourth event condition was visual vowel identification with noise (VAnoi). The procedure was similar to that described for VAcon, with the exception that noise was presented instead of speech. The fifth event condition was the motor control condition (CTLv), in which there was no visual or auditory stimulus except for the cross-hair. The subjects were instructed to press the thumb button when the color of the cross-hair changed. The inter-trial interval (ITI) was fixed at 2 s. Each condition was repeated 40 times, so the total number of events was 200 per session. The experimental protocol used was a rapid event-related design, which maximized the efficiency for the contrasts of interest [14]. Each session was repeated twice.

The first three volumes from each fMRI session were discarded to allow for the stabilization of the magnetization. In total, 800 volumes per subject were included in the analysis. Imaging data were analyzed using statistical parametric mapping (SPM99, Wellcome Department of Cognitive Neurology, London, UK) implemented in Matlab (Mathworks, Sherborn, MA, USA). The EPI images were realigned, spatially normalized into stereotaxic space, and smoothed with an isotropic Gaussian kernel of 8 mm full width at half maximum.

In the individual analyses, the signal time course for each participant was modeled using a delta function convolved with a hemodynamic response function, session effect, trial effect, and high-pass filtering (60 s). The explanatory variables were centered to zero. To test hypotheses about regionally-specific trial effects, the estimates for each model parameter were compared with the linear contrasts. First, we delineated the areas activated in the A, AVcon, AVdis, and AVclo conditions, and the V, VAcon, VAdis, and VAnoi conditions, as compared with those activated during the CTL periods of the equivalent sessions. The resulting set of voxel values constituted a statistical parametric map of the *t*-statistic, SPM{*t*}. Second, multimodal areas were depicted through the intersection of A and V (masking procedure) with a statistical threshold of $Z > 3.09$ and a cluster size > 10 voxels (80 mm^3). Within these multimodal areas, the cross-modal areas were defined as those that were more prominently activated during the discordant condition than the concordant condition. The statistical threshold for the discordance effects within the multimodal areas were set at a false discovery rate (FDR) corrected *P*-value of <0.05.

In the group analysis using the random-effect model, the weighted sum of the parameter estimates in the individual analysis constituted the "contrast" images. The contrast images obtained via the individual analysis represented the normalized task-related increment of the MR signal of each subject. For each contrast, a one-tailed one-sample *t*-test was performed for every voxel to obtain population inferences. The discordant effects were depicted with the same procedures and statistical threshold as in individual analysis.

During task performance, the percentage of correct responses was higher for the auditory vowel identification sessions than for the visual sessions (modality effect; $F\ [1,27] = 13.4$; $P = 0.001$). The condition effects ($F\ [2.1, 57.4] = 13.1$; $P < 0.001$; repeated-measures analysis of variance [ANOVA] with Greenhouse-Geisser correction) and the modality × condition interaction ($F\ [3,81] = 6.9$; $P = 0.001$) were also significant. Specifically, performance on A was better than that on V ($P < 0.05$; paired *t*-test), performance on AVdis was better than that on VAdis ($P < 0.05$; paired *t*-test), and performance on AVclo was better than that on VAnoi ($P < 0.05$; paired *t*-test). AVcon and VAcon showed similar performance levels ($P = 0.47$; paired *t*-test; supplementary Table 1).

The median RT for correct responses was calculated for each subject (supplementary Table 1). The RT was longer for the auditory vowel identification sessions than for the visual

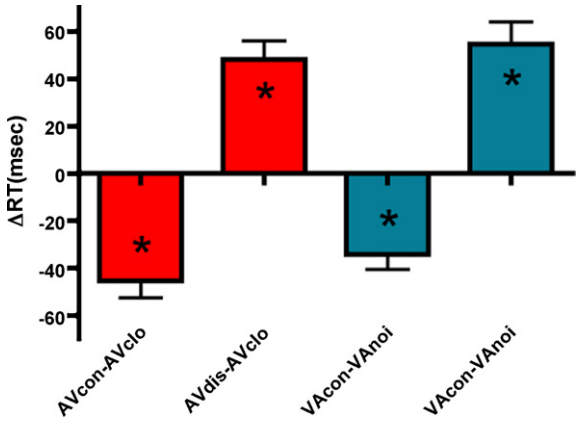

Fig. 2. Task performance. The RTs in the concordant and discordant conditions compared with those in the AV control conditions (AVclo and VAnoi). The error bar indicates the S.E.M. *$P < 0.001$ (one sample *t*-test).

sessions (modality effect; $F[1,27] = 46.4$; $P < 0.001$). The condition effects ($F[2.3, 62.9] = 106.4$; $P < 0.001$; repeated-measures ANOVA with a Greenhouse-Geisser correction) and the modality × condition interaction ($F[2.2, 59.2] = 7.0$; $P < 0.05$; repeated-measures ANOVA with a Greenhouse-Geisser correction) were also significant. Specifically, the RT of the A condition was 136 ms longer than that of V ($P < 0.001$; paired t-test). As the visual stimuli were presented 231 ms earlier than the auditory

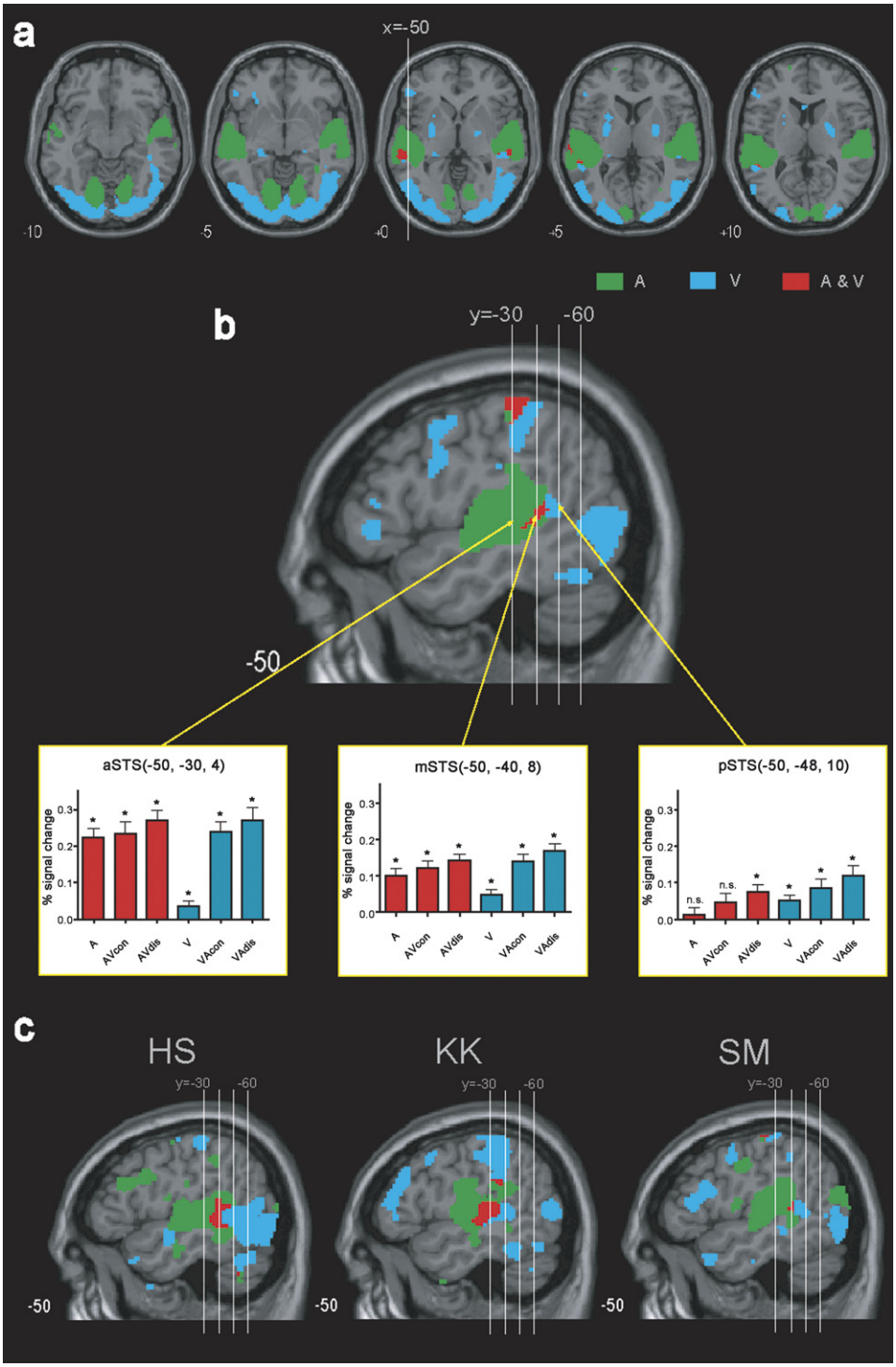

Fig. 3. Unisensory and polysensory activation revealed by group analysis with random-effect model (a and b) and individual analysis (c). The areas showing significant activation by A (green) and V (blue), and their overlap (red), are superimposed on a T1-weighted high-resolution MRI of (a) the transaxial slices and (b) the parasagittal section at $x = -50$ mm. The four white lines indicate the plane of the Montreal Neurological Institute (MNI) coordinates of $y = -30$, -40, -50, and -60 mm. Note that the polysensory STS (red) is located more anteriorly (around $y = -40$ mm), whereas the posterior STS located around $y = -40$ to -50 mm is mainly activated by V. (Bottom) The percentage signal change in the anterior (left), mid (middle), and posterior STS (right). *$P < 0.001$. No significant activation was found in the posterior STS (-50, -48, 10) during A ($P = 0.50$, one-sample t-test) or AVcon ($P = 0.07$, one-sample t-test). (c) Three representative activation maps by individual analysis (subjects HS, KK, and SM) are shown with the same format as group analysis.

stimuli, these findings might have been due to a cue effect. To rule out this confounding factor, comparisons with the AV control conditions (AVclo and VAnoi) were also performed (Fig. 2). Concordant conditions led to a significant shortening of the RT, whereas the values were higher during discordant conditions. Repeated-measures ANOVA revealed a significant concordance effect (F [1,27] = 208.1; $P < 0.001$), but neither a significant modality effect (AV versus VA; F [1,27] = 0.9; $P = 0.34$) nor a significant interaction (F [1,27] = 0.1; $P = 0.77$) were observed.

In the fMRI experiment, a discordance effect (Dis – Con) during the auditory and visual vowel identification session was observed in the polysensory areas (A & V; Fig. 3a and b) in the bilateral STS (Fig. 4). This activation was observed irrespective of the attended modality. A two-way repeated-measures ANOVA on the right STS (x, y, and z coordinates: 52, −34, 4) data from the AVcon, AVdis, VAcon, and VAdis conditions revealed that the main condition effect (concordant versus discordant) was significant (F[1,27] = 15.5; $P < 0.001$) whereas the main modality effect (AV versus VA; F[1,27] = 0.212; $P = 0.649$) and their interaction (F[1,27] = 0.023; $P = 0.882$) were not significant. In the left STS, the main condition effect (concordant versus discordant) was significant (F[1,27] = 14.2; $P < 0.001$) whereas the main modality effect (AV versus VA; F[1,27] = 0.1; $P = 0.73$) and their interaction (F[1,27] = 0.1; $P = 0.80$) were not.

The percentage of correct responses was higher during auditory vowel identification than visual vowel identification. Performance was equal only when concordant auditory stimuli were presented. This reflects the fact that lip movement is a less reliable stimulus than speech for vowel identification. The average RT for auditory vowel identification was 109 ms longer than that for visual vowel identification. As movement of the facial articulators typically precedes the onset of the acoustic signal by up to a few hundred milliseconds [19], visual speech might provide a direct clue for the on-line prediction of auditory

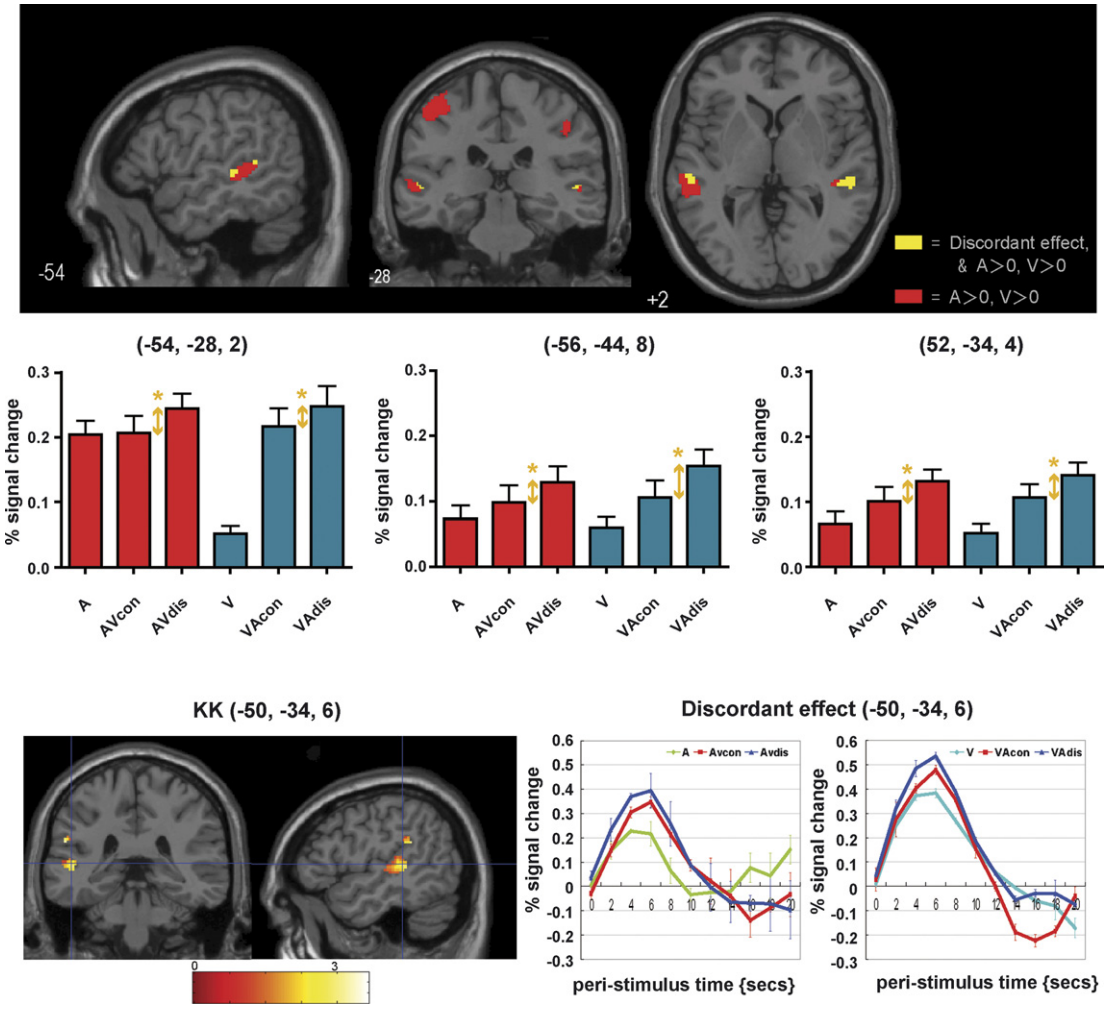

Fig. 4. Cross-modal interaction within the multisensory areas. (Top) The areas showing significant activation by both A and V (red; $P < 0.001$ uncorrected) are superimposed on the T1-weighted high-resolution MRI of the sagittal, coronal, and transaxial sections that cross at (−54, −28, 2). Within the multimodal areas, the regions showing a discordance effect (yellow; $P < 0.05$ FDR corrected) – that is, more prominent activation during the discordant condition (AVdis or VAdis) than the concordant condition (AVcon or VAcon) – are also superimposed. (Middle) The percentage BOLD signal changes in the STS at (−54, −28, 2) (left), (−56, −44, 8) (middle), and (52, −34, 4) (right). *$P < 0.05$ (FDR corrected). (Bottom) Individual analysis of the discordant effect in the STS (left, $P < 0.05$ uncorrected). The percentage signal changes during auditory attending sessions (Middle) and visually attending sessions (Right) were plotted against peri-stimulus time.

signals [19]. However, we observed a small but significant AV interaction that was not specific to a particular vowel. By discounting these nonspecific effects, the concordance/discordance effect was found not to be specific to the attended modality. The RT needed to identify the vowel was shorter when AV speech was concordant and longer when it was discordant (concordance effect), irrespective of the attended modality. These results also suggest that the time period required for vowel identification and cross-modal interaction was longer than the differences in the onsets for each modality.

Within the polysensory area characterized by multisensory convergence, the mid STS showed more prominent activation during the discordant condition than the concordant condition (Fig. 4). No region within the polysensory areas showed more prominent activation during the concordant condition than during the discordant condition. This may be caused by the vowel identification task that is potentially influenced by the long-term learning processes. A recent study showed that the STS was involved in audio-visual cross-modal associative learning [17]. Experience plays a critical role in forming the AV associations that underlie AV speech perception [15]. Learned cross-modal association may provide constraint on the internal construction of multisensory perceptual representations [19]. Therefore in discordant condition, more workload is needed to accomplish vowel identification against the constraint. The long-term learning process could be compared to the sharpening of neuronal tuning [8,13,19]. Thus, the stronger signal amplitudes for the discordant condition in the STS in the present study might reflect suboptimal tuning in the local network.

Visual speech primarily activated the right posterior STS region (40–55 mm posterior from the anterior commissure [AC]). Puce et al. [12] reported that nonlinguistic mouth movement elicited activation in the right posterior STS (50, −49, 3). Based on these findings, Wright et al. [20] concluded that the posterior STS area is a candidate visual speech processing area. It is important to inspect the unimodal responses in candidate integration regions [1,18,20]. Therefore, we plotted the responses in the unimodal areas and the area of overlap of these regions (the STS; Fig. 3b). The plot indicates an anterior–posterior gradient of the AV representation in the STS, in which auditory information is represented in more anterior regions, visual information is represented in more posterior regions, and AV polysensory representation occurs in-between. Thus, the representation of the audiovisual vowel identification involves the coordinated activity along the superior temporal sulcus. The mid STS may represent the auditory–visual convergence and interaction, i.e., cross-modal integration process.

Acknowledgments

This study was supported by a Grant-in-Aid for Scientific Research (S#17100003) to N.S. from the Japan Society for the Promotion of Science.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.neulet.2008.01.044.

References

[1] M.S. Beauchamp, See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex, Curr. Opin. Neurobiol. 15 (2005) 145–153.
[2] D.E. Callan, J.A. Jones, K. Munhall, A.M. Callan, C. Kroos, E. Vatikiotis-Bateson, Neural processes underlying perceptual enhancement by visual speech gestures, Neuroreport 14 (2003) 2213–2218.
[3] D.E. Callan, J.A. Jones, K. Munhall, C. Kroos, A.M. Callan, E. Vatikiotis-Bateson, Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information, J. Cogn. Neurosci. 16 (2004) 805–816.
[4] G.A. Calvert, Crossmodal processing in the human brain: insights from functional neuroimaging studies, Cereb Cortex. 11 (2001) 1110–1123.
[5] G.A. Calvert, M.J. Brammer, E.T. Bullmore, R. Campbell, S.D. Iversen, A.S. David, Response amplification in sensory-specific cortices during crossmodal binding, Neuroreport 10 (1999) 2619–2623.
[6] G.A. Calvert, E.T. Bullmore, M.J. Brammer, R. Campbell, S.C. Williams, P.K. McGuire, P.W. Woodruff, S.D. Iversen, A.S. David, Activation of auditory cortex during silent lipreading, Science 276 (1997) 593–596.
[7] G.A. Calvert, R. Campbell, M.J. Brammer, Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex, Curr. Biol. 10 (2000) 649–657.
[8] A. Hurlbert, Visual perception: learning to see through noise, Curr. Biol. 10 (2000) R231–R233.
[9] R.C. Oldfield, The assessment and analysis of handedness: the Edinburgh inventory, Neuropsychologia 9 (1971) 97–113.
[10] I.R. Olson, J.C. Gatenby, J.C. Gore, A comparison of bound and unbound audio-visual information processing in the human cerebral cortex, Brain Res. Cogn. Brain Res. 14 (2002) 129–138.
[11] J. Pekkola, V. Ojanen, T. Autti, I.P. Jaaskelainen, R. Mottonen, A. Tarkiainen, M. Sams, Primary auditory cortex activation by visual speech: an fMRI study at 3 T, Neuroreport 16 (2005) 125–128.
[12] A. Puce, T. Allison, S. Bentin, J.C. Gore, G. McCarthy, Temporal cortex activation in humans viewing eye and mouth movements, J. Neurosci. 18 (1998) 2188–2199.
[13] T. Raij, K. Uutela, R. Hari, Audiovisual integration of letters in the human brain, Neuron 28 (2000) 617–625.
[14] D.N. Saito, K. Yoshimura, T. Kochiyama, T. Okada, M. Honda, N. Sadato, Cross-modal binding and activated attentional networks during audio-visual speech integration: a functional MRI study, Cereb Cortex 15 (2005) 1750–1760.
[15] E.A. Schorr, N.A. Fox, V. van Wassenhove, E.I. Knudsen, Auditory-visual fusion in speech perception in children with cochlear implants, Proc. Natl. Acad. Sci. U.S.A. 102 (2005) 18748–18750.
[16] W.H. Sumby, I. Pollack, Visual contribution to speech intelligibility in noise, J. Acoust. Soc. America 26 (1954) 212–215.
[17] H.C. Tanabe, M. Honda, N. Sadato, Functionally segregated neural substrates for arbitrary audiovisual paired-association learning, J. Neurosci. 25 (2005) 6409–6418.
[18] N.M. van Atteveldt, E. Formisano, L. Blomert, R. Goebel, The effect of temporal asynchrony on the multisensory integration of letters and speech sounds, Cereb Cortex 17 (2007) 962–974.
[19] V. van Wassenhove, K.W. Grant, D. Poeppel, Visual speech speeds up the neural processing of auditory speech, Proc. Natl. Acad. Sci. U.S.A. 102 (2005) 1181–1186.
[20] T.M. Wright, K.A. Pelphrey, T. Allison, M.J. McKeown, G. McCarthy, Polysensory interactions along lateral temporal regions evoked by audiovisual speech, Cereb Cortex 13 (2003) 1034–1043.