Available online at www.sciencedirect.com
SCIENCE DIRECT®
Neuroscience Research 47 (2003) 277–287
Neuroscience Research
www.elsevier.com/locate/neures

Auditory-visual speech perception examined by fMRI and PET

Kaoru Sekiyama [a,b,*], Iwao Kanno [c], Shuichi Miura [c], Yoichi Sugita [a]

[a] *Neuroscience Research Institute, National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan*
[b] *Division of Cognitive Psychology, Future University-Hakodate, Hakodate 041-8655, Japan*
[c] *Department of Radiology and Nuclear Medicine, Akita Research Institute of Brain and Blood Vessels, Akita, Japan*

Received 17 March 2003; accepted 27 June 2003

Abstract

Cross-modal binding in auditory-visual speech perception was investigated by using the McGurk effect, a phenomenon in which hearing is altered by incongruent visual mouth movements. We used functional magnetic resonance imaging (fMRI) and positron emission tomography (PET). In each experiment, the subjects were asked to identify spoken syllables ('ba', 'da', 'ga') presented auditorily, visually, or audiovisually (incongruent stimuli). For the auditory component of the stimuli, there were two conditions of intelligibility (High versus Low) as determined by the signal-to-noise (SN) ratio. The control task was visual talker identification of still faces. In the Low intelligibility condition in which the auditory component of the speech was harder to hear, the visual influence was much stronger. Brain imaging data showed bilateral activations specific to the unimodal auditory stimuli (in the temporal cortex) and visual stimuli (in the MT/V5). For the bimodal audiovisual stimuli, activation in the left temporal cortex extended more posteriorly toward the visual-specific area in the Low intelligibility condition. The direct comparison between the Low and High audiovisual conditions showed increased activations in the posterior part of the left superior temporal sulcus (STS), indicating its relationship with the stronger visual influence. It was discussed that this region is likely to be involved in cross-modal binding of auditory-visual speech.

Keywords: Cross-modal binding; Auditory-visual integration; Speech perception; The McGurk effect; fMRI; PET; Superior temporal sulcus

1. Introduction

The visual cues from a speaker's mouth movements play an important role in speech perception. They facilitate speech perception when auditory speech is degraded (e.g. Sumby and Pollack, 1954; Rosen et al., 1981). Furthermore, the visual cues alter what the perceiver hears when incongruent visual and auditory cues are presented, as demonstrated in the McGurk effect (McGurk and MacDonald, 1976).

Recent psychophysical research has shown that in various perceptual domains, the brain tends to bind cross-modal inputs not only when they are congruent, but also when they are incongruent or in an ambiguous relationship. For example, sound localization is displaced by incongruent visual source information (Bertelson et al., 2000), visual motion perception is altered by an additional sound (Sekuler et al., 1997), and visual frequency judgment is distorted by sounds (Shams et al., 2000). The McGurk effect, speech perception altered by discrepant mouth movements (McGurk and MacDonald, 1976), can be also seen as an example of humans' ubiquitous propensity to bind cross-modal inputs.

In this study, we used the McGurk effect to investigate the cross-modal processing in auditory-visual speech perception. The McGurk effect demonstrates an influence of discrepant visual input on auditory speech perception (McGurk and MacDonald, 1976; MacDonald and McGurk, 1978). When incongruent auditory and visual inputs are presented in synchrony, the perceiver often reports hearing a syllable distinct from the auditory one (e.g. audio /pa/ + video /na/ results in the perception of "ta"). In this case, the percept is an integrated product of information from the two sensory modalities. This illusion depends much on the complementary nature of the two modalities (Binnie et al., 1974). That is, visual speech is advantageous to conveying the information about the place of articulation (e.g. at the lips or inside of the mouth) while auditory speech is robust for conveying the rest of the information (the manner of articulation and voicing).

* Corresponding author. Tel.: +81-138-34-6327; fax: +81-138-34-6301.
E-mail address: sekiyama@fun.ac.jp (K. Sekiyama).

0168-0102/$ – see front matter
doi:10.1016/S0168-0102(03)00214-1

According to a magnetoencephalographic (MEG) study by Sams et al. (1991), it seems that the McGurk effect is related to the 'supratemporal' cortex. In each trial, their subjects were presented either auditory-visual congruent "pa" or auditory /pa/ combined with discrepant visual [ka]. Of these congruent or discrepant (McGurk) stimuli, one was presented more frequently (84% of stimuli) and the other infrequently (16%). They found a 'mismatch response' to the infrequent stimulus around the superior temporal and/or inferior frontal cortices. Unfortunately, the implication of the result is not straightforward due to the mismatch paradigm. The fact that they recorded only from the left hemisphere gives limited information. We used functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) to investigate brain activation in the McGurk effect.

Our goal was to compare the brain activation for two audiovisual conditions in which the intelligibility of the auditory component of stimuli differs. Although the tendency to integrate incompatible auditory and visual speech has been well documented in English speaking countries (for review, Summerfield, 1992; Campbell et al., 1998; Massaro, 1998), some Asian peoples such as the Japanese are less subject to the McGurk effect (Sekiyama and Tohkura, 1993; Kuhl et al., 1994). Whereas native speakers of English show a strong McGurk effect for highly intelligible auditory speech, the Japanese do not use visual cues as much as do the native speakers of English unless auditory speech has some ambiguity due to added noise or foreign accent (Sekiyama and Tohkura, 1991; Kuhl et al., 1994; Sekiyama, 1994). We used this tendency of the Japanese to unravel the integration process in speech perception. The same audiovisual stimuli were presented in two different levels of auditory intelligibility. Whereas the substantial McGurk effect was anticipated in the Low auditory intelligibility condition, only a limited McGurk effect was in the High auditory intelligibility condition. The brain activation was compared between the two conditions. We also tried to examine the relationship between unimodal activation for auditory-alone or visual-alone stimuli and bimodal activation for audiovisual stimuli.

We used fMRI in Experiment 1, and PET in Experiment 2. Since we intended to compare bimodal activation between the High and Low auditory intelligibility conditions, the PET experiment was to obtain a larger difference in auditory intelligibility between the two conditions. As is well known, PET scans can be conducted with much less scanner noise than fMRI scans. Therefore, a noiseless audiovisual stimulation can be realized only in the PET High intelligibility condition. It was anticipated that when the brain activation data in the Low intelligibility audiovisual condition is contrasted to those in the High intelligibility condition, the difference would be larger in the PET experiment than in the fMRI experiment.

The results of the two experiments were basically in good agreement although fMRI and PET target two different measures, that is, regional cerebral blood flow (rCBF) in PET and blood oxygen level dependent (BOLD, which is a by-product of cerebral blood flow) in fMRI.

2. Materials and methods

2.1. Subjects

The fMRI data were from eight native speakers of Japanese. They were healthy right-handed volunteers (aged 22–46 years; seven male and one female) with normal hearing, and normal or corrected to normal vision. The PET data were collected from another comparable group of 10 Japanese subjects (aged 20–46 years; all male). All subjects gave informed written consent for participation.

2.2. Stimuli and tasks

Stimuli were created from "ba", "da", and "ga" uttered by three female talkers (Fig. 1A). The utterances were videotaped (with the talker's full face), digitized, and edited on a computer (Sony PCV-S720) for audio-only (A), video-only (V), and audiovisual (AV) stimuli. Video digitizing was done at 29.97 frames/s in 640 × 480 dots, and audio digitizing was at 32 kHz in 16 bit. Each stimulus was created as a 2 s movie of a monosyllabic utterance. The duration of acoustic speech signals in each movie was approximately 330 ms in average. The movie files were edited with frame unit accuracy (33.3 ms), but the sound portion was additionally edited with 1 ms accuracy. Each natural utterance was first cut such that the onset of its acoustic energy was at some point in the 19th frame of the movie. Then the position of the acoustic

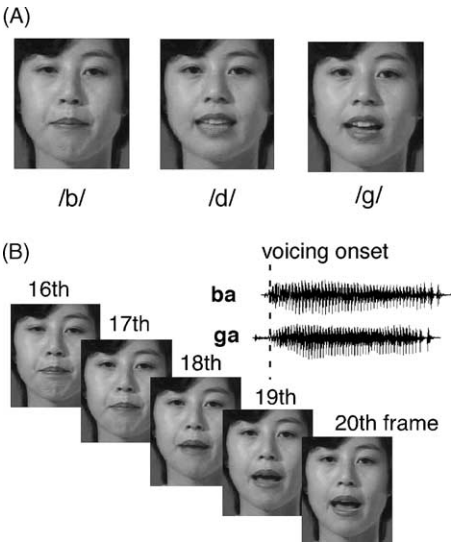

Fig. 1. Stimulus samples. (A) A video frame of consonant constriction for /ba/, /da/, and /ga/, shown for one of the three talkers. These video frames were one or two frames earlier than the acoustic onset of voicing. (B) A part of video sequence for /ba/. The acoustic onset of voicing was at 600 ms in all sound files. In off-line editing, the voicing onset was synchronized with the beginning of the 19th video frame.

signals was adjusted so that the onset of voicing was at the beginning of the 19th frame (i.e. 600 ms from the onset of the movie file).

The A stimuli (/ba/, /da/, and /ga/ of the three talkers) were used to present only the auditory component of speech, but were combined with the talker's still face with mouth neutrally closed (thus no linguistic information in the face). The V stimuli ([ba], [da], and [ga] of the three talkers) consisted of only the video component of speech, providing projections of a silent talking face. The AV stimuli were within-talker combinations of the synchronized auditory and visual speech. All AV stimuli were the McGurk-type stimuli consisting of discrepant auditory and visual syllables (audio /ba/ was combined with video [da] or [ga], and audio /da/ and /ga/ were combined with video [ba]). To make the dubbing, the onset of the voicing was used as the synchronization reference (Fig. 1B).

For the A, V, and AV stimuli, the subject's task was syllable identification. The subjects were instructed to watch and listen to the talkers speaking and were asked to report what they perceived by choosing a syllable from three alternatives ('ba', 'da', and 'ga'). Although the audiovisual stimuli included so-called "combination presentations" (auditory /da/ or /ga/ combined with visual [ba] which often produce "combination responses", for example, "bda" or "bga"), we did not allow such responses based on our previous results that Japanese perceivers rarely reported combination responses in an open-choice task (Sekiyama and Tohkura, 1991). For these stimuli, "ba" responses were anticipated for visually influenced responses. The subjects were not informed of the incompatibility of the AV stimuli at all. They were to signal each of the three alternatives by using fingers. There was also a control (C) condition for which the static neutral faces (visual component of the A stimuli) were presented silently for a visual talker identification task. The subjects were asked to signal which of the three talkers they saw by using fingers. We assumed that the C condition yields processing of talker information only, whereas the A, V, and AV conditions involve linguistic processing as well as visual talker processing. Note that the C condition was equalized with the A, V, and AV conditions in the processes of response selection and response execution.

2.3. fMRI experiment (Experiment 1)

2.3.1. Procedure

The stimuli were presented from the computer. The visual stimuli were projected onto a rear screen that the subject viewed through a mirror attached to the head coil. The auditory stimuli were presented on a loud speaker (Bose 121V) located outside of a shielded room in which the subject was being tested. The output of the loud speaker was conveyed through a pipe (10 cm in diameter, 2 m long) from the outside loud speaker to the subject (in the middle of both ankles). We used the loud speaker rather than a headset because the sound quality was better with the loud speaker than with the air-tube headset provided in the magnet. The sound intensity was adjusted by using an audio amplifier (Onkyo A-924(N)) placed between the computer and the loud speaker.

The subject's head was scanned while a sequence (A-V-C-AV) of the stimuli was presented in two levels of sound intensity. The intensity levels of speech sounds (altered by using the audio amplifier) were approximately 112 dB sound pressure level (SPL) for the High intelligibility runs and 102 dB SPL for the Low intelligibility runs. The timing of the speech was set such that the speech was always presented during the MRI scan noise (about 105 dB SPL). Thus, the signal-to-noise ratios were +7 dB and −3 dB. Although the sound intensities were rather high compared with usual speech perception experiments, they were not uncomfortable to the subjects whose auditory system soon became adapted to the MRI scan noise. Also note that the MRI scanner we used was relatively quiet compared with MRI scanners of higher magnetic fields.

2.3.2. Data acquisition

Echo planar MRI data were obtained with a 1-T Siemens Magnetom system with a standard circular-polarized (CP) head coil. T2*-weighted functional images were acquired for 10 axial noncontiguous 6 mm thick slices with 3 mm interslice gap. The following parameter settings were used: Repetition time (TR) was 3.95 s, echo time (TE) 66 ms, field of view (FOV) 200 mm, and spatial resolution 3.13 mm × 3.13 mm (matrix = 64 × 64). These axial slices, covering approximately two thirds of the cortex in height, included most of the temporal, occipital, and parietal cortices. The stimuli were presented in a blocked design by alternating four stimulus conditions in an A-V-C-AV pattern. During each condition epoch, eight stimuli were presented with a stimulus onset asynchrony (SOA) of 3.95 s. The functional runs consisted of 102 volumes (6 dummy + 96). Thus, with the 3.95 s TR, one functional run took about 7 min. Each subject participated in two runs for each of the High and Low intelligibility conditions. The order of the two conditions was counterbalanced among the subjects.

2.4. PET experiment (Experiment 2)

2.4.1. Procedure

The stimulus set and the task were identical to those in the fMRI experiment. The subjects were given the AV, A, V, and C blocks. The stimuli were presented from a computer (Sony PCG-Z505). The visual component of the stimuli was presented on a CRT display (Sony Multiscan 17GS), which was attached to a display arm suspended from the ceiling. The viewing distance was approximately 50 cm. The auditory component was presented through earphones (via an audio mixer, Pioneer DJM-300) due to the restriction for the head stabilization. In each block, 21 stimuli were presented in random order with a fixed SOA of 3.1 s (for 65 s stimulation).

For each of the AV and A conditions, there were two conditions depending on the SN ratio of the auditory

component of the stimuli (High and Low). Thus, there were six blocks in total (AV, nAV, A, nA, V, and C, with the prefix 'n' indicating 'noise-dominant' due to a lower SN ratio). The intensity level of speech sounds was approximately 75 dB SPL. To vary the SN ratio, we added white noise where intensity was either 60 or 80 dB. Thus, the SN ratios were +15 dB (in the AV and A conditions) and −5 dB (in the nAV and nA conditions). The SN ratio of +15 dB was regarded as 'noise-free' because previous performance data in our laboratory showed that the effect of the added noise on auditory-visual interaction is virtually null if the SN ratio is higher than +12 dB. This weaker level of noise was to mask unexpected variation of the room noise. The order of the six blocks was randomized for each subject.

2.4.2. Data acquisition

PET scans were performed using a 3 module-ring PET scanner (Shimadsu Headtome V), operated in its 3D acquisition mode. It provides 47 image planes with an axial field of view of 150 mm (Iida et al., 1998). The effective in-plane resolution was 8 mm full width at half maximum (FWHM) after reconstruction with a Butterworth filter, and the effective axial resolution was 6 mm after combining adjacent images. The transmission scan for attenuation correction was done using a 68Ge-68Ga rod source prior to succeeding emission scans. The PET data were acquired with a bolus injection of $H_2{}^{15}O$ (Raichle et al., 1983; Kanno et al., 1987) but without the arterial blood sampling. Corresponding to the six experimental conditions, six PET scans were repeated with a 10 min $H_2{}^{15}O$ injection interval. The $H_2{}^{15}O$ bolus injection (10 mCi per scan) was commenced, and the data acquisition was started at 15 s and continued for 90 s. We applied the so-called "stop paradigm" (Cherry et al., 1993). The stimulation was started at the same time as the $H_2{}^{15}O$ injection and continued for 65 s, followed by the rest condition. The reconstructed images were obtained in 128 × 128 format with each 2 mm × 2 mm pixel.

2.5. Analyses of BOLD and rCBF data

All activation data were processed using the SPM99 software package (Wellcome Department of Cognitive Neurology, Friston et al., 1995). Standard linear image realignment, linear normalization to the stereotactic anatomical space, and spatial smoothing (3D Gaussian kernel, 6 mm FWHM for the fMRI data, and 16 mm FWHM for the PET data) were successively performed for each subject. All subjects were pooled together and group comparisons were performed using a fixed-effect general linear model. The A, V, and AV data were compared with those of the control task. A significant increase was tested for with t statistics and displayed as statistical parametric maps. The threshold for significance was set at voxel-level $P < 0.05$ (corrected for multiple comparisons) for the fMRI data. For the PET data, however, this threshold was too strict to see differences between the High and Low intelligibility AV conditions. Because we intended to examine differences in brain activation between the two conditions where behavioral data usually show a sharp difference, we displayed the results at a threshold of $P < 0.005$ (uncorrected) and then took clusters with their peak-thresholds $P < 0.002$ (uncorrected) to be significant.

3. Results

3.1. Behavioral performance

As anticipated, the subjects showed a much stronger McGurk effect for lower SN ratios than they did for higher SN ratios. In Fig. 2, percent of auditorily correct responses, averaged across stimuli, is shown for each condition. The size of the McGurk effect (a visual effect) is roughly given as a decrease of auditory responses due to the additional incongruent visual cues (A minus AV). If the size of the visual effect were constant across the SN ratios, the interaction between 'SN (High versus Low)' and 'modality (AV versus A)' would have been null. In fact, the interaction was statistically significant ($F(1, 7) = 9.53$, $P < 0.05$ for fMRI; $F(1, 9) = 11.65$, $P < 0.01$ for PET; two factor ANOVAs), indicating that the visual influence was stronger when speech was harder to hear.

3.2. Brain activation

Each condition (A, V, and AV) was contrasted to the C condition (visual talker identification of still faces). The bimodal audiovisual condition showed noticeable differences between the High and Low intelligibility conditions, whereas the unimodal auditory condition showed essentially identical activations for the two intelligibility conditions (Fig. 3 for fMRI, Fig. 4 for PET).

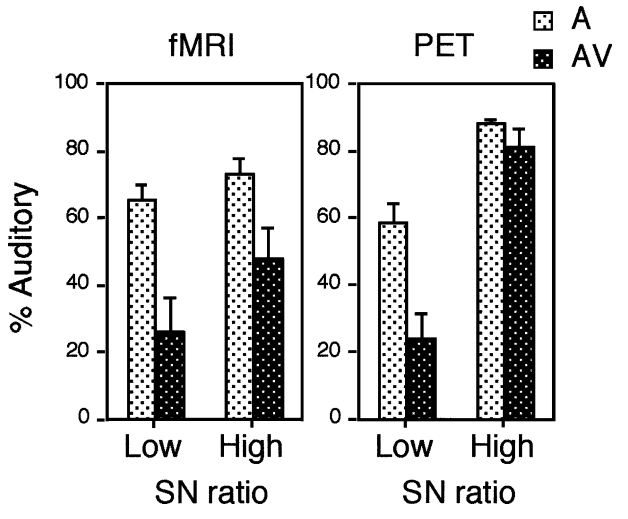

Fig. 2. Behavioral performance in each experiment. Percent auditorily correct responses for the audio-only (A; lighter bars) and incompatible audiovisual (AV; darker bars) stimuli is shown for the Low and High SN ratios (−3 dB and +7 dB for the fMRI, −5 dB and +15 dB for the PET) of the auditory component of the stimuli. Error bars show standard errors. SN: signal-to-noise.

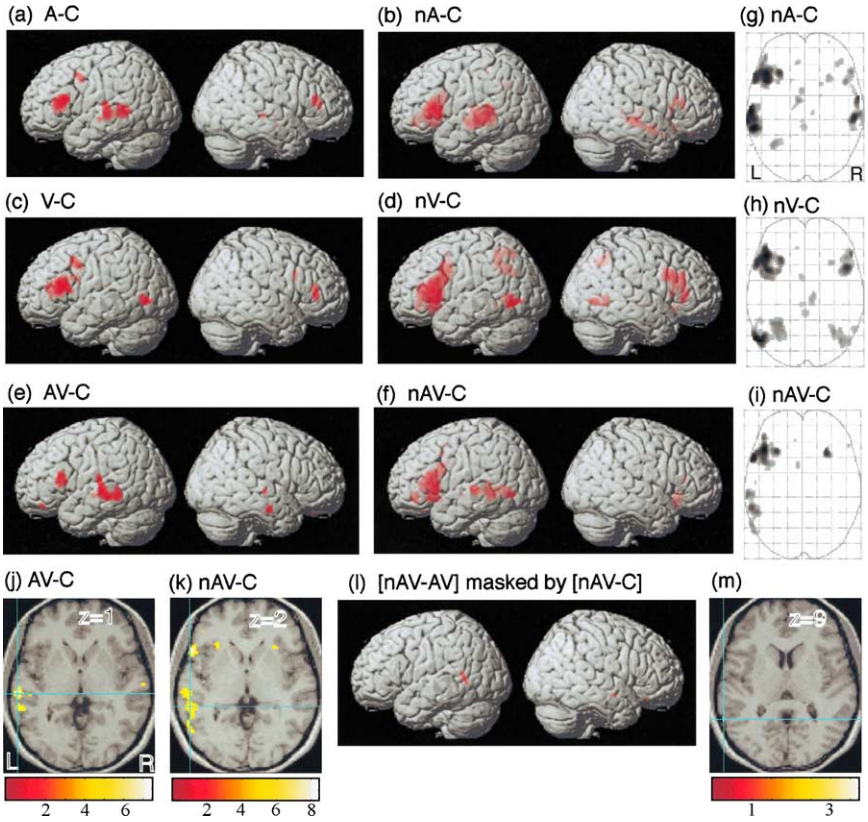

Fig. 3. Suprathreshold voxels in the fMRI experiment ($P < 0.05$, corrected). (a–f) 3D-rendered surface images for the A, nA, V, nV, AV, and nAV conditions contrasted to the C condition. (g, h) Axial projections for the nA, nV, and nAV conditions. (j, k) Axial images for the AV and nAV conditions sectioned at their local maxima in the temporal cortex. (l, m) Results of 'nAV-AV' contrast inclusively masked by 'nAV-C', indicating significant increase in the nAV condition relative to the AV condition ($P < 0.001$, uncorrected) among voxels that were significantly activated in the nAV condition relative to the C condition ($P < 0.05$, uncorrected). The peak coordinates of the cluster in the left temporal cortex was (−56, −49, 9). A: audio-only, V: video-only, AV: audiovisual, nAV: noise-dominant audiovisual, C: control.

In the audio-only conditions (A-C, nA-C), activation common to the fMRI and PET experiments was observed in the temporal cortex bilaterally (Fig. 3a, b and g, Fig. 4a, b and g). The activation was along the superior temporal sulcus (STS), including Brodmann's area (BA) 22, overlapping the so-called 'Wernicke's area'. In the PET experiment, the cluster in the temporal cortex contained the primary auditory cortex (Table 2). In the fMRI nA condition, the angular gyrus (BA 39) was additionally activated (Table 1). Broca's area (BA 44, 45) in the frontal cortex was activated in most audio-only conditions, but not in the PET A condition.

The V and C stimuli differed in the presence or absence of visual speech information given as motion pictures. For the 'V-C' contrast, activation common to the visual-only conditions was seen at the MT/V5 area (BA 37/19), Broca's area (BA 44, 45), and the premotor area (BA 6), mostly bilateral (Figs. 3c, d, h and 4d, h). The cerebellum was often activated (fMRI-nV, PET-V; Tables 1 and 2). Additional cortical activation was seen in the intraparietal sulcus (fMRI-nV, Fig. 3d), middle temporal gyrus (fMRI-nV, Fig. 3d), visual prestriate area (PET-V, Fig. 4d), and superior temporal gyrus (PET-V).

In the audiovisual conditions, the areas activated differed depending on the speech intelligibility. Compare the AV (AV-C) with nAV (nAV-C) activations (Figs. 3e versus f and 4c versus e). In the AV condition, the activated areas were almost the same as those for the audio-only stimuli (A or nA) in both the fMRI and PET experiments (Figs. 3e and 4c).

On the other hand, the activation for the nAV stimuli included a part of the visual-specific area observed in the unimodal V condition (Figs. 3f and 4e). In the nAV condition, the activation in the left temporal cortex extended more posteriorly toward the visual (MT) area (e.g. Figs. 3j versus k and 4j versus k). This is consistent with the behavioral data of the stronger visual influence in the nAV condition than in the AV condition. The large cluster in the left temporal cortex in the nAV condition included the superior temporal gyrus (BA 22, along with the STS) and the lateral occipitotemporal gyrus (BA 37) in both the fMRI and PET experiments. Close to this activation, the left angular gyrus was also activated in the PET experiment along the ascending branch of the STS (BA 22/39, Fig. 4e). It was noticeable that the extended activation from the temporal to more posterior areas was confined to the left hemisphere. This is in striking

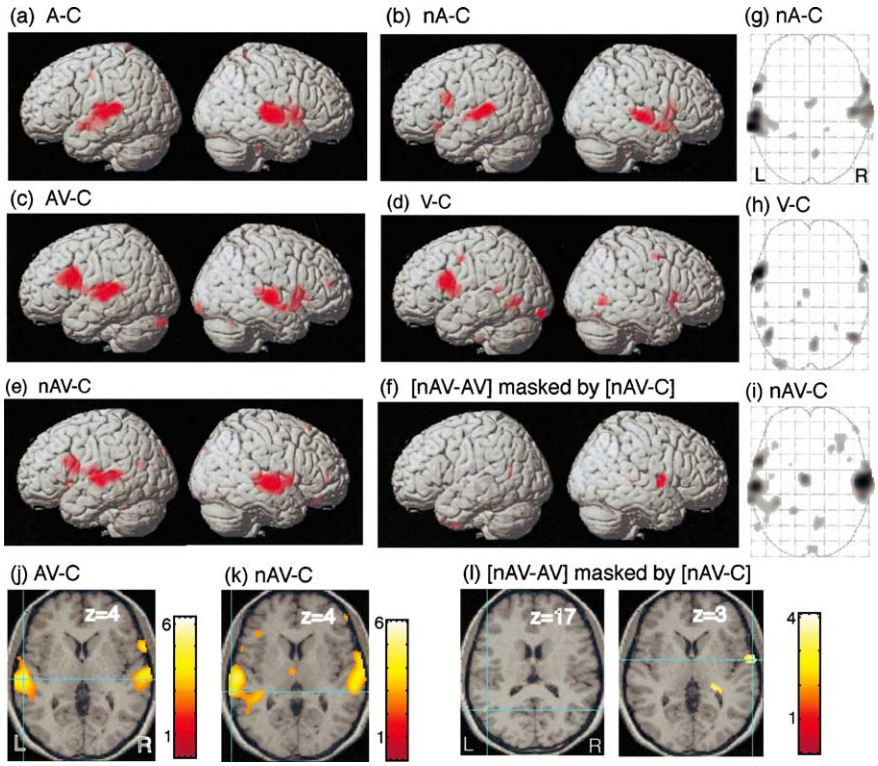

Fig. 4. Suprathreshold voxels in the PET experiment ($P < 0.005$, uncorrected). (a–e) 3D-rendered surface images for the A, nA, AV, V, and nAV conditions contrasted to the C condition. (g–i) Axial projections corresponding for the nA, nV, and nAV conditions. (j, k) Axial images for the AV and nAV conditions sectioned at their local maxima in the let temporal cortex. (f, l). Results of 'nAV-AV' contrast inclusively masked by 'nAV-C', indicating significant increase in the nAV condition relative to the AV condition ($P < 0.005$, uncorrected) among voxels that were significantly activated in the nAV condition relative to the C condition ($P < 0.05$, uncorrected). The peak coordinates of the cluster in the left temporal cortex was (−43, −55, 17). There were also other significant increases in the right temporal cortex, thalamus, and cerebellum. A: audio-only, V: video-only, AV: audiovisual, nAV: noise-dominant audiovisual, C: control.

contrast to the fact that corresponding unimodal activations, that is, in the temporal cortex (the A condition) or in the MT area (the V condition), were mostly observed bilaterally.

Finally, the difference between the nAV and AV conditions were directly tested. The nAV condition was contrasted to the AV condition ($P < 0.001$, uncorrected for fMRI, $P < 0.005$, uncorrected for PET) and masked by the 'nAV-C' contrast inclusively so that the comparison is made using only voxels that reached significance ($P < 0.05$, uncorrected) in the 'nAV-C' contrast. In the fMRI experiment, a significant increase was found in the ventral bank of the posterior STS (BA 21/22) in the left hemisphere (see Fig. 3l and m, Table 3). In the PET experiment, a similar increase was seen in a slightly more posterior region along the ascending branch of the left STS (BA 22/39), as well as in the right superior temporal gyrus (BA 22), thalamus, and cerebellum (see Fig. 4f and l, Table 4).

4. Discussion

4.1. Unimodal processing

For the A stimuli, activation common to the fMRI and PET experiments was observed in the superior temporal gyrus (BA 22) bilaterally. Although so-called Wernicke's area is roughly equivalent to BA 22 in the left hemisphere, recent imaging studies have shown that BA 22 is consistently involved bilaterally in perceiving the acoustic pattern of speech or vocal sounds (Belin et al., 2000; Binder et al., 2000; Scott et al., 2000; Wise et al., 2001). Thus, these bilateral activations in BA 22 are reasonable for auditory speech perception.

Although Broca's area (BA 44, 45) in the frontal cortex was often activated for the A stimuli, this area was activated for the V stimuli as well, indicating that it is not auditory-specific. Moreover, Broca's area was not activated in the PET A condition. These results suggest that the observed activation in Broca's area is related to subvocal rehearsals (Zatorre et al., 1996; Paulesu et al., 1993).

Excluding this Broca's area, the MT area seemed to be a main visual-specific area in the lip-reading task. As the MT area has been implicated in visual motion processing (e.g. Zeki, 1993), this activation is reasonable for visual speech perception.

4.2. Cross-modal binding

By altering the intelligibility of auditory speech, we produced two perceptually different AV situations. In both the

Table 1
Significant activation for each contrast in the fMRI experiment

Anatomical region (BA)	Side	Talairach coordinates (mm)			Volume	*Z*
		x	*y*	*z*		
A-C						
Superior temporal gyrus (22)	L	−55	−20	1	298	6.71
Inferior frontal gyrus (45)	L	−41	26	13	363	6.12
Middle frontal gyrus (6)	L	−43	2	37	41	5.44
Inferior frontal gyrus (45)	R	45	26	17	80	5.23
Superior temporal sulcus (22/21)	R	47	−24	1	22	5.17
V-C						
Inferior frontal gyrus (44/45)	L	−47	18	16	555	7.43
Middle frontal gyrus (6)	L	−43	2	37	82	6.22
Lateral occipitotemporal gyrus (37)	L	−47	−63	5	84	5.65
Inferior frontal gyrus (45/46)	R	41	32	8	114	5.57
Inferior frontal gyrus (44)	R	38	10	27	25	5.26
Inferior frontal gyrus (44)	L	−38	8	23	17	4.94
AV-C						
Superior temporal sulcus (22/21)	L	−55	−22	1	335	7.38
Inferior frontal gyrus (44)	L	−47	16	13	115	6.87
Middle temporal gyrus (21)	R	52	−8	−18	59	5.95
Middle temporal gyrus (21)	L	−45	−28	−8	53	5.57
Superior temporal gyrus (22)	R	57	−12	0	13	5.09
nA-C						
Superior temporal gyrus (22)	L	−50	−34	5	1117	Infinite
Inferior frontal gyrus (44/45)	L	−38	22	10	1540	Infinite
Middle temporal gyrus (21)	R	50	−22	−6	410	7.66
Inferior frontal gyrus (45)	R	47	24	13	194	6.05
Thalamus	L	−4	−3	7	56	5.84
Angular gyrus (39)	L	−31	−53	32	94	5.83
Inferior frontal gyrus (47)	R	24	16	−13	57	5.57
Anterior cingulate gyrus (24)	L	−3	3	39	34	5.41
Posterior cingulate gyrus (23)	L	−1	−32	22	34	5.3
nV-C						
Inferior frontal gyrus (44/45)	L	−38	24	10	2267	Infinite
Lateral occipitotemporal gyrus (37/19)	L	−45	−65	−7	306	Infinite
Inferior frontal gyrus (44)	R	41	10	23	706	7.59
Intraparietal sulcus (7)	L	−26	−45	36	515	6.22
Intraparietal sulcus (7)	R	29	−55	40	138	5.93
Intraparietal sulcus (7)	L	−22	−53	52	73	5.91
Prestriate area (19)	R	36	−65	−5	166	5.90
Thalamus	R	6	−20	−4	108	5.56
Inferior frontal gyrus (47)	R	27	16	−9	37	5.47
Lateral occipitotemporal gyrus (37)	R	52	−51	−4	53	5.38
Middle temporal gyrus (21)	L	−45	−41	−7	15	5.32
Thalamus	L	−1	−34	21	48	5.19
Intraparietal sulcus (7)	L	−33	−61	47	19	5.11
Anterior cingulate gyrus (24)	L	−3	1	35	27	5.08
Cerebellum	L	−3	−24	−19	12	4.86
nAV-C						
Inferior frontal gyrus (47/45)	L	−48	14	−4	981	Infinite
Inferior frontal gyrus (47)	R	24	20	−8	128	7.14
Inferior frontal gyrus (47)	L	−26	22	−4	137	7.12
Middle temporal gyrus (21)	L	−50	−35	2	454	6.58
Inferior frontal gyrus (47)	L	−38	35	−9	66	6.41
Anterior cingulate gyrus (24)	L	−4	4	47	16	5.17

Note: *x*, *y*, *z* coordinates refer to Talairach coordinates of maxima (converted from MNI coordinates of SPM) for activated clusters (thresholded at voxel-level $P < 0.05$, corrected). Clusters of more than 10 voxles are shown here. BA: Brodmann's area.

Table 2
Significant activation for each contrast in the PET experiment

Anatomical region (BA)	Side	Talairach coordinates (mm)			Volume	*Z*
		x	*y*	*z*		
A-C						
Superior temporal gyrus (22)	R	59	−16	4	2355	4.64
Superior temporal gyrus (42/22)	L	−59	−30	8	3052	4.35
Thalamus	L	−3	−6	12	166	3.57
Anterior cingulate gyrus (24)	R	4	−8	27	32	3.15
Prestriate area (19)	L	−1	−84	36	31	3.09
Cerebellum	R	17	−32	−31	56	3.07
Middle frontal gyrus (6/9)	L	−52	−3	40	31	3.04
Superior parietal lobule (7)	R	15	−43	60	36	2.94
nA-C						
Superior temporal gyrus (42/22)	L	−57	−30	8	1151	4.30
Inferior frontal gyrus (44)	L	−54	6	18	213	3.79
Superior temporal gyrus (22)	R	63	−8	4	1251	3.47
Thalamus	L	−1	−4	11	65	3.08
Cerebellum	R	6	−59	−12	56	3.05
Temporal pole (38)	L	−47	16	−12	62	2.85
V-C						
Inferior frontal gyrus (44)	L	−50	6	19	755	4.32
Inferior frontal gyrus (47)	R	56	12	0	164	3.71
Lateral occitotemporal gyrus (37/19)	R	43	−57	−4	204	3.50
Lateral occitotemporal gyrus (37/19)	L	−40	−55	−6	271	3.46
Cerebellum	R	4	−65	−17	244	3.25
Middle frontal gyrus (6/9)	L	−50	−1	40	61	3.24
Prestriate area (18)	L	−29	−88	−18	125	3.12
Cerebellum	L	−10	−20	−38	13	3.02
Superior temporal gyrus (22)	L	−47	−40	12	80	2.97
Thalamus	L	−3	−6	9	52	2.96
Cerebellum	R	54	−4	−42	79	2.92
AV-C						
Inferior frontal gyrus (44)	L	−48	8	21	2890	5.12
Superior temporal gyrus (22)	R	56	−12	0	1581	4.17
Prestriate area (18)	R	24	−94	−17	123	3.41
Cerebellum	L	−26	−80	−30	163	3.08
Cerebellum	R	6	−57	−13	47	3.00
Middle frontal gyrus (46)	R	45	43	14	33	2.96
Cerebellum	R	34	−57	−13	33	2.91
nAV-C						
Superior temporal gyrus (22)	R	54	−10	−2	1801	4.98
Superior temporal gyrus (22)	L	−54	−16	4	2109	4.55
Thalamus	L	−4	−6	12	169	3.85
Cerebellum	L	−15	−41	−30	72	3.28
Cerebellum	R	10	−55	−32	90	3.17
Temporal pole (38)	R	31	22	−31	182	3.17
Cerebellum	R	6	−57	−10	89	3.11
Middle frontal gyrus (46)	R	47	43	7	48	3.07
Prestriate area (19)	R	3	−84	36	52	3.02
Angular gyrus (39)	L	−52	−61	22	46	3.02
Superior frontal gyrus (6)	R	20	20	57	36	2.88
Inferior frontal gyrus (45/44)	L	−34	24	8	33	2.86

Note: *x*, *y*, *z* coordinates refer to Talairach coordinates of maxima (converted from MNI coordinates of SPM) for activated clusters (thresholded at voxel-level $P < 0.005$, uncorrected). Clusters of $P < 0.002$ are shown here. BA: Brodmann's area.

fMRI and PET experiments, we could observe differences in brain activation between the High and Low intelligibility conditions although a less strict threshold was needed for the PET data. When speech was easier to hear, the activated areas were mainly in the temporal cortex, being almost the same as those for the A stimuli. When speech was harder to hear, the activation in the left temporal cortex extended more posteriorly toward the visual-specific activation (the MT area).

Comparing the nAV condition with the AV condition directly, the increased activations were localized in the posterior STS of the left hemisphere for both the fMRI and PET

Table 3
Significant increase in nAV-AV contrast in the fMRI experiment

Anatomical region (BA)	Side	Talairach coordinates (mm)			Volume	Z
		x	y	z		
[nAV-AV] × [nAV-C]						
Superior temporal sulcus (21/22)	L	−56	−49	9	25	3.7

Note: The threshold was voxel-level $P < 0.001$ (uncorrected), masked by nAV-C contrast ($P < 0.05$, uncorrected) inclusively. Clusters of more than 10 voxles are shown here.

experiments although the locations were slightly different between the two experiments. These increases seem to be related to the stronger visual influence (stronger visual attention and/or stronger auditory-visual interaction) that we observed behaviorally.

According to data from monkeys, there are neuroanatomically identified areas within the STS that receive convergent inputs from visual, auditory, and somatosensory cortices (Jones and Powell, 1970; Seltzer and Pandya, 1978). Electrophysiological studies have shown that the STS contain cells that respond to stimulation in more than one sensory modality (Desimone and Gross, 1979; Hikosaka et al., 1988). Although our current results were slightly different between the fMRI and PET experiments, both results indicated inclusion of the posterior region of the STS for the stronger visual influence in auditory-visual speech perception. Our observation of the increased activations in the STS was confined to the left language hemisphere. This suggests that the cross-modal binding occurred as a linguistic event. Therefore, these increases may be related to stronger auditory-visual interaction underlying the McGurk effect, rather than only reflecting a stronger visual attention.

This argument may be supported by some other brain activation studies that employed somewhat different methodologies (Calvert et al., 2000; Callan et al., 2001; Calvert, 2001). Calvert et al. (2000) used audiovisual and unimodal presentation of connected speech in which auditory and visual speech were either congruent or unrelated. They defined brain regions for cross-modal binding as those which show response enhancement to matched audiovisual inputs and response depression to mismatched inputs. Applying these criteria to their fMRI data, they found a cluster of such voxels in the ventral bank of the STS in the left hemisphere ($x = -49$, $y = -50$, $z = 9$). This location is very close to our fMRI result ($x = -56$, $y = -49$, $z = 9$) Callan et al. (2001) reported a single-sweep EEG case study on auditory-visual speech perception. They used audiovisual presentation of spoken words in which auditory and visual speech was concordant or unrelated. These audiovisual stimuli were presented with or without auditory noise. Whereas the unrelated stimuli showed no activation changes due to the noise, the concordant stimuli showed significant activation enhancement in the noise around 200 ms post-stimulus onset at electrodes in temporal and occipital lobes. According to their current source density analysis for this enhancement, there was a component localized in the left superior temporal gyrus. These results, together with the previous results on the McGurk effect by Sams et al. (1991), generally agree with our results. Therefore, the left posterior STS, possibly including its ascending branch, may be related to cross-modal binding of auditory-visual speech.

In our PET experiment, there were also other regions in which the Low intelligibility condition showed stronger activation than the High intelligibility condition (the right BA 22, thalamus, and cerebellum). These activations may indicate larger efforts for auditory speech perception (the right BA 22) and lip-reading (cerebellum) in the Low intelligibility bimodal condition relative to the High condition. The reason why these regions were activated only in the PET experiment may be that the intelligibility difference between the High and Low conditions was much larger in the PET than in the fMRI experiment. The implication of the activation in the thalamus is not clear because it was often observed in the unimodal conditions as well (Tables 1 and 2).

Table 4
Significant increase in nAV-AV contrast in the PET experiment

Anatomical region (BA)	Side	Talairach coordinates (mm)			Volume	Z
		x	y	z		
[nAV-AV] × [nAV-C]						
Superior temporal sulcus (22/6)	R	57	−1	3	180	3.68
Thalamus	R	22	−30	5	61	3.36
Cerebellum	R	12	−49	−32	90	3.3
Superior temporal sulcus (22/39)	L	−43	−55	17	29	2.9

Note: Significant increase was thresholded at $P < 0.005$ (uncorrected), masked by nAV-C contrast ($P < 0.05$, uncorrected) inclusively. Clusters of $P < 0.002$ are shown here.

4.3. Some discrepancies in lip-reading related areas

In the present study, the visual-specific activation for lip-reading was found mainly in the MT area (bilaterally in the PET experiment and in the left hemisphere in the fMRI experiment) that is well documented for its function of visual motion processing (Zeki, 1993). A case study is reported in which a patient with bilateral lesions in area V5/MT was unable to lip-read multisyllabic utterances (Campbell et al., 1997). Thus, the MT may play an essential role in lip-reading.

In the literature, however, brain areas activated for lip-reading in normal hearing are somewhat variable. Calvert and colleagues reported an fMRI study indicating activation in the primary auditory cortex (BA 41) as well as the superior temporal gyrus (BA 22) and visual areas around the MT (Calvert et al., 1997). Later, the activation in the primary auditory cortex was not replicated (Campbell et al., 2001; Bernstein et al., 2002; but also see Calvert and Campbell, 2003), nor was it in this study. The MT area is often activated (Calvert et al., 1997; Campbell et al., 2001) as in the present study. Activation in the superior temporal gyrus (BA 22, often with BA 21) is frequently observed (Calvert et al., 1997; Campbell et al., 2001; Bernstein et al., 2002), and it was also observed in the fMRI-nV condition and the PET experiment in this study. Broca's area is sometimes activated (Campbell et al., 2001) as in our study. The causes of these inconsistencies are not clear due to many experimental differences such as speech stimuli (meaningless monosyllables versus words), the nature of the control condition, and subjects' language background.

4.4. Generality across languages

In the present study, brain activation for the AV stimuli of High intelligibility (+15 dB or +7 dB in SN ratio) was almost identical to that for the audio-only stimuli, lacking the evidence for auditory-visual integration. Although it is in accordance with the relatively weak McGurk effect observed, a question arises. Is this result specific to Japanese speaking subjects? It has been reported that native speakers of Japanese rely on the auditory input more than native speakers of English, showing only a weak McGurk effect unless the auditory speech has some ambiguity (Sekiyama and Tohkura, 1991; Kuhl et al., 1994; Sekiyama, 1994, but also see Massaro et al., 1993). Concerning the data from the High intelligibility condition, generality of the present results across languages remains for further research.

In conclusion, by using Japanese speakers who are less subject to the McGurk effect for intelligible auditory speech, we found a brain region seemingly related to the McGurk effect for less intelligible auditory speech. The region was located in the posterior STS in the left hemisphere. Although our experimental design does not rule out a possibility that this activation reflects stronger visual attention to cope with the Low intelligibility, its location was reasonably close to the area recently reported for cross-modal binding of auditory-visual speech, suggesting that the posterior STS in the left hemisphere plays a role in the McGurk effect.

Acknowledgements

This work was supported by grants from the Science and Technology Agency, Japan, the Ministry of Education, Science, Sports and Culture, Japan, and H. Nakayama Science Promotion Foundation, to the first author. We are grateful to the staff of PET facilities at Akita Research Institute of Brain and Vessels for their support during the PET experiment, and two anonymous reviewers for their helpful comments and suggestions on an earlier version of this article.

References

Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., Pike, B., 2000. Voice-selective areas in human auditory cortex. Nature 403, 309–312.

Bernstein, L.E., Auer Jr., E.T., Moore, J.K., Ponton, C.W., Don, M., Singh, M., 2002. Visual speech perception without primary auditory cortex activation. NeuroReport 13, 311–315.

Bertelson, P., Vroomen, J., de Gelder, B., Driver, J., 2000. The ventriloquist effect does not depend on the direction of deliberate visual attention. Percept. Psychophys. 62, 321–332.

Binder, J.R., Frost, J.A., Hammeke, T.A., Bellgowan, P.S.F., Springer, J.A., Kaufman, J.N., Possing, E.T., 2000. Human temporal lobe activation by speech and nonspeech sounds. Cereb. Cortex 10, 512–528.

Binnie, C.A., Montgomery, A.A., Jackson, P.L., 1974. Auditory and visual contributions to the perception of consonants. J. Speech Hear. Res. 17, 619–630.

Callan, D.E., Callan, A., Kroos, C., Vatikiotis-Bateson, E., 2001. Multimodal contribution to speech perception revealed by independent component analysis: a single-sweep EEG case study. Cogn. Brain Res. 10, 349–353.

Calvert, G.A., 2001. Crossmodal processing in the human brain: insights from functional neuroimaging studies. Cereb. Cortex 11, 1110–1123.

Calvert, G.A., Campbell, R., 2003. Reading speech from still and moving faces: the neural substrates of visible speech. J. Cogn. Neurosci. 15, 57–70.

Calvert, G.A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C.R., McGuire, P.K., Woodruff, P.W.R., Iversen, S.D., David, A.S., 1997. Activation of auditory cortex during silent lipreading. Science 276, 593–596.

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, 649–657.

Campbell, R., Zihl, J., Massaro, D., Munhall, K., Cohen, M.M., 1997. Speechreading in the akinetopsic patient, L.M. Brain 120, 1793–1803.

Campbell, R., Dodd, B., Burnham, D., 1998. Hearing by Eye II. Psychology Press, Hove, UK.

Campbell, R., MacSweeney, M., Surguladze, S., Calvert, G., McGuire, P., Suckling, J., Brammer, M.J., David, A.S., 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). Cogn. Brain Res. 12, 233–243.

Cherry, S.R., Woods, R.P., Mazziotta, J.C., 1993. Improved signal-to-noise in activation studies by exploiting the kinetics of oxygen-15 labelled water. In: Uemura, K. et al. (Eds.), Qualification of Brain Function: Tracer Kinetic and Image Analysis in Brain PET. Elsevier Science Publishers, Tokyo, pp. 79–87.

Desimone, R., Gross, C.G., 1979. Visual areas in the temporal cortex of the macaque. Brain Res. 178, 363–380.

Friston, K., Holmes, A., Worsley, K., Poline, J.B., Frith, C.D., Heather, J.D., Frackowiak, R.S.J., 1995. Statistical parametric maps in functional imaging: a general approach. Hum. Brain Mapp. 2, 189–210.

Hikosaka, K., Iwai, E., Saito, H., Tanaka, K., 1988. Polysensory properties of neurons in the anterior bank of the caudal superior temporal sulcus of the macaque monkey. J. Neurophysiol. 60, 1615–1637.

Iida, H., Miura, S., Shoji, Y., Ogawa, T., Kado, H., Narita, Y., Hatazawa, J., Eberl, S., Kanno, I., Uemura, K., 1998. Non-invasive quantitation of CBF using oxygen-15-water and a dual-PET system. J. Nucl. Med. 39, 1789–1798.

Jones, E.G., Powell, T.P.S., 1970. An anatomical study of converging sensory pathways within the cerebral cortex of the monkey. Brain 93, 793–820.

Kanno, I., Iida, H., Miura, S., Murakami, M., Takahashi, K., Sasaki, H., Inugami, A., Shishido, F., Uemura, K., 1987. A system for cerebral blood flow measurement using an $H_2{}^{15}O$ autoradiographic method and positron emission tomography. J. Cereb. Blood Flow Metab. 7, 143–153.

Kuhl, P.K., Tsuzaki, M., Tohkura, Y., Meltzoff, A.N., 1994. Human processing of auditory-visual information in speech perception: potential for multimodal human-machine interfaces. In: Acoust. Soc. Japan (Ed.), Proceedings of the International Conference on Spoken Language Processing. Acoust. Soc. Japan, Tokyo, pp. 539–542.

MacDonald, J., McGurk, H., 1978. Visual influence on speech perception processes. Percept. Psychophys. 24, 253–257.

Massaro, D.W., 1998. Perceiving Talking Faces: From Speech Perception to Behavioral Principle. MIT Press, Cambridge, MA.

Massaro, D.W., Tsuzaki, M., Cohen, M.M., Gesi, A., Heredia, R., 1993. Bimodal speech perception: an examination across languages. J. Phonetics 21, 445–478.

McGurk, H., MacDonald, J., 1976. Hearing lips and seeing voices. Nature 264, 746–748.

Paulesu, E., Frith, C.D., Frackowiak, R.S., 1993. The neural correlates of the verbal component of working memory. Nature 362, 342–345.

Raichle, M.E., Martin, W.R., Herscovitch, P., Mintun, M.A., Markham, J., 1983. Brain blood flow measured with intravenous $H_2{}^{15}O$. Part II. Implementation and validation. J. Nucl. Med. 24, 790–798.

Rosen, S.M., Fourcin, A.J., Moore, B.C., 1981. Voice pitch as an aid to lipreading. Nature 291, 150–152.

Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O.V., Lu, S., Simola, J., 1991. Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. Neurosci. Lett. 127, 141–145.

Scott, S.K., Blank, C.C., Rosen, S., Wise, R.J.S., 2000. Identification of a pathway for intelligible speech in the left temporal lobe. Brain 123, 2400–2406.

Sekiyama, K., 1994. Differences in auditory-visual speech perception between Japanese and Americans: McGurk effect as a function of incompatibility. J. Acoust. Soc. Jpn. (E) 15, 143–158.

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, 1797–1805.

Sekiyama, K., Tohkura, Y., 1993. Inter-language differences in the influence of visual cues in speech perception. J. Phonetics 21, 427–444.

Sekuler, R., Sekuler, A.B., Lau, R., 1997. Sound alters visual motion processing. Nature 385, 308.

Seltzer, B., Pandya, D.N., 1978. Afferent cortical connections and architectonics of the superior temporal sulcus and surrounding cortex in the rhesus monkey. Brain Res. 149, 1–24.

Shams, L., Kamitani, Y., Shimojo, S., 2000. What you see is what you hear. Nature 408, 788.

Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. 26, 212–215.

Summerfield, Q., 1992. Lipreading and audio-visual speech perception. Phil. Trans. R. Soc. Lond. B 335, 71–78.

Wise, R.J.S., Scott, S.K., Blank, S.C., Mummery, C.J., Murphy, K., Warburton, E.A., 2001. Separate neural subsystems within 'Wernicke's area'. Brain 124, 83–95.

Zatorre, R.J., Meyer, E., Gjedde, A., Evans, A.C., 1996. PET studies of phonetic processing of speech: review, replication, and reanalysis. Cereb. Cortex 6, 21–30.

Zeki, S., 1993. A Vision of the Brain. Blackwell Scientific Publications, Oxford.