6-2004

Multisensory Integration Sites Identified by Perception of Spatial Wavelet Filtered Visual Speech Gesture Information

Daniel E. Callan
ATR International

Jeffery A. Jones
Wilfrid Laurier University, jjones@wlu.ca

Kevin Munhall
Queen's University - Kingston, Ontario

Christian Kroos
Munich University

Akiko M. Callan
ATR International

See next page for additional authors

Multisensory Integration Sites Identified by Perception of Spatial Wavelet Filtered Visual Speech Gesture Information

Daniel E. Callan[1], Jeffery A. Jones[1,2], Kevin Munhall[1,3], Christian Kroos[4], Akiko M. Callan[1], and Eric Vatikiotis-Bateson[1,5]

Abstract

■ Perception of speech is improved when presentation of the audio signal is accompanied by concordant visual speech gesture information. This enhancement is most prevalent when the audio signal is degraded. One potential means by which the brain affords perceptual enhancement is thought to be through the integration of concordant information from multiple sensory channels in a common site of convergence, multisensory integration (MSI) sites. Some studies have identified potential sites in the superior temporal gyrus/sulcus (STG/S) that are responsive to multisensory information from the auditory speech signal and visual speech movement. One limitation of these studies is that they do not control for activity resulting from attentional modulation cued by such things as visual information signaling the onsets and offsets of the acoustic speech signal, as well as activity resulting from MSI of properties of the auditory speech signal with aspects of gross visual motion that are not specific to place of articulation information. This fMRI experiment uses spatial wavelet bandpass filtered Japanese sentences presented with background multispeaker audio noise to discern brain activity reflecting MSI induced by auditory and visual correspondence of place of articulation information that controls for activity resulting from the above-mentioned factors. The experiment consists of a low-frequency (LF) filtered condition containing gross visual motion of the lips, jaw, and head without specific place of articulation information, a midfrequency (MF) filtered condition containing place of articulation information, and an unfiltered (UF) condition. Sites of MSI selectively induced by auditory and visual correspondence of place of articulation information were determined by the presence of activity for both the MF and UF conditions relative to the LF condition. Based on these criteria, sites of MSI were found predominantly in the left middle temporal gyrus (MTG), and the left STG/S (including the auditory cortex). By controlling for additional factors that could also induce greater activity resulting from visual motion information, this study identifies potential MSI sites that we believe are involved with improved speech perception intelligibility. ■

INTRODUCTION

Although most studies concerning perception focus on a single sensory channel, human perceptual experience involves simultaneous stimulation through multiple sensory channels. An investigation of the processes underlying the integration of information between different sensory channels is important given the influence multisensory stimulation has on perception. Presentation of concordant multisensory information relative to unimodal information is known to improve stimulus detection as well as speed up reaction time (Giard & Peronnet, 1999; Hershenson, 1962). Furthermore, perception of a stimulus through a single sensory channel can be altered considerably under conditions in which the same stimulus is presented in conjunction with information from an additional sensory channel. A single flash of light is perceived as multiple flashes when accompanied by multiple auditory beeps (Shams, Kamitani, & Shimojo, 2002). An audio phoneme stimulus when presented with discordant visual phoneme information is perceived as a completely different phoneme than that specified by either the audio or visual channels—''McGurk effect'' (McGurk & MacDonald, 1976). Research concerned with multimodal perception may give insight into the underlying neural systems involved with perceptual experience that cannot be gleaned from investigation of single modalities in isolation.

A potential means by which the brain may integrate information from multiple sensory channels is by convergence. Multisensory integration (MSI) sites that receive converging unimodal input and display distinct neural response properties have been identified in subcortical and cortical brain regions (Stein & Meredith, 1993; Wallace, Meredith, & Stein, 1992) of nonhuman mammals. Rules governing the response properties of neurons involved with MSI have been defined (Stein &

[1]ATR International, [2]Wilfrid Laurier University, [3]Queens University, [4]Munich University, [5]University of British Columbia

Meredith, 1993): The *spatial rule* states that superadditive enhancement (subadditive depression) of multisensory stimulation is dependent on the degree of spatial overlap (divergence) of the receptive fields of unimodal stimulation (superadditivity refers to a response to multisensory stimulation that is greater than the additive combination of unimodal stimulation alone; subadditivity refers to a response to multisensory stimulation that is less than the additive combination of unimodal stimulation alone). The *temporal rule* states that superadditive enhancement of multisensory stimulation is most effective when there is overlap between the peak periods of unimodal stimulation. The *inverse effectiveness rule* states that superadditive enhancement of multisensory stimulation is greatest when the unimodal stimuli are least effective. It is believed that the facilitative effect on perception afforded by concordant stimulation through multiple sensory channels is mediated by the enhancement in response properties of neurons in MSI sites (Stein, Huneycutt, & Meredith, 1988; Stein & Meredith, 1993).

It is well known that the addition of visual speech information improves speech intelligibility over that of auditory speech information alone (Grant & Braida 1991; Sumby & Pollack, 1954). Although the greatest improvement in intelligibility occurs when the audio channel is degraded, improvement in intelligibility also occurs when the audio channel is perfectly clear (Reisberg, Mclean, & Goldfield, 1987) and is worse when audio and visual speech information is incongruent (Dodd, 1977). It has been conjectured that this improvement in speech intelligibility for audiovisual speech may be mediated in part by an MSI site located in the left superior temporal sulcus (STS) (Calvert, Campbell, & Brammer, 2000). The superior temporal gyrus and sulcus (STG/S) responds to auditory speech stimulation (Wise et al., 2001; Binder et al., 2000; Scott, Blank, Rosen, & Wise, 2000), visual speech stimulation (''speechreading'') (Olson, Gatenby, & Gore, 2002; Campbell et al., 2001; MacSweeney et al., 2000, 2001; Calvert et al., 1997; Calvert & Campbell, 2003), as well as audiovisual speech stimulation (Mottonen, Krause, Tiippana, & Sams 2002; Callan, Callan, Kroos, & Vatikiotis-Bateson, 2001; Calvert et al., 2000; Sams et al., 1991). Using fMRI, Calvert et al. (2000) demonstrated that the left STS shows properties consistent with those of an MSI site in that activity to congruent audiovisual speech is superadditive (showing a greater response relative to the sum of the responses of audio and visual speech information presented alone), and that activity to incongruent audiovisual speech is subadditive (showing a reduced response relative to the sum of the responses of audio and visual speech information presented alone). It was pointed out by Calvert (2001) that the inclusion of subadditivity may be too conservative a criteria for MSI given that electrophysiological studies indicate that not all cortical MSI cells that show response enhancement also show response depression (Wallace et al., 1992). Although auditory and visual cortices did not show subadditive responses to incongruent audiovisual speech they did show superadditive responses to congruent audiovisual speech (Calvert et al., 2000), suggesting that they may also potentially serve as multimodal integration sites. These results are consistent with findings showing altered evoked responses to audiovisual speech in the auditory cortex (Mottonen et al., 2002; Sams et al., 1991). Although some studies have reported activity in the auditory cortex to visual speech stimuli alone, supporting the possibility that it may serve as an MSI site (MacSweeney et al., 2001; Calvert et al., 1997; Calvert & Campbell, 2003), this is not true for all studies (Paulesu et al., 2003; Bernstein et al., 2002; Olson et al., 2002; Campbell et al., 2001).

Additional evidence that the STG/S region, including auditory cortex, may be a site of MSI comes from an EEG case study (Callan et al., 2001) as well as an fMRI study (Callan et al., 2003) in which the property of inverse effectiveness was demonstrated for audiovisual speech presented with and without background audio noise. For the EEG study (Callan et al., 2001), spectrotemporal analysis revealed greater enhancement of high-frequency activity (45–70 Hz) for audiovisual stimuli presented with audio noise (AVN) over that of all other conditions: audio only with noise (AON); audiovisual without noise (AV); audio only without noise (AO) (Callan et al., 2001). The site of enhancement was localized to the STG region using current source density analysis constrained by individual specific volume conductor and source models constructed from anatomical MRI data (Callan et al., 2001). Similarly, the fMRI study (Callan et al., 2003) showed significant activity in the middle temporal gyrus (MTG) and the STG/S, including the auditory cortex, in response to the interaction of (AVN–AON)–(AV–AO). Consistent with the principle of inverse effectiveness, these studies (Callan et al., 2001, 2003) demonstrate that enhancement of audiovisual speech information is greatest when the unimodal audio stimuli, due to the addition of audio noise, are least effective.

Although the studies reviewed above (Callan et al., 2001, 2003; Calvert et al., 2000) support the claim that the STG/S, including the auditory cortex, may serve as MSI sites for audiovisual speech perception, some studies have failed to demonstrate properties of MSI in the STG/S including the auditory cortex (Jones & Callan, 2003; Calvert et al., 1999). In an fMRI study of the McGurk effect conducted by Jones and Callan (2003) greater responses in the STS/G for congruent audiovisual stimuli (/aba/) were not observed over incongruent audiovisual stimuli (audio /aba/ paired with visual /ava/), as one would predict for an MSI site. Calvert et al. (1999) failed to show greater activation in an fMRI study in the STS for audiovisual stimuli (consisting of numbers between 1 and 10) over that of audio-only stimuli; however, greater activation was observed in the auditory cortex. In

an fMRI study conducted by Olson et al. (2002), significantly greater activity in synchronized audiovisual speech was not found from that of temporally delayed audiovisual speech, suggesting a violation in the temporal rule for a site of MSI. However, the audiovisual stimuli used in the experiment were designed to induce a McGurk effect (audio and visual signals were incongruent) and thus may not meet the criteria sufficient to induce enhanced activity in an MSI site. Differences between studies that support the STG/S as an MSI site (Callan et al., 2001, 2003; Calvert et al., 2000) and those that do not (Jones and Callan, 2003; Calvert et al., 1999) appear to lie in the nature of the stimuli used (e.g., sentence level vs. words or nonwords) or the presence of background noise (thought to induce a greater enhancement effect due to the rule of inverse effectiveness).

The purpose of this study was to induce activity in multisensory sites selective to speech gesture information signaling place of articulation. The speech gesture is defined as biological motion of the various articulators (e.g., jaw, lips, tongue, larynx) that specify vocal tract shape. A direct relationship is known to exist between vocal tract shape, speech acoustics, and deformation of the face (Munhall & Vatikiotis-Bateson, 1998; Yehia, Rubin, & Vatiktiotis-Bateson, 1998; Vatikiotis-Bateson, Munhall, Hirayama, Lee, & Terzopoulos, 1996). Observation of the articulators during speech can give direct information regarding place of articulation. In addition, speech gesture information signals the onset, offset, and rate of change of the acoustic speech signal. It also gives information concerning the overall amplitude contour (Grant & Seitz, 2000) as well as the spectral region of the acoustic signal (Grant, 2001). Even head motion itself is highly correlated with changes in fundamental frequency (Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004; Yehia et al., 1998).

All the above-mentioned correspondences between the auditory signal and visual speech gesture information could potentially be processed in MSI sites in the STG/S. It is entirely possible that enhanced activity in the STG/S reported by Calvert et al. (2000) and Callan et al. (2001, 2003) may be the result of MSI of gross properties of visual motion of the head, jaw, and lips specifying the onset, offset, and rate of change of the acoustic speech signal and/or changes in fundamental frequency information instead of by visual speech gesture information signaling place of articulation (thought to be the dominant cue used during speechreading). The finding that the STS shows properties of MSI to nonspeech audiovisual stimuli (Calvert, Hansen, Iversen, & Brammer, 2001; Giard & Peronnet, 1999) as well as to auditory–somatosensory (Foxe et al., 2002) stimuli testifies to the general strength that correspondence to gross properties in different sensory channels has in activating this region. It should also be pointed out that enhanced activation of the STG/S may not reflect MSI of visual information at all, but rather may reflect greater attention (Jancke, Mirzazade, & Shah, 1999) to the onset and/or offset of the auditory stimuli cued by visual information.

In this fMRI study we attempted to control for gross properties of visual motion that may enhance activity due to greater attentional modulation and/or may activate MSI sites not specific to speech gesture place of articulation information. In order to enhance the likelihood of activation of MSI sites, sentences were used as stimuli and were presented with background multispeaker babble (thought to induce a greater enhancement effect due to the rule of inverse effectiveness). In a series of experiments from our laboratory, spatial frequency wavelet band-pass filtered sentence stimuli were presented to English-speaking subjects in the presence of auditory noise (Munhall, Kroos, & Vatikiotis-Bateson, 2002). The stimuli were constructed using the procedure specified by Kroos, Kuratate, and Vatikiotis-Bateson (2002). All of the experiments (Munhall et al., 2002) demonstrated that visual speech information is present across a number of bands and that watching these bands significantly enhanced the perception of speech in noise. The performance demonstrated an inverted U function with a peak in the 10–15 cycles per face (cpf) range. The stimuli used in the fMRI study reported here consisted of Japanese audiovisual sentences with low predictability presented with background multispeaker babble. The sentences were spatially wavelet band-pass filtered using the procedure specified by Kroos et al. (2002). The experiment was conducted in Japan so Japanese stimuli were used. Although native Japanese speakers show a smaller McGurk effect than that of native English speakers, Japanese speakers do show a strong McGurk effect for audiovisual speech in the presence of noise (Sekiyama & Tohkura, 1991) indicating that Japanese speakers do use visual speech information during speech perception especially under noisy conditions. Our experiment consisted of the following four conditions (Figure 1): an unfiltered (UF) audiovisual condition; a midfrequency (MF) 11-cpf audiovisual condition; a low-frequency (LF) 2.7-cpf, audiovisual condition; and a baseline condition consisting of a video of a still face presented without multispeaker babble. The baseline condition was used to control for activation of visual areas of the brain resulting from visual presentation of a face. The task for the subjects was to passively identify as many phonemes presented as possible. Only gross properties of lip, jaw, and head movement with very little place of articulation information is probably available in the LF condition (2.7 cpf) because behavioral results using English subjects and sentences showed that performance was no different than audio only with noise (Munhall et al., 2002). The midfrequency MF condition (11 cpf) consisted of detailed properties of lip and face movement such that place of articulation information was preserved. Behavioral results using English subjects and sentences show that performance

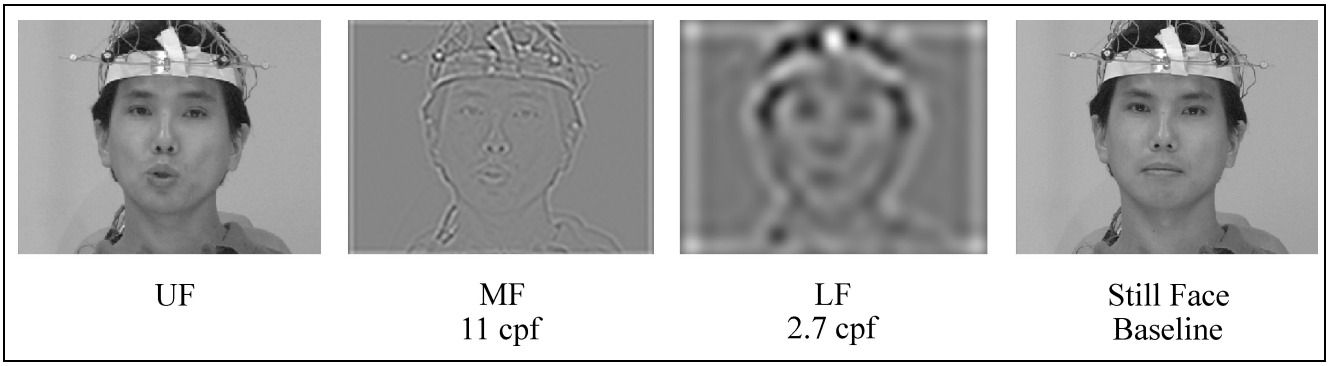

Figure 1. Representative images of the conditions used in this study. The unfiltered (UF) condition contains all visual speech gesture motion. The spatial midfrequency (MF) wavelet band-pass filtered condition maintains place of articulation information and the spatial low-frequency (LF) wavelet band-pass filtered condition consists of mainly gross properties of movement of the lips, jaw, and head. Also shown is the still face baseline condition. The object worn on the head of the speaker in the video contained optical markers used for head tracking necessary for wavelet filtering using the method of Kroos et al. (2002). cpf = cycles per face.

is higher than all other wavelet band-pass filtered stimuli but are not as high as unfiltered stimuli (Munhall et al., 2002). It is possible that the UF condition contains some place of articulation information that is degraded for the MF condition.

Utilizing the spatial wavelet band-pass filtered stimuli discussed above, one finds it is possible to discern brain activity potentially reflecting MSI, considered to be induced by correspondence between properties of the auditory speech signal and visual speech gestures containing place of articulation information. The UF condition contains all types of visual speech motion including gross properties of visual motion of the head, jaw, and lips specifying the onset, offset, and rate of change of the acoustic speech signal and/or changes in fundamental frequency as well as visual speech gesture information signaling place of articulation (Munhall et al., 2002, 2004). The MF condition is believed to contain visual speech gesture information signaling place of articulation to a much greater extent than the LF condition, which is thought to contain predominantly gross properties of visual motion of the head, jaw, and lips specifying the onset, offset, and rate of change of the acoustic speech signal and/or changes in fundamental frequency (Munhall et al., 2002, 2004). MSI sites coding for correspondence between speech gesture information and the auditory speech signal can be discerned by locating activity common to both the UF and MF conditions that is greater than the LF condition (activity restricted to visual speech motion information in the spatial frequency range of the MF condition that does not result from gross properties of visual motion). This study is restricted to investigation of MSI sites responsive to visual speech gesture information present in the MF condition as well as in the UF condition. This study could potentially miss activity reflecting MSI of speech gesture information that codes for place of articulation present in the UF but not in the MF condition. However, this is considered acceptable in order to control for activity resulting from MSI of properties of the auditory speech signal with aspects of gross visual motion that are not specific to place of articulation information, as well as activity resulting from attentional modulation cued by such things as visual information signaling the onsets and offsets of the acoustic speech signal. Based on the results of Calvert et al. (2000) and Callan et al. (2001, 2003) it is predicted that MSI sites that code for the correspondence between visual speech gestures depicting place of articulation information and the auditory speech signal will be found in the STG/S.

RESULTS

Behavioral Performance

Although objective measures of behavioral speech perception performance were not taken, subjects were asked to qualitatively rate the degree of benefit that each of the visual speech motion conditions afforded. Nine of the 11 subjects progressively rated the UF condition as affording the most benefit followed by the MF condition and then the LF condition. One subject reported the UF condition affording benefit and the MF and LF not affording benefit. Another subject reported no difference in benefit between any of the conditions. Results of a chi-square test indicate that the number of cases falling into each of the three observed patterns of classification were significantly different from the expected number based on chance ($\chi^2 = 11.6$, $p < .005$, $df = 2$). It should be pointed out that for the purpose of this experiment it is not imperative that enhanced behavioral performance be demonstrated given that MSI is expected to occur for concordant information in audio and visual channels whether it affords benefit or not. However, because many electrophysiological studies conducted on nonhuman animals have reported behavioral improvement associated with activity in MSI

sites (Stein et al., 1988; Stein & Meridith, 1993), the results of qualitative reports made by subjects suggesting that the UF condition affords the most perceptual enhancement followed by the MF condition, which shows enhancement over the LF condition are relevant. The qualitative results reported here are consistent with the objective measures of behavioral performance using similar stimuli with native English speakers that showed significantly greater intelligibility for the UF condition over that of the MF condition and for the MF condition over that of the LF condition (Munhall et al., 2002).

Brain Imaging

Regional brain activity for the various conditions was assessed using statistical parametric mapping SPM (SPM2b, Wellcome Department of Cognitive Neurology, University College London) in which a general linear model was employed, using a boxcar function convolved with a hemodynamic response function with time derivatives. Additionally, global normalization and grand mean scaling was carried out. A fixed-effect analysis was first employed for all contrasts of interest across data from each subject. The baseline, still face condition was implicitly modeled in the design. Following this procedure, a random-effects level, between-subjects analysis of the UF condition was conducted. All further analyses were restricted to voxels significant in the UF condition to ensure that only suprathreshold differences in activity relative to the baseline still face condition were determined. Otherwise, differences in contrasts involving the various conditions could be a result of subthreshold activity with respect to the baseline condition being less for one condition over the other. The location of activity for the UF–LF and the UF–MF contrasts was restricted by finding the intersection of statistically significant voxels with those of the UF contrast: (UF–LF ∩ UF) and (UF–MF ∩ UF). The location of activity for the MF–LF contrast was restricted by means of inclusive masking with significant voxels for the UF contrast. The use of inclusive masking for this contrast but not the others is deemed acceptable because the MF–LF contrast and the UF contrast are orthogonal. Significance values were corrected with reference to the mask instead of the whole brain as were carried out in the other analyses. Multiple comparisons for all contrasts were controlled for by adjusting the T threshold using the false discovery rate (FDR) procedure (Genovese, Lazar, & Nichols, 2002). A spatial extent of five voxels was used for all contrasts of interest. Inclusive masking was based on $p < .05$ FDR corrected threshold. The location of active voxels was determined by reference to the Talairach atlas (Talairach & Tournoux, 1988) after transforming from the MNI to the Talairach coordinate system (www.mrc-cbu.cam.ac.uk/Imaging/mnispace.html). Activity in the auditory cortex was identified with reference to the atlas given by Rademacher et al. (2001).

Figure 2 and Table 1 depict the results of the SPM analysis for all contrasts of interest: (A) UF relative to the baseline still face condition; (B) UF–LF ∩ UF; (C) UF–MF ∩ UF; (D) MF–LF inclusively masked by UF; and (E) (UF–LF ∩ UF) ∩ (MF–LF inclusively masked by UF) (referred to from here on as MF–LF ∩ UF–LF). All contrasts of interest show activity in the MTG and the STG/S including the auditory cortex (the MF–LF ∩ UF–LF contrast shows considerably more left- than right-hemisphere activity). The UF and the UF–LF ∩ UF contrasts additionally show activity in the middle occipital gyrus (MOG), the hippocampus, and the anterior insula claustrum region (Figure 2A and B, Table 1). The UF and the UF–MF ∩ UF contrasts additionally show activity in the right middle temporal occipital region, thought to be visual motion processing area MT+ (Figure 2A and C, Table 1). The UF, the UF–LF ∩ UF, and the UF–MF ∩ UF contrasts all show activity in the premotor cortex (PMC) (Figure 2A–C, Table 1).

DISCUSSION

The finding of primary interest in this study is the presence of activity consistent with a site of MSI induced by auditory and visual correspondence of place of articulation information (MF–LF ∩ UF–LF) located in the mid to anterior MTG and the STG/S (including the auditory cortex) (Figure 2E, Table 1). In agreement with aspects of findings reported by Calvert et al. (1999, 2000) and Callan et al. (2001, 2003) the results of this study support the proposal that the MTG and STG/S may be sites of MSI of the properties of the auditory speech signal and visual speech gesture information that codes for place of articulation. The presence of strongly left lateralized activity for the MF–LF ∩ UF–LF contrast, thought to reflect MSI of auditory and visual place of articulation information, is consistent with the idea of left-hemisphere dominance for language processing (Figure 2E). It is interesting to point out that the same left mid to anterior STS region found to be a site of MSI in this study was also found to be selectively responsive to the presence of phonetic features over that of acoustically complex signals (Scott et al., 2000). Based on the results of this experiment, it is possible that this region of the mid to anterior STS may not only be selective for processing of phonetic information but may be multisensory in nature. The activity found in the right MOG for the MF–LF ∩ UF–LF contrast probably reflects differences in visual processing of MF aspects of the stimulus that are also present in the UF condition but not present in the LF condition (Figure 2E, Table 1). A limitation of the findings reported here as well as in other studies investigating MSI (Callan et al., 2003; Calvert et al., 1999, 2000) is that activity at a specific site cannot be discerned as being directly the result of MSI or the result of

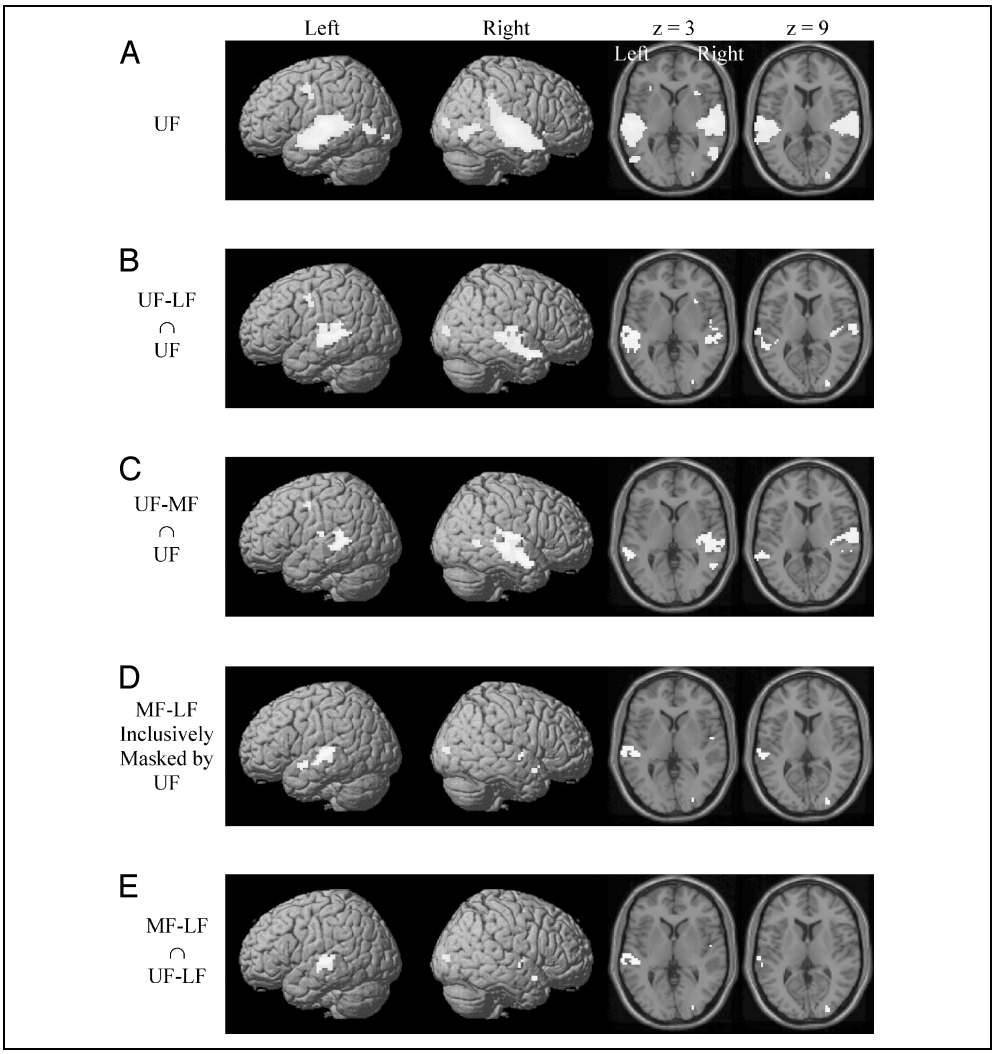

Figure 2. Statistical parametric maps showing brain activity (white) for the contrasts of interest ($p < .05$, FDR corrected, spatial extent threshold = 5 voxels, $df = 10$). (A) Unfiltered (UF) relative to the baseline contrast. (B) The intersection of the UF relative to low-frequency (LF) contrast with that of the UF relative to the baseline contrast. (C) The intersection of the UF relative to midfrequency (MF) contrast with that of the UF relative to the baseline contrast. (D) The MF relative to the LF contrast inclusively masked by the UF relative to the baseline contrast. (E) Activity thought to represent sites of MSI for place of articulation information: Activity present for both the MF–LF and the UF–LF contrasts; the intersection of results from (B) and (D). The horizontal slices at the Talairach coordinate of $z = 3$ and $z = 9$ shows auditory cortex activity 3(A through E) and anterior insula/claustrum activity (only A and B).

modulation from an active MSI site located elsewhere in the brain.

The results (Figure 2, Table 1) suggest that the activity found is not due to MSI of aspects of gross visual motion with properties of the acoustic speech signal. Correlation between visualized head movement with changes in fundamental frequency (Yehia et al., 1998), related to the intonation contours of speech, may be processed by MSI sites. Furthermore, it has been found that visual observation of head movement during a speech in noise task significantly enhances perceptual performance (Munhall et al., 2004). It is possible that findings reported in previous studies (Callan et al., 2001; Calvert et al., 2000) result from activation of MSI sites coding for correlation between head movement and fundamental frequency rather than between visual place of articulation gesture information and the auditory speech signal. One reason why this is unlikely to be the case is because stimuli used in the experiments (Callan et al., 2001, 2003; Calvert et al., 2000) probably do not contain much head movement resulting from the unnatural recording conditions used to make them. In any case, the potential confound of activating an MSI site involved with correlation between gross visual movement of the lips, jaw, and head with that of properties of the auditory speech signal can be accounted for in this study because this information is present in all conditions.

The results (Figure 2, Table 1) further suggest that the activity found is not due to attentional modulation cued by such things as visual information signaling the onsets and offsets of the acoustic speech signal. Attention has been shown to enhance activity in the auditory cortex and the STG (Jancke et al., 1999). Visual cues that signal the onset and offset of speech stimuli may be used by the listener to attend to specific aspects of the auditory signal, resulting in greater auditory cortex and STG activity. In this study, because the LF condition consists of gross visual properties of lip, jaw, and head movement sufficient to cue the onsets and offsets of the auditory speech signal, attentional modulation to these cues is controlled for. However, it is possible that activity found in this study that is thought to reflect MSI is actually the result of greater attentional modulation to auditory correlates of visual motion cues present in the

Table 1. Talairach Coordinates

	UF (Figure 2A)				*UF–LF ∩ UF (Figure 2B)*				*UF–MF ∩ UF (Figure 2C)*				*MF–LF Inc. Masked UF (Figure 2D)*				*MF–LF ∩ UF–LF (Figure 2E)*			
Brain Region	*x*	*y*	*z*	*p<*	*x*	*y*	*z*	*p<*	*x*	*y*	*z*	*p<*	*x*	*y*	*z*	*p<*	*x*	*y*	*z*	*p<*
STG/S	−56	−20	7	.001	−50	−26	1	.033	−63	−39	15	.017	−48	−26	−1	.047	−48	−26	−1	.047
	50	−20	1	.001	48	−26	−1	.033	48	−26	−1	.008	53	−6	0	.047	53	−9	−2	.047
MTG	−53	−15	−12	.001	−52	−32	−1	.04	−63	−33	0	.034	−62	−15	−4	.047	−62	−15	−4	.047
	53	−15	−7	.003	56	8	−18	.033	54	−24	−6	.03	−53	0	−8	.047				
													59	8	−13	.047				
Auditory cortex	−49	−23	5	.001	−49	−23	5	.034					−49	−23	5	.047	−49	−23	5	.047
	46	−24	5	.001	46	−24	5	.038	45	−29	5	.034	46	−24	5	.047				
Ant. insula	−30	26	0	.033																
claustrum	33	21	4	.028	30	18	5	.042												
PMC	−62	−7	36	.012	−62	−7	36	.033	−53	−4	41	0.03								
OT junct. (MT+)	−48	−73	1	.008																
	56	−67	3	.006					53	−52	8	0.017								
MOG	27	−87	10	.027	27	−87	15	.033					27	−90	10	.047	27	−90	10	.047
IOG	−39	−88	−3	.033																
Hipp.	−33	−24	−9	.009	−33	−27	−9	.036												

Note: Talairach (*x*, *y*, *z*) coordinates for each brain region were selected based on the location of local maxima of brain activity falling within regions defined by Talairach and Tournoux (1988). Left-hemisphere activity is denoted by negative *x* values; right-hemisphere activity is denoted by positive *x* values. False discovery rate (FDR) corrected *p* values are given for each coordinate. UF = unfiltered; MF = midfrequency 11 cpf; LF = low-frequency; Inc. masked = inclusively masked; ∩ = intersection (logical and); MT+ = brain region in the middle temporal area near the occipito-temporal junction (OT junct.) associated with visual motion processing. STG/S = superior temporal gyrus/sulcus; MTG = middle temporal gyrus; Ant. insula = anterior insula; MOG = middle occipital gyrus; IOG = inferior occipital gyrus; Hipp. = hippocampus.

UF and MF conditions that are not present in the LF condition. It is unlikely that modulation resulting from attention to general motion is entirely responsible for activity in sites thought to reflect MSI in this study because one would also expect attentional modulation to occur in brain regions involved with visual motion perception (middle temporal MT+ area near the occipito-temporal junction). Yet this is not the case for the contrast reflecting MSI (MF–LF ∩ UF–LF) (Figure 2E, Table1). This, however, does not rule out attentional modulation specific to biological motion that may occur in potential MSI sites in the mid to anterior MTG and STG/S and may be misconstrued as MSI itself.

Biological motion related to speech (Olson et al., 2002; Campbell et al., 2001; Macsweeney et al., 2000, 2001; Calvert et al., 1997; Calvert & Campbell, 2003) as well as nonspeech movement (reported in Allison, Puce, & McCarthy, 2000) has been found to activate regions implicated as MSI sites in this and other studies (Callan et al., 2001; Calvert et al., 2000). Studies investigating visual observation of biological motion of speech-related movement (Olson et al., 2002; Campbell et al., 2001; Macsweeney et al., 2000, 2001; Calvert et al., 1997; Calvert & Campbell, 2003) as well as non-speech-related movement of the hands, eyes, body, and mouth (Allison et al., 2000) show activity in the STG/S. This region appears to be responsive to both actual biological motion and implied biological motion, such as in the case of stimuli consisting of still pictures of speech articulation (Calvert & Campbell, 2003; Nishitani & Hari, 2002). Although it has been claimed that there is a distinct region in the STS that is selectively responsive to speech gestures (Calvert et al., 1997), in a review of the biological motion literature, Allison et al. (2000) suggests that this may not be the case. Given that studies investigating unimodal visual observation of biological motion related to speech (Olson et al., 2002; Campbell et al., 2001; Macsweeney et al., 2000, 2001; Calvert et al., 1997; Calvert & Campbell, 2003) indicate activity in the same regions (STG/S) as audiovisual speech perception (Callan et al., 2001; Calvert et al., 2000) it is difficult to discern activity reflective of MSI of biological motion related to speech.

Some evidence that activity in the mid to anterior MTG and STG/S is a result of MSI of audio and visual speech information rather than resulting from observation of biological motion comes from studies conducted by Calvert et al. (2000) and Callan et al. (2001, 2003). Calvert et al. controlled for observation of biological motion by requiring that the audiovisual condition show superadditivity characteristic of a multisensory site, such that the audiovisual condition was greater than the sum of the audio and visual conditions presented unimodally. Callan et al. controlled for observation of biological motion by demonstrating the principle of inverse effectiveness such that enhancement to audiovisual speech is greatest when the audio channel is degraded by noise. In the Callan et al. studies, given that the visual speech stimuli were the same in both audiovisual conditions, one would not expect a difference in enhancement by visual observation of biological motion to be dependent on the presence or absence of audio noise. Although it is possible that activity in the mid to anterior MTG and STG/S found in this study may result from visual observation of biological motion, the studies reported above lend some support to the conclusion that STG/S is a site of MSI. However, attentional modulation specific to observation (visual and/or auditory) of speech-related biological motion occurring in the STG/S may be responsible for activity found in this study, as well as that of other studies reported above (Callan et al., 2001, 2003; Calvert et al., 2000), rather than resulting from processes related solely to MSI.

Another potential confound that occurs in this study as well as others (Callan et al., 2001, 2003; Calvert et al., 2000) is the activation of regions of the STG/S in response to processes related to better retrieval from semantic memory (Wise et al., 2001) as well as processes related to better phonetic perception (Scott et al., 2000) resulting from enhanced intelligibility due to properties of MSI or other factors such as attentional modulation of speech-specific biological motion information. Wise et al. (2001) identified a site in the posterior STS that acts as an interface between perception and retrieval of information from semantic memory. Activity reflecting a site of MSI in the mid to anterior STG/S (Figure 2E, Table 1) found in this study is unlikely to be a result of processes related to semantic retrieval, which is reported to involve more posterior regions of STG/S (Wise et al., 2001). Significant differences in activity are present for the UF–MF ∩ UF contrast (Figure 2C) in the posterior STG/S region but not in the mid to anterior STG/S region found to be a site of MSI (Figure 2E). On the other hand, some portion of the left mid to anterior STG/S region considered a site of MSI of audiovisual place of articulation information in our study (Figure 2E, Table 1) is also selectively activated by the presence of phonetic features over that of acoustically complex signals (Scott et al., 2000). Although activation of this region of the mid to anterior STG/S as a result of better intelligibility cannot be ruled out, it is entirely possible that this region is also multisensory in nature.

Although activity was found consistent with sites of MSI of place of articulation information in the mid to anterior MTG and the STG/S, including the auditory cortex, there was also considerable activity found for the UF–LF ∩ UF contrast and UF–MF ∩ UF contrast that is thought to be the result of different processes (Figure 2, Table 1). Regions activated by the UF condition to a greater extent than either the LF or the MF conditions include portions of the posterior MTG, STG/S, and the PMC. Regions activated only for the UF–LF ∩ UF contrast additionally include the anterior insula claustrum region, and the hippocampus, whereas regions

activated only for the UF–MF ∩ UF contrast additionally include the middle temporal area near the occipito-temporal junction. For the MF–LF inclusively masked by UF, contrast activity was present in the anterior STG/S. There are many potential causes that could be responsible for this enhanced activity.

One of the more obvious reasons for enhanced activity concerns differences in general properties of the visual stimuli for the various conditions. Greater activity for the UF–MF ∩ UF contrast in the occipital–temporal junction in region MT+ thought to reflect general visual motion processing (Figure 2, Table 1) is likely the result of greater visual motion information in the UF condition relative to the MF conditions in which some of this information is likely to be degraded.

Enhanced activity for the UF over the MF and LF conditions (Figure 2B and C) may result from MSI of the correlation between properties of the auditory signal with that of visual aspects of biological motion (present in the UF but not the MF and LF conditions) that are not related to place of articulation information. Several studies have reported activity consistent with MSI for speech (Olson et al., 2002; Callan et al., 2001; Calvert et al., 1999, 2000) and nonspeech stimuli (Foxe et al., 2002; Bushara, Grafman, & Hallett, 2001; Calvert et al., 2001; Banati, Goerres, Tjoa, Aggleton, & Grasby, 2000; Lewis, Beauchamp, & DeYoe, 2000; Hadjikhani & Roland, 1998) in the MTG, STG/S, as well as the insula (claustrum) region. The claustrum contains topographic maps of the auditory, visual, and somatosensory cortices and has been conjectured to be involved with relaying information between these regions (Olson et al., 2002; Ettinger & Wilson, 1990). The insula has been conjectured to be involved with processing of the temporal coherence of multisensory stimulation (Bushara et al., 2001).

Enhanced phonetic intelligibility and/or retrieval from semantic memory may be responsible for greater activity in the UF condition relative to the MF and LF conditions (Figure 2B and C, Table 1). Based on subjective reports of intelligibility in this study as well as objective behavioral performance tests using similar stimuli (Munhall et al., 2002), the UF condition enhances perceptual intelligibility to a greater extent than the MF or the LF conditions. Better performance for the UF condition may be reflected by enhanced activity in the posterior STG/S and MTG, as well as PMC in regions known to be involved with processing of phonetic information (Scott et al., 2000; Zatorre & Binder, 2000) and retrieval from semantic memory (Wise et al., 2001). Activity was present for the MF–LF inclusively masked by UF contrast in the anterior STG/S in a region shown by Scott et al. (2000) to be selectively responsive to intelligible phonetic information relative to stimuli with equivalent acoustic complexity.

Biological motion present to a greater extent for the UF condition than for the MF and LF conditions may be responsible for the enhanced activity observed (Figure 2B and C, Table 1). It is entirely possible that degrading the visual stimuli by spatial wavelet band-pass filtering reduces the amount and/or the type of biological motion in the stimuli. Greater activity for the UF condition relative to the MF and LF conditions is found in regions of the STG/S known to be responsive to biological motion of speech (Olson et al., 2002; Campbell et al., 2001; Macsweeney et al., 2000, 2001; Calvert et al., 1997; Calvert & Campbell, 2003) and nonspeech stimuli (Allison et al., 2000).

Several studies have reported activity in brain regions involved with planning and execution of speech production (Broca's area, PMC, anterior insula) (Callan, Callan, Honda, & Masaki, 2000; Kent & Tjaden, 1997) in response to visual speech gesture information (Callan et al., 2003; Paulesu et al., 2003; Calvert & Campbell, 2003; Bernstein et al., 2002; Nishitani & Hari, 2002; Olson et al., 2002; Campbell et al., 2001). However, this is not true of all studies (MacSweeney et al., 2001; Calvert et al., 1997, 1999, 2000). Activity found in speech motor regions in response to implied (Calvert & Campbell, 2003; Nishitani & Hari, 2002) as well as actual (Bernstein et al., 2002; Olson et al., 2002; Campbell et al., 2001) visual biological motion of speech gestures is interesting, given recent claims of involvement of the ''mirror neuron system'' (Callan et al., 2003; Paulesu et al., 2003; Calvert & Campbell, 2003; Nishitani & Hari, 2002; Campbell et al., 2001). The mirror neuron system is thought to reflect processing in brain regions involved with producing certain gestures during perception of these same or similar gestures (Rizzolatti & Arbib, 1998). With regard to speech, it is possible that brain regions involved with the planning and execution of articulation are used to simulate the intended speech act of the observed speaker given visual speech gesture information. Activity found in this study in the PMC and the anterior insula are consistent with this hypothesis. The anterior insula has been implicated with processes related to speech production planning (Dronkers, 1996). Although only right anterior insula activity is denoted in Figure 2B and Table 1 a cluster of four voxels was also present in the left anterior insula at coordinates $x = -30$, $y = 23$, $z = 2$. The UF relative to baseline contrast shows activity in both left and right anterior insula (Figure 2A, Table 1). The site of PMC activity found in this study (Figure 2A–C, Table 1) has been identified as being activated during speech production tasks (Wildgruber, Ackermann, & Grodd, 2001; Lotze, Seggewies, Erb, Grodd, & Birbaumer, 2000). It is unclear why Broca's area was not found to be active in this study, even for the UF relative to baseline contrast (Figure 2A, Table 1), when other studies investigating visual speech gesture perception have found it (Callan et al., 2003; Paulesu et al., 2003; Calvert & Campbell, 2003; Bernstein et al., 2002; Nishitani & Hari, 2002; Olson et al., 2002; Campbell et al., 2001). One possible

explanation is that there are multiple parallel pathways by which visual speech gesture information can be processed. Two possible pathways that may exist consist of one involving internal simulation in brain regions involved with planning and execution of speech production and one involving sites of MSI when both audio and visual speech gesture information is present. Under conditions in which perception is facilitated by MSI there may be less reliance on brain regions involved with internal simulation. It is interesting to point out that some studies reporting activity in brain regions involved with planning and execution of speech production present visual speech gesture information without a concordant audio speech signal (Calvert & Campbell, 2003; Bernstein et al., 2002; Nishitani & Hari, 2002; Olson et al., 2002; Campbell et al., 2001). However, brain regions involved with planning and execution of speech production are also activated by listening to auditory speech information, especially during tasks that require considerable phonetic processing (Zattore & Binder, 2000). It is possible that the degree to which speech gesture information available through visual or auditory channels is used to internally simulate the speech act is dependent on the demands of the perceptual task. Further research needs to be conducted to determine under what conditions brain regions involved with planning and execution of speech production are activated by perception of visual speech gesture information.

This study indicates sites of MSI in the MTG and STG/S, including the auditory cortex (Figure 2E, Table 1), thought to reflect processing of place of articulation information present in auditory and visual stimulation. By comparison of activity present for both the MF and UF conditions relative to the LF condition it is possible to control for the presence of activity resulting from attentional modulation cued by such things as visual information signaling the onsets and offsets of the acoustic speech signal, as well as MSI of aspects of gross visual motion with properties of the auditory speech signal. Additional research is needed to rule out attentional modulation specific to visual and/or auditory observation of speech-related biological motion to ensure that activity found in the mid to anterior STG/S and MTG is indeed a result of MSI. It is entirely possible that regions of the STG/S and the MTG involved with processing of speech related biological motion have some properties characteristic of MSI sites such as superadditivity of auditory and visual information.

METHODS

Subjects

Eleven right-handed native Japanese speakers (nine men and two women) participated in this study. Subjects were between 21 and 43 years of age (mean 27.4). All subjects volunteered to participate in the study and gave informed written consent for experimental procedures, approved by the ATR Human Subject Review Committee.

Stimuli

The stimuli consisted of 28 sentences spoken by a native Japanese speaker selected from the ATR Japanese Sentences database. The sentences are low in predictability. The duration of the sentences ranged from 2627 to 3916 msec with a mean of 3479 msec. The video signal of the sentences was degraded using one-octave band-pass wavelet spatial filters centered at 2.7 cpf, and 11 cpf using the procedure reported in Kroos et al. (2002). In this experiment, stimuli consisted of unfiltered sentences (UF), 11-cpf wavelet band-pass filtered sentences, and 2.7-cpf wavelet band-pass filtered sentences. The 2.7-cpf condition will be referred to as the LF condition and the 11cpf condition will be referred to as the MF condition. The sentences were recorded onto video laser disk for later stimulus presentation. A single frame of the same speaker's face in a neutral position was used for the baseline still face condition. Video was presented in black and white.

Audio noise used in the experiment consisted of a commercial English multispeaker babble track (Audiotec, St. Louis, MO) that was mixed with the speech signal. The signal-to-noise ratio (approximately −8 dB) was determined by pilot work and held constant for all conditions for all subjects.

Procedure

The fMRI procedure consisted of a block design in which seven sentences were presented (approximately 85–90 dB SPL) in each of the 16 blocks with a block duration of 35.37 sec. The total duration of sentences within a block was equated and the order of sentences within a block was randomized. The same sentences were used for all three experimental conditions and all sentences were used once in each condition. The presentation duration of the baseline still face condition was matched to be the same as the experimental conditions. Order of conditions was balanced by Williams square and was the same for each subject. The task was to passively identify as many phonemes as possible in the sentences presented. All subjects had practice with the various types of stimuli prior to fMRI scanning.

Video and audio (both speech and multispeaker babble noise) were presented from laser disk. Audio was presented via MR-compatible headphones (Hitachi Advanced Systems' ceramic transducer headphones). Video was presented by a projector located outside of the MR room to a mirror positioned inside of the head coil just above the subjects' eyes. Stimulus presentation

was controlled by specialized computer hardware–software that can drive the laser disk player.

fMRI Data Collection and Preprocessing

Functional brain imaging data were collected using the Shimadzu-Marconi Magnex Eclipse 1.5T PD250 at the ATR Brain Activity Imaging Center. Functional T2*-weighted images were acquired using a gradient echo-planar imaging sequence (echo time 55 msec; repetition time 3930 msec; flip angle 90°). A total of 37 contiguous axial slices covering the cortex and cerebellum were acquired with a 4 × 4 × 4-mm voxel resolution. Images were preprocessed using programs within SPM2b. Differences in acquisition time between slices were accounted for, movement artifact was removed, images were spatially normalized to a standard space using a template EPI image (3 × 3 × 3-mm voxels), and were smoothed using an 8 × 8 × 8-mm FWHM Gaussian kernel.

Acknowledgments

This research was conducted as part of 'Research on Human Communication' with funding from the Telecommunications Advancement Organization of Japan. We would like to thank the fMRI technicians Yasuhiro Shimada and Ichiro Fujimoto at the Brain Activity Imaging Center as well as Takaaki Kuratate for his help with stimulus construction.

Reprint requests should be sent to Daniel E. Callan, PhD, 2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan, or via e-mail: dcallan@atr.co.jp.

The data reported in this experiment have been deposited in the fMRI Data Center (http://www.fmridc.org). The accession number is 2-2004-115MP.

REFERENCES

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: Role of the STS region. *Trends in Cognitive Sciences, 4,* 267–278.

Banati, R. B., Goerres, G. W., Tjoa, C., Aggleton, J. P., & Grasby, P. (2000). The functional anatomy of visual–tactile integration in man: A study using positron emission tomography. *Neuropsychologia, 38,* 115–124.

Bernstein, L., Auer, E., Moore, J., Ponton, C., Don, M., & Singh, M. (2002). Visual speech perception without primary auditory cortex activation. *NeuroReport, 13,* 311–315.

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P., Springer, J. A., Kaufman, J. N., & Possing, E. T. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex, 10,* 512–520.

Bushara, K., Grafman, J., & Hallet, M. (2001). Neural correlates of auditory–visual stimulus onset asynchrony detection. *Journal of Neuroscience, 21,* 200–304.

Callan, D. E., Callan, A. M., Honda, K., & Masaki, S. (2000). Single-sweep EEG analysis of neural processes underlying perception and production of vowels. *Cognitive Brain Research, 10,* 173–176.

Callan, D. E., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2001). Multimodal contribution to speech perception revealed by independent component analysis: A single-sweep EEG case study. *Cognitive Brain Research, 10,* 349–353.

Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C., & Vatikiotis-Bateson (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *NeuroReport*, *14,* 2213–2218.

Calvert, G. A. (2001). Crossmodal processing in the human brain: Insights from functional neuroimaging studies. *Cerebral Cortex, 11,* 1110–1123.

Calvert, G. A., Brammer, M., Bullmore, E., Campbell, R., Iversen, S. D., & David, A. (1999). Response amplification in sensory-specific cortices during crossmodal binding. *NeuroReport, 10,* 2619–2623.

Calvert, G. A., Bullmore, E., Brammer, M. J., Campbell, R., Williams, S., McGuire, P., Voodruff, P., Iversen, S., & David, A. (1997). Activation of auditory cortex during silent lipreading. *Science, 276,* 593–596.

Calvert, G. A., & Campbell, R. (2003). Reading speech from still and moving faces: The neural substrates of visible speech. *Journal of Cognitive Neuroscience, 15,* 57–70.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology, 10,* 649–657.

Calvert, G. A., Hansen, P. C., Iversen, S. D., & Brammer, M. J. (2001). Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. *Neuroimage, 14,* 427–438.

Campbell, R., MacSweeney, M., Surguladze, S., Calvert, G. A., 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). *Cognitive Brain Research, 12,* 245–264.

Dodd, B. (1977). Lip reading in infants: Attention to speech presented in-and-out-of-synchrony. *Cognitive Psychology, 11,* 478–484.

Dronkers, N. F. (1996). A new brain region for coordinating speech articulation. *Nature, 384,* 159–161.

Ettinger, G., & Wilson, W. (1990). Cross-modal performance: Behavioural processes, phylogenetic considerations and neural mechanisms. *Behavioural Brain Research, 40,* 169–192.

Foxe, J. J., Wylie, G. R., Martinez, A., Schroeder, C. E., Javitt, D. C., Guilfoyle, D., Ritter, W., & Murray, M. M. (2002). Auditory–somatosensory multisensory processing in auditory association cortex: An fMRI study. *Journal of Neurophysiology, 88,* 540–543.

Genovese, C. R., Lazar, N. A., & Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage, 15,* 870–878.

Giard, M., & Peronnet, F. (1999). Auditory–visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. *Journal of Cognitive Neuroscience, 11,* 473–490.

Grant, K. W. (2001). The effect of speechreading on masked detection thresholds for filtered speech. *Journal of the Acoustical Society of America, 109,* 2272–2275.

Grant, K. W., & Braida, L. D. (1991). Evaluating the Articulation Index for audiovisual input. *Journal of the Acoustical Society of America, 89,* 2952–2960.

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America, 108,* 1197–1208.

Hadjikhani, N., & Roland, P. (1998). Cross-modal transfer of

information between the tactile and visual representations in the human brain: A positron emission tomography study. *Journal of Neuroscience, 18,* 1072–1084.
Hershenson, M. (1962). Reaction time as a measure in intersensory facilitation. *Journal of Experimental Psychology, 63,* 289–293.
Jancke, L., Mirzazade, S., & Shah, N. (1999). Attention modulates activity in the primary and the secondary auditory cortex: A functional magnetic resonance imaging study in humans. *Neuroscience Letters, 266,* 125–128.
Jones, J., & Callan, D. (2003). Brain activity during audiovisual speech perception: An fMRI study of the McGurk effect. *NeuroReport, 14,* 1129–1133.
Kent, R. D., & Tjaden, K. (1997). Brain functions underlying speech. In W. J. Hardcastle & J. Lavers (Eds.), *The handbook of the phonetic sciences* (pp. 220–255). London: Blackwell.
Kroos, C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). Video-based face motion measurement. *Journal of Phonetics, 30,* 569–590.
Lewis, J. W., Beauchamp, M. S., & DeYoe, E. A. (2000). A comparison of visual and auditory motion processing in human cerebral cortex. *Cerebral Cortex, 10,* 873–888.
Lotze, M., Seggewies, Erb, W., Grodd, W., & Birbaumer, N. (2000). The representation of articulation in the primary sensorimotor cortex. *NeuroReport, 11,* 2985–2989.
MacSweeney, M., Amaro, E., Calvert, G. A., Campbell, R., David, A., McGuire, P., Williams, S., Woll, B., & Brammer, M. (2000). Silent speechreading in the absence of scanner noise: An event-related fMRI study. *NeuroReport, 11,* 1729–1733.
MacSweeney, M., Campbell, R., Calvert, G., McGuire, P., Davie, A., Suckling, J., Andrew, C., Woll, B., & Brammer, M. J. (2001). Dispersed activation in the left temporal cortex for speech-reading in congenitally deaf people. *Proceedings of the Royal Society of London, Series B: Biological Sciences, 268,* 451–457.
McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746–748.
Mottonen, R., Krause, C., Tiippana, K., & Sams, M. (2002). Processing of changes in visual speech in the human auditory cortex. *Cognitive Brain Research, 13,* 417–425.
Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science., 15,* 133–137.
Munhall, K. G., Kross, C., & Vatikiotis-Bateson, E. (2002). Audiovisual perception of band-pass filtered faces. *Journal of the Acoustical Society of Japan, 21,* 519–520.
Munhall, K. G., & Vatikiotis-Bateson, E. (1998). The moving face during speech communication. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye, part 2: Advances in the psychology of speechreading and auditory-visual speech* (pp. 123–139). Sussex: Taylor & Francis–Psychology Press.
Nishitani, N., & Hari, R. (2002). Viewing lip forms: Cortical dynamics. *Neuron, 36,* 1211–1220.
Olson, I. R., Gatenby, J. G., & Gore, J. C. (2002). A comparison of bound and unbound audio-visual information processing in the human cerebral cortex. *Cognitive Brain Research, 14,* 129–138.
Paulesu, E., Perani, D., Blasi, V., Silani, G., Borghese, N. A., De Giovanni, U., Sensolo, S., & Fazio, F. (2003). A functional–anatomical model for lip-reading. *Journal of Neurophysiology, 90,* 2005–2013.
Rademacher, J., Morosan, P., Schormann, T., Schleicher, A., Werver, C., Freund, H., & Zilles, K. (2001). Probabilistic mapping and volume measurement of human primary auditory cortex. *Neuroimage, 13,* 669–683.
Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lipreading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97–113). Hillsdale, NJ: Erlbaum.
Rizzolatti, G., & Arbib, M. (1998). Language within our grasp. *Trends in Neurosciences, 21,* 188–194.
Sams, M., Aulanko, R., Hamalainen, M., Hari, R., Lounasmaa, O. V., Lu, S. T., & Simola, J. (1991). Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters, 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., & Tohkura, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *Journal of the Acoustical Society of America, 90,* 1797–1805.
Shams, L., Kamitani, Y., & Shimojo, S. (2002). Visual illusion induced by sound. *Cognitive Brain Research, 14,* 147–152.
Stein, B., Huneycutt, W., & Meredith, M. (1988). Neurons and behavior: The same rules of multisensory integration apply. *Brain Research, 448,* 355–358.
Stein, B., & Meredith, M. (1993). *The merging of the senses.* Cambridge: MIT Press.
Sumby, W., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26,* 212–215.
Talairach, J., & Tournoux, P. (1988). *Co-planar stereotactic atlas of the human brain.* New York: Thieme.
Vatikiotis-Bateson, E., Munhall, K. G., Hirayama, M., Lee, Y. C., & Terzopoulos, D. (1996). The dynamics of audiovisual behavior in speech. In D. Stork & M. Hennecke (Eds.), *Speechreading by humans and machines* (Vol. 150, pp. 221–232). Berlin: Springer-Verlag.
Wallace, M. T., Meredith, M. A., & Stein, B. E. (1992). Integration of multiple sensory modalities in cat cortex. *Experimental Brain Research, 91,* 484–488.
Wildgruber, D., Ackermann, H., & Grodd, W. (2001). Differential contributions of motor cortex, basal ganglia, and cerebellum to speech motor control: Effects of syllable repetition rate evaluated by fMRI. *Neuroimage, 13,* 101–109.
Wise, R. J. S., Scott, S. K., Blank, C., Mummery, C. J., Murphy, K., & Warburton, E. A. (2001). Separate neural subsystems within 'Wernicke's area.' *Brain, 124,* 83–95.
Yehia, H. C., Rubin, P. E., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication, 26,* 23–44.
Zatorre, R., & Binder, J. (2000). Functional and structural imaging of the human auditory system. In A. Toga & J. Mazziotta (Eds.), *Brain mapping the systems* (pp. 365–402). San Diego: Academic Press.