Interactions between Auditory and Visual Semantic Stimulus Classes: Evidence for Common Processing Networks for Speech and Body Actions

Georg F. Meyer[1], Mark Greenlee[2], and Sophie Wuerger[1]

Abstract

■ Incongruencies between auditory and visual signals negatively affect human performance and cause selective activation in neuroimaging studies; therefore, they are increasingly used to probe audiovisual integration mechanisms. An open question is whether the increased BOLD response reflects computational demands in integrating mismatching low-level signals or reflects simultaneous unimodal conceptual representations of the competing signals. To address this question, we explore the effect of semantic congruency within and across three signal categories (speech, body actions, and unfamiliar patterns) for signals with matched low-level statistics. In a localizer experiment, unimodal (auditory and visual) and bimodal stimuli were used to identify ROIs. All three semantic categories cause overlapping activation patterns. We find no evidence for areas that show greater BOLD response to bimodal stimuli than predicted by the sum of the two unimodal responses. Conjunction analysis of the unimodal responses in each category identifies a network including posterior temporal, inferior frontal, and premotor areas. Semantic congruency effects are measured in the main experiment. We find that incongruent combinations of two meaningful stimuli (speech and body actions) but not combinations of meaningful with meaningless stimuli lead to increased BOLD response in the posterior STS (pSTS) bilaterally, the left SMA, the inferior frontal gyrus, the inferior parietal lobule, and the anterior insula. These interactions are not seen in premotor areas. Our findings are consistent with the hypothesis that pSTS and frontal areas form a recognition network that combines sensory categorical representations (in pSTS) with action hypothesis generation in inferior frontal gyrus/premotor areas. We argue that the same neural networks process speech and body actions. ■

INTRODUCTION

The integration of sensory information from multiple modalities is a fundamental requirement for the recognition of human actions. Behavioral data show that temporally, spatially, and semantically congruent information has a facilitatory effect on performance such that bimodal stimuli are detected or discriminated faster or more accurately than incongruent bimodal stimuli (e.g., see Meyer, Wuerger, Röhrbein, & Zetzsche, 2005; Laurienti, Kraft, Maldjian, Burdette, & Wallace, 2004). The facilitatory effect of spatial and temporal congruence can be explained by early neural integration stages that have, for instance, been demonstrated in the superior colliculus of cat (e.g., Meredith & Stein, 1996; Meredith, Nemitz, & Stein, 1987), but early, signal-statistic-dependent integration of sensory signals (we will use the term *sensory* integration for this), which presumably draws on neural systems that do not yet provide a representation of the stimulus semantics, cannot account for the behavioral consequences of semantic congruency (semantic integration). Semantic effects have been demonstrated for biological motion perception (e.g., Brooks et al., 2007) as well as in speech perception (e.g., Soto-Faraco, Navarra, & Alsius, 2004) in EEG (e.g., Teder-Salejärvi, Di Russo, McDonald, & Hillyard, 2005) and fMRI studies (e.g., Werner & Noppeney, 2009; van Atteveldt, Formisano, Goebel, & Blomert, 2004) and are increasingly used to study audiovisual integration mechanisms (Szycik, Jansma, & Münte, 2009; Szycik, Tausche, & Münte, 2008).

Recent neuroimaging data support the view that sensory and semantic multimodal congruency selectively affects the BOLD response in different areas of the brain. Sadaghiani, Maier, and Noppeney (2009) showed a shift in the effect of natural (sensory), metaphoric, and linguistic (semantic) congruency for motion processing along the cortical processing hierarchy. The exact constituents of this hierarchy, however, are not yet well defined. Werner and Noppeney (2010a, 2010b) argue that superadditive multisensory effects in STS and intraparietal sulcus (IPS) are correlates of object categorization, whereas Doehrmann, Weigelt, Altmann, Kaiser, and Naumer (2010) describe repetition-related effects in object categorization at earlier stages of visual (lateral occipital cortex) and auditory (middle superior temporal gyrus [STG]) processing. Superadditive effects in the posterior STS (pSTS) are also the basis for the claim that this structure is an audiovisual binding site (e.g., Beauchamp, Argall, Bodurka, Duyn, &

[1]University of Liverpool, UK, [2]Universität Regensburg, Germany

Martin, 2004; Calvert, 2001), which may not require semantics at all.

One way to discriminate between semantic and sensory representations is to study the effect of congruency in audiovisual component signals that differ in semantics but have similar signal descriptors. Stimuli containing incongruent semantics should cause more metabolic activity in areas that process semantic representations but not at low-level, sensory, binding sites.

The purpose of the present fMRI study is to explore the effect of semantic congruency of audiovisual motion sequences within and across three signal categories: speech, body actions, and unfamiliar patterns. All signals were designed to have very similar underlying statistics so that they should be treated similarly by processes that use purely statistical characteristics for multisensory integration but come from two semantic categories for which strong theoretical claims for specialized processing are made (e.g., Liberman, 1996, "Speech is Special"; and Troje and Westhoff, 2006, special "life detectors" process biological motion). We would therefore expect significant differences between congruent and incongruent representations of the signals in regions where semantic categorization takes place and predict differential effects for signals that consist of two conflicting meaningful patterns (for instance, auditory speech paired with visual body action) compared with signals that consist of pairs of meaningful and meaningless unimodal patterns (e.g., visual speech presented with a scrambled auditory signal that is matched in low-level statistics).

The use of speech and body action signals also affords us the opportunity to directly compare activation patterns for these two stimulus classes within the same experiment and therefore address the question to what extent processing of the two "special signal" categories differs.

Specialized Representations for Speech and Biological Motion?

Speech and biological motion perception share much more than their respective claims to draw on special processing systems: Both signal types require the translation of hierarchically and temporally structured actions plans into kinematics to generate actions and, equally importantly, require these action plans to be recovered from multiple simultaneous and temporally overlapping ("coarticulated" in speech terminology) articulator movements to recognize the intended gestures. Both signal types require a close link between perceptual and motor systems for learning and self-monitoring.

The main motivation for this study, however, is that previous research in speech and biological motion perception (whether visual or auditory) identify the same brain areas: particularly the posterior superior temporal areas (pSTS, Brodmann's area [BA] 22) as well as Broca's area (the left inferior frontal gyrus [IFG], BA 44/45) and the premotor areas (BA 6), which will be reviewed in more detail in the following sections. A direct comparison of responses to the two signal types with matched stimuli and tasks within a single experiment not only addresses the question to what extent biological motion and speech perception are based on specialized processes but may also contribute to our understanding of the ontogeny and phylogeny of language by providing evidence that speech perception and action recognition share common processing substrates as a basis for development and evolution. Showing shared neural processing in the perception of body action and speech may help to develop and unify cognitive models in the two research domains.

Brain Areas Involved in Speech and Body Action Processing

In the following sections, we briefly review representative data to show that speech and body action processing predominantly draw on two linked brain areas: the posterior part of the STS bilaterally and a frontal cluster including the ventral premotor cortex and the IFG (Broca's area), where activation is dominant in the left hemisphere.

Posterior STS

The primary site for biological motion perception is the pSTS imaging data, which show selective responses to visual (e.g., Saygin, Wilson, Hagler, Bates, & Sereno, 2004; Puce & Perrett, 2003; Vaina, Solomon, Chowdhury, Sinha, & Belliveau, 2001; Grossman et al., 2000), auditory (Bidet-Caulet, Voisin, Bertrand, & Fonlup, 2005; Pizzamiglio et al., 2005), audiovisual (e.g., Stevenson, Kim, & James, 2009; Beauchamp, 2005), and imagined (Grossman & Blake, 2001) biological motion signals. The imaging data are supported by lesion studies (Vaina et al., 2001) and TMS data (Grossman, Batelli, & Pascual-Leone, 2005) and are consistent with anatomical studies in monkey (Seltzer & Pandya, 1989) and electrophysiological data (Barraclough, Xiao, Baker, Oram, & Perrett, 2005; Hikosaka, Iwai, Sato, & Tanaka, 1988).

Superior temporal areas, among them the pSTS, are also key speech areas and respond strongly to heard (for reviews, see, e.g., Hein & Knight, 2008; Hickok & Poeppel, 2007; Cabeza & Nyberg, 2000), seen (e.g., Skipper, Nusbaum, & Small, 2005), read (e.g., Cabeza & Nyberg, 2000), and subvocally reproduced (Hickok & Buchsbaum, 2003) speech. Most recent studies identify the STS as the site of prelexical speech categorization (for reviews, see Obleser & Eisner, 2009; Scott, McGettigan, & Eisner, 2009; Hickok & Poeppel, 2007). The pSTS demarcates the lower bank of Wernicke's area so that there is a wealth of literature linking speech perception deficits to lesions at this site (for a review, see, e.g., Damasio & Geschwind, 1984; but see Pulvermüller, 2005).

Although there is considerable evidence for the involvement of pSTS in audiovisual processing, there is debate about its precise role. One interpretation of the activation

of the pSTS in response to visual and auditory stimuli is that it acts as an audiovisual binding site. The basis for this hypothesis is the observation that the BOLD signal in response to bimodal stimuli is larger than the sum of the responses to the unimodal visual and auditory stimuli (e.g., Beauchamp, Lee, Argall, & Martin, 2004; Calvert, Campbell, & Brammer, 2000). In this interpretation, the increased BOLD response for bimodal compared with the summed unimodal response is caused by neural populations that selectively activate for bimodal signals or actively integrate visual and auditory signals. This selective activation could be seen for stimuli that are integrated on the basis of their signal properties without drawing on categorical representations.

An alternative hypothesis is that pSTS activation reflects supramodal processing on the basis of learned, conceptual representations that emerge from the integration of visual and auditory inputs. The basis for this argument is that the response to mismatching auditory and visual signals is typically larger than for congruent bimodal signals (e.g., Szycik et al., 2008, 2009; Skipper, Goldin-Meadow, Howard, Nusbaum, & Small, 2007; van Wassenhove, Grant, & Poeppel, 2005). Hocking and Price (2008) argue that when the task, attention, and stimuli are carefully controlled, the responses in a bimodal conceptual matching task are the same as seen in unimodal tasks, whereas the response to semantically incongruent bimodal stimuli is much higher than for congruent signals. A bimodal semantic matching task of course also places differential demands on audiovisual binding mechanisms. For semantically incongruent signals, increased metabolism in the pSTS is caused by the demands placed by conceptual matching or representations rather than multisensory signal fusion. This view is consistent with the suggestion that the pSTS may be a phonological buffer (Wise et al., 2001) and more recent proposals by Jacquemot and Scott (2006), who propose a role of the pSTS in short-term perceptual memory.

Premotor Cortex and Broca's Area

A second major complex that has been shown to be selectively activated by speech and biological motion signals includes the IFG and premotor areas.

Saygin et al. (2004) showed selective activation of human premotor cortex by visual point light stimuli. This finding is mirrored by action observation studies using natural video sequences (e.g., Pelphrey, Morris, Michelich, Allison, & McCarthy, 2005; Hamzei et al., 2003) that show differential activation in effector-specific sectors of Broca's area and premotor cortex (Buccino, Binkowski, & Riggio, 2004). Further evidence for the independent involvement of premotor areas as well as temporal sites in visual biological motion perception is provided by lesion data (Saygin, Driver, & de Sa, 2008).

A wealth of evidence links the IFG (Broca's area) and the premotor cortex to speech production (for a review, see Hickok & Poeppel, 2007), but both areas have also been shown to be active in the perception of spoken (e.g., Pulvermüller et al., 2006; Wilson, Saygin, Sereno, & Iacoboni, 2004; Watkins, Strafella, & Paus, 2003), audiovisual speech (Skipper, Goldin-Meadow, et al., 2007; Skipper et al., 2005), and environmental (nonspeech) sounds (Lewis et al., 2004 [fMRI]; Pizzamiglio et al., 2005 [EEG]). This is consistent with data showing that TMS over premotor cortex (Meister, Wilson, Deblieck, Wu, & Iacoboni, 2007) and IFG (Fadiga & Craighero, 2006) disrupts speech perception.

A possible neural substrate that explains activation for observed and executed speech and body action are "mirror neurons," which were first demonstrated in the ventral premotor cortex (area F5) of the macaque (Buccino et al., 2004; Gallese, Fadiga, Fogassi, & Rizzolatti, 1996). These neurons respond while the monkey performs goal-directed actions and, critically, also when the animal observes similar actions performed by others. Although the response is selective for specific actions, the modality of observation is not critical (Keysers et al., 2003; Kohler et al., 2002). This suggests that the mirror neuron system is a "supramodal" rather than a multisensory integration site.

The human homologue of macaque area F5 is Broca's area (e.g., Rizzolatti & Arbib, 1998; Petrides & Pandya, 1997); its implication in speech production and perception as well as action recognition has led to suggestions of shared circuitry for language and motor behavior in general (Meister & Iacoboni, 2007). This view is supported by experimental data that show overlapping activation maps (Aziz-Zadeh, Wilson, Rizzolatti, & Iacoboni, 2006; Pulvermüller, 2005; Grezes, Armony, Rowe, & Passingham, 2003) and common modulation of neural activation (e.g., Meister et al., 2003) in linguistic and action observation tasks. To our knowledge, there are currently no studies that directly compare the responses to speech and body action stimuli with matched stimulus and task complexity.

Given the striking similarity of the neural networks involved in speech and biological motion perception, it is interesting to determine the extent to which integrative mechanisms involved in the processing of speech and body action share a common neural basis. A direct comparison of activation patterns is one way of addressing the question. A second, complementary, approach is to present congruent and incongruent audiovisual signals that draw on both signal classes to study interactions between the stimulus types. Significant interactions would provide further evidence for colocalization of representations of speech and body actions.

Direct Comparison of Speech and Nonspeech Stimuli

Santi, Servos, Vatikiotis-Bateson, Kuratate, and Munhall (2003) directly compared visible speech and walking figures as point light stimuli and found speech-selective activation in a network of motor-related areas (Broca's area, premotor, primary motor, and SMA) for the speech stimuli. The task for the observers was to lip-read three word

sentences in the speech condition and to discriminate between jumping and walking for the body action stimuli. Although the speech-related activation of the motor circuit is consistent with other findings (e.g., Skipper et al., 2005) and models (e.g., Hickok & Poeppel, 2004), one might have expected to also see some activation in action observation networks for the body action stimuli (Saygin, 2007). Puce, Allison, Bentin, Gore, and McCarthy (1998) report overlapping bilateral activation of the pSTS for eye and mouth movements but not for checkerboard control patterns. A direct comparison of face and hand movements also identified selective activation in the same segments of the pSTS (Thompson et al., 2007), leading to the conclusion that the response of the pSTS is not body part specific. Although the mouth movements in the latter two studies were not speech movements, they suggest that visual speech and other actions may well also cause colocalized activation in the pSTS.

Meister and Iacoboni (2007) compared responses with hand–object interactions and related linguistic and perceptual tasks. The linguistic tasks activated a subset of the fronto-parietal network active during action perception.

Speech recognition and biological motion perception have been associated with claims of specialization to achieve closely related goals: the recovery of invariant, semantic representations from highly variable sensory input signals that represent human actions and are constrained by human motor dynamics. In both cases, the task is achieved in a largely modality-independent fashion, and independent imaging studies identify very similar processing networks for both tasks.

The experiments reported here have two complementary aims:

(1) To explore mechanisms of audiovisual semantic integration by comparing audiovisual (in)-congruency effects for two meaningful signals with those for meaningful and meaningless signals. We predict significantly stronger congruency effects between pairs of meaningful signals than pairs of meaningful and meaningless signals. We call this experiment our "main" experiment.
(2) To systematically compare the activation patterns of the three signal types to identify specialized or shared processing and to determine ROIs to perform our congruency tests. We call this experiment our "localizer" experiment, although the scope goes significantly beyond a simple ROI definition.

METHODS

Stimuli

We reviewed a number of experiments that contrast speech signals with nonspeech actions. A particular challenge for these comparisons is the selection of stimuli that are comparable in terms of their underlying signal statistics. Visual point light displays are ideally suited to create signals with matched underlying statistics, whereas auditory signals can also be conditioned to have matching long-term spectra and temporal structure (envelope).

A second consideration is that the operational task has to be comparable for the different stimulus types: An obvious consideration is task difficulty, perhaps less obvious, but equally important is the depth of analysis required to perform the task. Presentation of spoken sentences, for instance, will invoke syntactic and semantic representations that are not needed to recognize body actions such as walking or jumping (Santi et al., 2003). We argue that isolated speech syllables, such /aga/, are a good basis for comparison because they can be chosen not to invoke lexical or syntactic representations and have the same kinematic structure, a move from the vowel toward the consonantal target and back again, as simple body actions such as steps, jumps, and so forth.

With these considerations in mind, we used three categories of stimuli: body actions (BM), that is, whole-body visual actions and corresponding sounds, prelexical speech (SP), and scrambled (SCR) signals, which were novel but highly distinctive signals. Example videos are given in the Supplementary Material.

Visual Signals

All visual stimuli were point light displays that consisted of 13 white points on a black background. The body action stimuli were taken from the CRS biomotion examples and displayed using a ViSaGe system (Cambridge Research Systems Ltd., Kent, England); they represent distinctive actions, including a person walking, jumping, cycling, rowing, sawing, chopping wood, serving a tennis ball, playing pool, and so forth. We used one set of nine different signals for the localizer experiment and another nine examples for the main experiment.

The visual speech signals were generated by extracting the location of 13 markers on the speaker's lips that were recorded while producing prelexical vowel–consonant–vowel syllables, consisting of the three vowels /a, i, u/ and three consonants /b, d, g/. The same vowel was always used at the start and at the end of the syllable. Nine different combinations of the three vowels with the three consonants were used. A female speaker was used for the localizer experiment, whereas a male speaker was used for the main experiment.

Scrambled visual signals were created by taking a random selection of six points from one body action signal and seven points from one speech signal. The start points of the resulting 13 points were randomized inside a kernel corresponding to the size of the speech and body action stimuli, but each of the points followed its own local motion pattern during the stimulus presentations. A point in the top left quadrant of the scrambled pattern might therefore follow the local motion pattern of the right elbow of a tennis player, whereas an adjacent point might follow the trajectory of the left corner of a mouth saying /ubu/. The

resulting sequences are meaningless but highly distinctive motion patterns (see Figure 1 and Supplementary Material for example videos).

The speech and the body action stimuli were scaled to match the average local motion signal amplitude (the average distance travelled by all points over all examples) in the two data sets. All sequences were sampled at 30 Hz and lasted for 1.3 sec.

Auditory Signals

The auditory body action signals consisted of recordings of real sounds reflecting the visual action. The sound signals were carefully synchronized to ensure that impacts in the visual signals coincided with the onset of the auditory impact sounds. Auditory and visual speech signals were recorded at the same time.

To produce distinctive but novel auditory scrambled sounds, we rotated the spectrum of the speech sounds so that low- and high-frequency components are mirrored between 0 and 4000 Hz (Blesser, 1969). The resulting inverted spectra were filtered to match the long-term spectrum of speech sounds, and finally, signals were amplitude modulated to follow the envelope of the body action component. This procedure resulted in a set of novel but distinctive auditory signals that covaried with the visual scrambled signals (Figure 1).

Stimulus Presentation

Visual stimuli (13 white points on a dark background) were back projected on a screen inside the MR scanner with a ViSaGe system (Cambridge Research Systems Ltd.) driving a D-ILA LCD projector (JVC Corp., Japan) with a frame refresh rate of 60 Hz. The screen size subtended 16.4° × 21.7° of visual angle; the diameter of the visual stimulus was approximately 10°.

The auditory signals were presented via a TDT RM1 (Tucker-Davies Technologies, Alachua, FL) signal processor and an MR Confon MR-compatible headphones (MR Confon GmbH, Magdeburg, Germany). We calibrated all sounds to have the same root mean square level and measured the absolute signal levels in the scanner using an optical microphone (MR Confon GmbH), which we could calibrate outside the scanner. Our participants wore earplugs in addition to the MR Confon sound attenuating headphones. Assuming an average attenuation of 30 dB because of the earplugs, we estimate absolute signals levels of 68 db(A) while the EPI sequence resulted in sound levels of 63 dB(A); the background noise level measured during "silent" periods in the sparse imaging run was approximately 50 dB(A). The signal-to-noise ratio in the localizer experiment was 6 dB, whereas it was 18 dB in the main sparse imaging experiment. On the basis of previous experiments that used white noise as a masker, we expected the stimuli to be highly intelligible in both conditions (e.g., Meyer & Morse, 2003).

Participants

Eleven participants (9 women; mean age = 22.5 years, age range = 19–28 years) with normal or corrected-to-normal vision and no history of hearing difficulty or any other neurological disorders were recruited from the student population at Regensburg University. Participants received direct payment or course credit for their participation and gave informed consent in accordance with the guidelines of the Regensburg University Ethics Committee. The same subjects took part in two separate recording sessions for the localizer and main experiment.

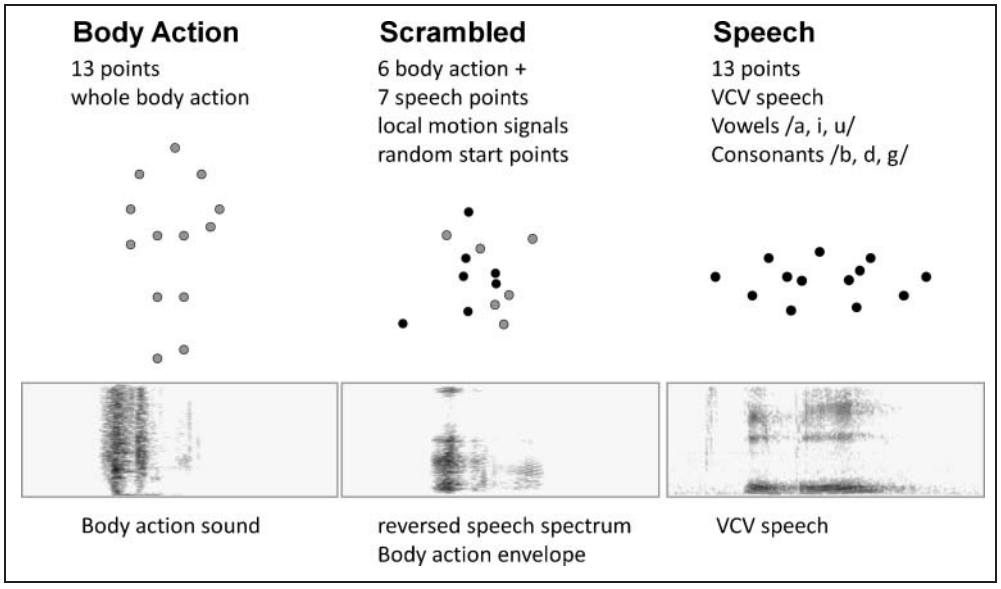

Figure 1. Schematic representation of the three stimulus types used. All stimuli lasted for 1.3 sec (39 frames at 30 Hz). Visual stimuli consisted of 13 moving points, whereas auditory stimuli were matching recordings. Body action stimuli represented different actions such as walking, jumping, sweeping, and so forth; speech stimuli were meaningless vowel–consonant–vowel syllables. Scrambled stimuli were generated by combining scrambled features of one speech and one body action example to produce novel but distinctive stimuli.

Behavioral Task

In both experiments, we used a one-back task where participants had to press a button when they perceived the same action twice in direct succession. In the localizer experiment, the actions were defined in the auditory modality, in the visual modality, or in both; in the main experiment, actions were always defined in both modalities. The behavioral task was chosen to ensure that subjects attend to and categorize the stimuli independent of the modality of presentation.

MRI Acquisition

All functional images were acquired with a 3-T head scanner (Siemens Allegra) with a birdcage head coil at the MR Imaging Centre of the University of Regensburg. Their heads were additionally secured by soft foam pads.

In the localizer experiment, BOLD responses were measured using a T2*-weighted EPI sequence (echo time = 30 msec, volume repetition time = 2.0 s, resolution = 3 × 3 mm, number of slices = 34 interleaved, slice thickness = 3 mm, distance factor = 15%, flip angle = 90°). The localizer session consisted of three experimental runs (audiovisual, auditory, and visual, in random order) using a block design (16 sec stimulation with seven randomly chosen stimuli, followed by 16 sec of rest). Button presses were modeled as separate events.

At the end of the localizer session, structural images of the whole brain were acquired using a T1-weighted MPRAGE sequence (repetition time = 2250 msec, echo time = 2.6 msec, resolution = 1 × 1 × 1 mm^3).

The main experiment used the same MRI scanner parameters except that a sparse sampling paradigm and an event-related design was used (Bunzeck, Wuestenberg, Lutz, Heinze, & Jäncke, 2005; Zaehle, Wüstenberg, Meyer, & Jäncke, 2004; Gaab, Gaser, Zaehle, Jancke, & Schlaug, 2003; Hall et al., 1999). We used a repetition time of 4 sec so that a 2-sec (noisy) acquisition period alternated with 2 sec or relative silence, during which period the stimuli were presented. The precise stimulus onset time within the 2-sec window was randomly jittered to enhance the estimate of the hemodynamic response. Responses consequently were sampled on average 2.65, 6.65, and 10.65 sec after stimulus onset. The experimental design included 10% "null events" where the no stimulus, other than the fixation point, was presented. All stimulus types were presented in a pseudorandom sequence.

Data analysis was conducted using SPM5 (Wellcome Department of Imaging Neuroscience, London, UK, http://www.fil.ion.ucl.ac.uk/spm/). Functional images of each participant were corrected for residual head motion, realigned to the first image, and corrected for slice timing. Subsequently, all functional images were coregistered and normalized to the Montreal Neurological Institute 152 template and resampled to a 2 × 2 × 2-mm^3 spatial resolution. Spatial smoothing was applied to the functional images using an isotropic Gaussian kernel with an FWHM of 8 mm. A general linear model was constructed for each participant to analyze the hemodynamic responses captured by the functional images.

For the localizer experiment, which used a block design, general linear model regressors were generated by convolving the canonical hemodynamic function with a boxcar function representing a particular section of the experiment. A high-pass filter (1/128 Hz) was applied to remove low-frequency drifts. The t test contrasts between the stimulus conditions were calculated individually and averaged across participants using random-effect analysis.

The main experiment was an event-related design so that regressors were generated by convolving unit impulses with the canonical hemodynamic function and also with the temporal derivative of this function (e.g., Henson, Rugg, & Friston, 2001). Null events and responses were treated as separate events (e.g., Friston, Holmes, Price, Buchel, & Worsley, 1999).

The centers of suprathreshold activation regions were localized using the SPM anatomy toolbox (Eickhoff et al., 2005, 2007; Eickhoff, Heim, Zilles, & Amunts, 2006). Where BAs were not provided by the anatomy toolbox, Caret (Van Essen et al., 2001) was used. Reported p values pertain to the cluster.

RESULTS

Behavioral Performance

In the localizer experiment, audiovisual, visual, and auditory stimuli were presented in blocks. Subjects were asked to press a button when they perceived the same signal twice in direct succession. In total, 274 repeated signals were presented of which 238 (86.9%) were correctly identified (see Table 1).

A repeated measures ANOVA for the two factors Modality (visual, auditory, and audiovisual) and Action (body action, speech, and scrambled) showed that there were no significant main effects of Modality, $F(2, 98) = 0.82$, $p = .46$, or Action, $F(2, 98) = 1.99$, $p = .16$. There was also no significant interaction between the two factors, $F(4, 98) = 2.087$, $p = .088$.

It is important to note that the performance on the scrambled data is very similar to that for the meaningful signals. The low score for visual speech is expected because the point light display does not provide the information necessary to distinguish alveolar stops (/d/) from velar stops (/g/), whereas bilabial stops (/b/) can easily be visually identified.

In the main experiment, subjects were asked to press a button when they perceived the same action twice in direct succession and were told that in this experiment the action could be presented in different modalities: The subject might hear a tennis serve in one trial and then see it in the next.

The average correct identification scores are displayed in terms of the content of the second signal (Table 2). Along

Table 1. Mean Behavioral Performance Measures for the Signals Used in the localizer Task

	Body Action			*Speech*			*Scrambled*		
	A	*V*	*AV*	*A*	*V*	*AV*	*A*	*V*	*AV*
% Correct	83.3	91.0	95.5	100	67.3	93.8	74.8	87.5	88.9
SD	16.85	32.3	10.8	0	42.5	13.5	24.1	30.34	26.15
Mean RT (sec)	1.25	1.21	1.07	1.25	1.11	1.13	1.06	1.53	1.17

The table shows the percentage of correctly recognized actions and standard deviation and the mean RT in seconds. RTs were computed only for correctly identified signals.

the diagonal are signals that have matching visual and auditory components (congruent trials), whereas other entries show signals that consist of signal pairs that are drawn from two action categories. For these categories, it is possible that a semantic target in one modality precedes the same category in the other modality (e.g., subjects first hear steps, then see a walker). To account for the effect of cross-modality matching, we present the statistics twice, once with all cross-modality matches included (within and across in Table 2) and once excluding across modality matches.

Subjects recognize between 55.7% (scrambled) and 78.4% (body action) of repeated signals. This level of performance is lower than that in the localizer experiment, which did not include incongruent trials (Table 2B).

A repeated measures ANOVA of the recognition rates with the two factors Auditory (BM, SP, and SCR) and Visual (BM, SP, and SCR) showed no significant main effects of the auditory stimulus identity, $F(2, 98) = 3.63, p = .08$, visual, $F(2, 98) = 1.23, p = .30$, or interaction effects, $F(4, 98) = 0.25$ $p = .91$, when cross-modal trials are discounted.

When cross-modal trials are included in the analysis (Table 2A), an ANOVA shows significant main effects for the auditory stimulus identity, $F(2, 98) = 4.46$, $p = .01$, and significant interactions, $F(4, 98) = 2.78$, $p = .03$, but no main effect of visual category, $F(2, 98) = 2.25$, $p = .11$.

When one or two components in an incongruent signal have been previously presented, recognition performance is lower except where auditory speech and visual body action are presented (70.8%). Performance for visual speech and auditory body action was much lower (38.9%), which may be expected from the unimodal data that showed that the auditory modality was much better recognized for speech whereas visual information was more salient for body action.

Table 2. Behavioral Performance Measures for the Signals Used in the Main Experiment

	Visual		
	BM	*SP*	*SCR*
A. Within and Across, % (SD)			
Audio			
BM	78.40 (40.35)	38.90 (26.85)	46.41 (27.59)
SP	70.80 (34.54)	67.42 (38.81)	38.05 (19.48)
SCR	29.09 (34.68)	34.68 (36.84)	55.67 (47.22)
B. Within Only, % (SD)			
Audio			
BM	78.40 (40.35)	54.55 (36.58)	62.92 (37.86)
SP	81.21 (30.23)	67.42 (38.81)	81.32 (31.45)
SCR	51.73 (43.47)	46.58 (35.11)	55.67 (47.22)

The table shows the percentage of recognized repeats (and standard deviation). All signals in the main experiment had an audio (rows) and visual (columns) component. The two sets of data on the left (A) show that average data for all signal presentations, whereas the data on the right (B) exclude trials where a signal was presented in one modality and then repeated in the other modality.

Imaging Results

Our study consisted of two parts, a localizer scan that was designed to identify candidate ROIs for the incongruency effects (main) experiment. The localizer experiment also affords us the opportunity to compare the activations with the three semantic audiovisual signal categories with matched underlying statistics and within the same experimental conditions and observers. Both experiments will be discussed in turn.

The localizer scan was designed to identify areas that show selective activation to bimodal compared with unimodal stimuli (audiovisual [AV] > auditory [A] + visual [V]) or to identify areas that are significantly activated by visual and auditory stimuli (a conjunction analysis: (A > Rest) ∩ (V > Rest)). This analysis was carried out for each of the three stimulus types, and although we found no instances of AV-selective activation, we identified a network of brain areas that respond to visual and auditory presentation of all three signal classes.

Experiment 1 (Localizer): Common Activation Patterns for Body Actions, Speech, and Scrambled Signals

In the localizer experiment, we observe broadly similar activation patterns for our three stimulus types for each of the modes of presentation and found that for all three

stimulus types, activation patterns for the bimodal stimuli were well described by the union of the two unimodal activation patterns. We therefore report the responses to bimodal stimuli here for brevity. Figure 2A shows extensive activation in bilateral occipital, posterior temporal, superior parietal, and posterior frontal areas for the three audiovisual stimuli when compared with rest.

Rather than discussing the detailed activation patterns for all stimuli and conditions, we describe those brain areas that respond to all three audiovisual stimulus conditions (Figure 2B) and discuss systematic differences in activation that are seen between stimulus categories (Figure 3). Activation patterns for unimodal stimuli and SPM analysis tables can be found in the Supplementary Materials.

We performed a conjunction analysis (Friston, Penny, & Glaser, 2005) to identify areas that were activated for all three stimulus types ($AV_{sp} \cap AV_{bm} \cap AV_{scr}$; Figure 2B). Areas with significant activation ($p < .05$, family-wise error [FWE] corrected) include visual and auditory sensory areas (extrastriate visual areas, middle occipital gyrus, and BA 18, 19, bilaterally; auditory cortical areas in STG and BA 22, 41) but also the superior parietal lobule (BA 7) bilaterally and premotor cortical sites (precentral and middle frontal gyrus, BA 6).

Stimulus-specific Differences in Activation Patterns

More interesting than commonalities in activation patterns are systematic differences between the three stimulus categories. We compare activation patterns for the audiovisual presentation of the three stimulus types (Figure 3; details are given in Table 2 in the Supplementary Material).

When speech is contrasted with body action and scrambled signals, significant increases in BOLD response are seen in the anterior temporal areas bilaterally. Speech causes selective activation ($p < .05$, FWE corrected) in superior and middle temporal (MT) gyrus (BA 21) bilaterally and at the right temporal pole (BA 38). Speech stimuli also cause increased activity at the left temporal pole compared with body action stimuli.

Scrambled signals cause significantly more activity than body action signals in the left superior medial gyrus, middle occipital gyrus bilaterally, and right insula and supramarginal gyrus. We could not identify any areas where body action causes significantly greater responses than scrambled signals at the conservative threshold of $p < .05$ (FWE corrected) used in all other contrasts.

Although speech signals cause significantly increased activation in anterior temporal areas, they simultaneously cause relatively less activity than both other signal types in MT and occipital areas that are associated with visual processing. Scrambled signals also activate superior parietal areas (SPL, BA 7) more than speech, whereas body action causes more activity than speech in the left MT gyrus (BA 22, 37, 39) and in the fusiform gyrus bilaterally.

No Evidence for Superadditive Responses to Bimodal Stimuli

Within each stimulus type (speech, body action, and scrambled), we looked for selective audiovisual activation

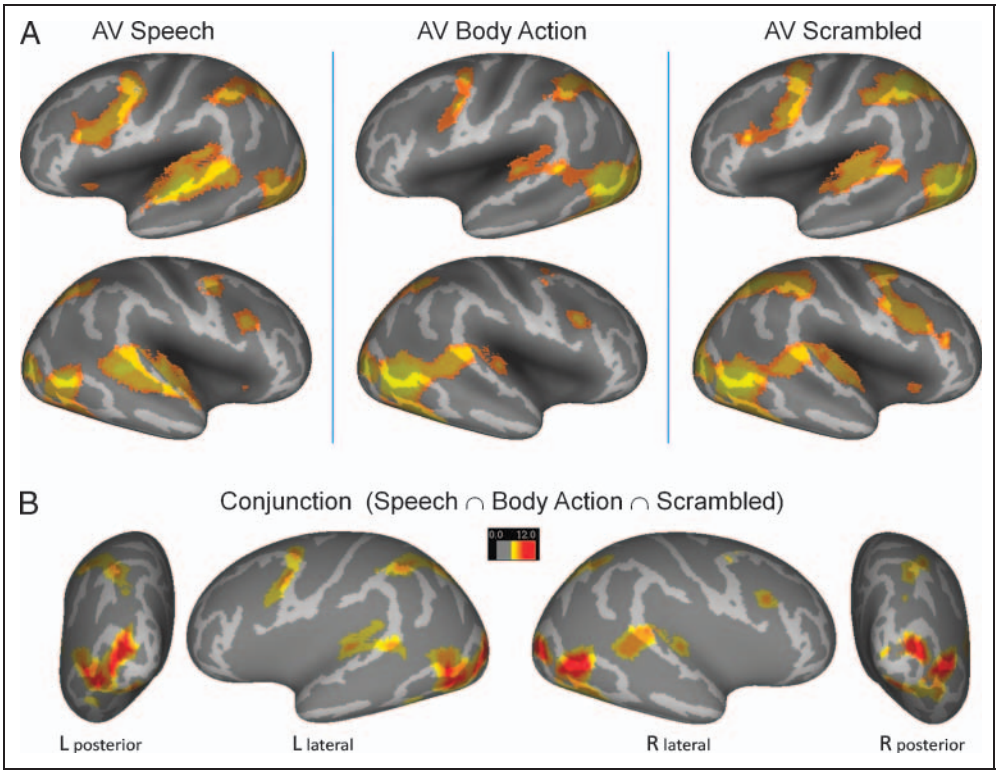

Figure 2. Activation clusters in response to the audiovisual localizer stimuli with $p < .05$ (FWE corrected), mapped onto "inflated brains" (A, top). All images are scaled to the same t value range. The images show that all three stimulus classes evoke activity in many shared brain areas. The bottom panel shows a conjunction analysis to highlight areas that are activated by all three audiovisual semantic stimulus classes. Areas involved are visual areas (LOC and MT+ in BA 19, BA 37, and BA 18), auditory areas in BA 42 and BA 22, superior parietal lobule (BA 7) bilaterally, and premotor cortical sites (precentral and middle frontal gyrus, BA 6).

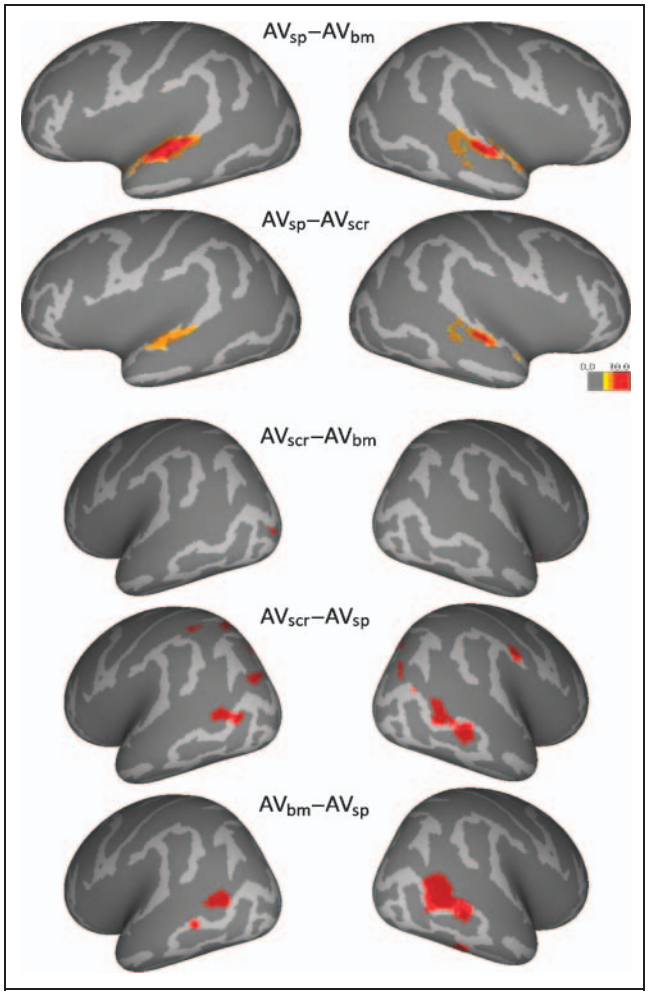

Figure 3. Differences ($p < .05$, FWE corrected) between the three types of signals used in this study. The AV_{bm}–AV_{scr} did not result in any suprathreshold clusters.

by computing the AV > (A + V) contrast, but we were unable to identify any areas where this contrast showed significant ($p < .05$, FWE corrected) activation after areas that showed deactivation in the unimodal conditions were discounted. This finding is consistent with the results of Hocking and Price (2008) but differs from earlier reports. This may be due to not only the more careful control of stimulus, task, and attentional factors and the relatively larger voxel size used in the analysis (cf. Beauchamp, 2005; Beauchamp, Argall, et al., 2004) but also the relatively conservative significance threshold we applied.

In the following section, we demonstrate the result of a conjunction analysis with a conservative threshold that identifies all the major areas previously identified in speech and body action processing as potential ROIs. This, in our view, vindicates our conservative threshold choice.[1]

Isolating Supramodal Representations

Looking for areas where bimodal stimulation exceeds the sum of the unimodal responses is a useful approach to identify areas that specifically represent or process bimodal rather than unimodal information. An alternative approach to select potential ROIs that code bimodal (or supramodal) representations is to identify those areas that respond to both visual and auditory stimuli. This approach roughly follows a suggestion by Szycik et al. (2008, 2009) for identifying areas responding to audiovisual speech; this approach identifies areas that are activated by either modality as supramodal.

We performed a separate conjunction analysis (Friston et al., 2005) for each of the three signal types to identify areas in the brain that respond to visual and auditory stimuli. There is an extensive overlap in the responses for the three signal types, so they are color coded in Figure 4 (the labels in the figure refer to the corresponding numbers in the next paragraph and in Table 3).

We see an extensive activation ($p < .001$, uncorrected) for the conjunction of unimodal visual and auditory signals in the left and right pSTS (BA 22, ❶ ❷ in Figure 4 and Table 3), the bilateral premotor areas (BA 6, extending ventrally into BA 44, ❸ ❹), the left SMA (BA 6, ❺), the left IFG (BA 44/45, ❻), and an area at the junction of the left temporal and parietal cortex (BA 40, ❼; referred to as area SPT by Hickok and Poeppel, 2007). We also find activity in the left dorsal inferior parietal lobule (IPL)/IPS (BA 40, ❽) and left anterior insula (❾). It is striking that the areas we identify in the conjunction analysis not only represent a "textbook" example of the areas involved in speech perception (e.g., the dorsal stream in Hickok & Poeppel, 2007) but also represent the extended mirror neuron network proposed by Pineda (2008).

Although some areas identified in the conjunction analysis specifically respond to one stimulus class, such as Broca's areas (IFG ❻ and area SPT ❼) that show significant activation in response to speech but not the other stimuli, other areas show significant activation for all stimulus types. This is most evident not only in the pSTS bilaterally but also in the SMA and anterior insula.

The conjunction analysis is a group analysis so that inter-individual anatomical variability may cause a blurring of observed activation patterns in standard space (e.g., Szycik et al., 2009) and consequently an apparent overlap of activation clusters in the group analysis that is not representative of individual data. We analyzed the imaging data on an individual basis and found significant overlap in the activation patterns for the three stimulus types in the bilateral pSTS regions in 10 of our 11 subjects (see Supplementary Material for the scans).

Activation in the right STS region was found to be much more extensive for all three stimulus types than that in the left hemisphere. This is borne out by the lateralization index (LI; Wilke & Lidzba, 2007). We computed for the STS region (LI = −0.74 for speech, LI = −0.79 for body action, and LI = −0.69 for scrambled signals).[2]

A very different picture emerges for the premotor area (❸ ❹). Here we see distinct activation patterns for the three stimulus types, with some overlap between body

action and scrambled and strong lateralization differences. Activation for speech stimuli is strongly left lateralized (LI = +0.889), whereas activation for body action (LI = −0.40) and scrambled (LI = −0.66) is predominantly right lateralized. Finding bilateral STS and left-lateralized frontal activity is consistent with previous speech data (for a review, see Hickok & Poeppel, 2007), whereas a right-lateralized response for body action and scrambled stimuli is consistent with recently reported data (Wuerger et al., under review).

Main Experiment: Semantic Congruency Effects in Regions Identified to Respond to Visual and Auditory Signal Presentation

The localizer experiment used a block design to identify areas of auditory–visual coactivation (ROIs). In a second recording session that used sparse sampling and an event-related design, we studied the effect of audiovisual congruency between the three stimulus categories within the ROIs. In this experiment, all stimuli were bimodal but did not have to be congruent: We used a factorial design where we presented all combinations of the three semantic stimulus categories visually and auditorily. All analysis was carried out in only the ROIs defined by the localizer (Figure 4, Table 3) using MarsBaR (Brett, Anton, Valabregue, & Poline, 2002).

We hypothesize that the presentation of incongruent stimuli containing speech and body action signals should lead to significant increases of BOLD response in areas that process these two semantic categories in comparison with congruent stimuli. Incongruent stimuli consisting of one meaningful and one scrambled signal should have less effect in areas where semantic representations are processed or maintained.

We computed the differences between congruent and incongruent stimulus pairs of the three stimulus types by looking for significant activation differences in the following contrasts:

(1) $(A_{bm}V_{bm} + A_{sp}V_{sp})$ versus $(A_{bm}V_{sp} + A_{sp}V_{bm})$
(2) $(A_{bm}V_{bm} + A_{scr}V_{scr})$ versus $(A_{bm}V_{scr} + A_{scr}V_{bm})$
(3) $(A_{sp}V_{sp} + A_{scr}V_{scr})$ versus $(A_{sp}V_{scr} + A_{scr}V_{sp})$

The left-hand term in the equation consists of congruent stimulus pairs; for Contrast 1, this would be all congruent AV body action $(A_{bm}V_{bm})$ and congruent AV speech $(A_{sp}V_{sp})$ trials. The second half contains incongruent pairs of the same auditory and visual component stimuli. $A_{bm}V_{sp}$, for example, represents stimuli that had an auditory body action and a visual speech component. If auditory and visual information is processed independently, then the BOLD responses for congruent and incongruent signal pairs should be equal because the stimuli on both sides of the equation are identical. Superadditive responses where bimodal stimulation is larger than the sum of unimodal stimulation (e.g., Calvert, 2001), perhaps because of specific bimodal representations or processes in the brain, should result in significantly greater responses for the congruent stimuli. Szycik et al. (2008) observe increased activity in the pSTS bilaterally when incongruent speech signals are compared with congruent signals and attribute this difference to audiovisual integration processes. Rather than presenting congruent and incongruent versions of the same stimulus type, we present congruent and incongruent stimulus categories.

In our data set, the only contrast that shows significant activation differences was Contrast 1, that is, the difference between incongruent and congruent speech and body action signals. We could not identify any significant effect of incongruency when contrasting biomotion with

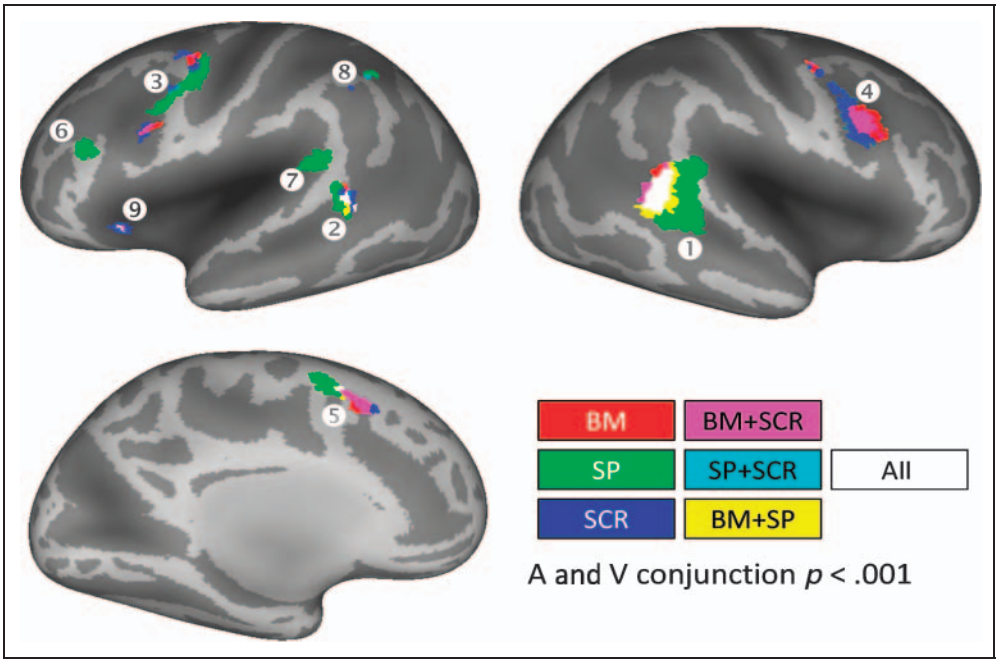

Figure 4. Summary of the localizer results. The areas identified by the conjunction analysis for the three stimulus types are shown on inflated brains. Areas that show significant ($p < .001$, uncorrected) activation for body action (BM, red), speech (SP, green), or scrambled (SCR, blue) or combination of two or three stimulus types are shown. The numbers refer to appropriate columns in Table 3 and the main text.

Table 3. Areas Activated a Conjunction of All Three Audiovisual Stimulus Types (Left) and *t*/*p* Values of the ROI Analysis within These Regions for the Three Semantic Stimulus Category Pairings

		Localizer Experiment				*Main Experiment Incongruent–Congruent ROI Analysis*		
Location		*Significance*	*Position*	*Voxels*	*T*	*SP/BM*	*SP/SCR*	*BM/SCR*
pSTS R (BA 22)	❶	SP	48 −40 4	714	6.69	**$t = 2.00, p = .02$**	$t = 0.52, p = .30$	$t = 0.00, p = .50$
		BM	56 −44 8	244	4.64	**$t = 2.13, p = .02$**	$t = 0.66, p = .25$	$t = 0.17, p = .43$
			(50 −40 14)		4.59			
		SCR	48 −44 8	158	5.02	**$t = 2.98, p < .01$**	$t = 0.24, p = .40$	$t = 0.07, p = .47$
pSTS L (BA 22)	❷	SP	−56 −44 8	107	4.52	**$t = 1.82, p = .04$**	$t = 0.90, p = .18$	$t = 0.23, p = .41$
			(−48 −42 8)		4.21			
		BM	−48 −44 10	28	3.67	$t = 1.36, p = .07$	$t = 0.67, p = .25$	$t = 0.14, p = .44$
		SCR	−50 −46 8	28	3.56	$t = 1.40, p = .08$	$t = 0.57, p = .28$	$t = 0.04, p = .49$
Precentral gyrus L (BA 6)	❸	SP	−48 −6 52	222	4.46	$t = 1.40, p = .08$	$t = 0.48, p = .32$	$t = 0.30, p = .38$
			(−54 −6 44)		3.90			
			(−44 2 30)		3.69			
		BM	−46 −6 50	36	3.59	$t = 1.11, p = .11$	$t = 0.64, p = .26$	$t = 0.27, p = .39$
			−38 −2 34	26	3.52	$t = 1.24, p = .09$	$t = 0.31, p = .38$	$t = 0.17, p = .43$
		SCR	−40 −2 36	49	3.63	$t = 1.41, p = .08$	$t = 0.49, p = .31$	$t = 0.35, p = .36$
			−38 −4 50	16	3.43	$t = 1.57, p = .06$	$t = 1.04, p = .15$	$t = 0.54, p = .30$
IFG pars opercularis L (BA 6/44)		BM	−50 8 24	12	3.42	$t = 1.58, p = .06$	$t = 0.41, p = .34$	$t = 0.89, p = .19$
		SCR	−50 8 24	21	3.52	$t = 1.52, p = .06$	$t = 0.57, p = .29$	$t = 0.04, p = .49$
Precentral gyrus R BA (6)	❹	SP	50 0 52	13	3.63	$t = .69, p = .27$	$t = 0.63, p = .27$	$t = 0.01, p = .50$
		BM	48 0 52	16	3.60	$t = 0.51, p = .27$	$t = 0.49, p = .31$	$t = -0.05, p = .52$
		SCR	46 6 24	343	4.92	$t = 1.24, p = .11$	$t = -0.22, p = .59$	$t = -0.24, p = .60$
			(48 6 38)		3.47	$t = 0.51, p = .30$	$t = 0.69, p = .24$	$t = -0.34, p = .63$
			48 2 52	22	3.91			
IFG/precentral gyrus R (BA 44/6)		BM	42 6 26	156	4.23	$t = 0.98, p = .16$	$t = -0.19, p = .57$	$t = -0.03, p = .51$
		SCR	44 4 28	67	3.99	$t = 1.24, p = .11$	$t = -0.29, p = .62$	$t = -0.39, p = .65$
			(48 4 38)		3.31			
SMA L (BA 6)	❺	SP	−4 2 60	92	4.38	$t = 1.45, p = .07$	$t = 0.31, p = .38$	$t = 0.42, p = .33$
			−6 12 46	4	3.30	**$t = 2.11, p = .02$**	$t = 0.93, p = .07$	$t = 1.04, p = .15$
		BM	−8 14 48	71	3.93	**$t = 2.55, p < .01$**	$t = 0.94, p = .18$	$t = 1.27, p = .10$
			(−6 8 56)		3.62			
		SCR	−6 12 50	113	4.26	**$t = 2.55, p < .01$**	$t = 0.77, p = .22$	$t = 1.03, p = .15$
IFG L (BA 44/45)	❻	SP	−38 24 22	92	4.11	**$t = 2.02, p = .02$**	$t = 0.69, p = .25$	$t = 0.36, p = .36$
		BM	−42 22 22	13	3.47	**$t = 2.88, p = .02$**	$t = 0.62, p = .27$	$t = 0.61, p = .27$
IPL L area SPT (BA 40)	❼	SP	−56 −42 22	78	4.29	**$t = 2.05, p = .02$**	$t = 0.73, p = .23$	$t = 0.06, p = .48$
IPL/IPS hIP1 L (BA 40)	❽	SP	−34 −48 40	54	3.69	$t = 0.94, p = .18$	$t = 1.05, p = .15$	$t = 0.09, p = .47$
		SCR	−44 −42 38	89	4.11	$t = 0.95, p = .17$	$t = 0.71, p = .25$	$t = 0.21, p = .42$
Anterior insula L	❾	SP	−28 24 2	4	3.28	**$t = 2.73, p < .01$**	$t = 1.34, p = .09$	**$t = 2.18, p = .02$**
		BM	−26 22 6	4	3.25	**$t = 2.17, p = .02$**	$t = 1.08, p = .14$	**$t = 1.83, p = .03$**
		SCR	−28 24 0	40	3.88	**$t = 2.73, p < .01$**	$t = 1.05, p = .18$	$t = 0.36, p = .37$

We find significant incongruency effects when speech and body actions are paired, but not when one of the two meaningful stimuli is paired with a meaningless, scrambled, signal.

scrambled or when contrasting speech with scrambled signals, except for two very small ROIs in the left anterior insula (Table 3).

An ROI analysis using MarsBaR (Brett et al., 2002) identified significant ($p < .05$, uncorrected) incongruency effects for the speech/body action pairing in the following areas: the pSTS bilaterally, the left SMA, the left IFG, the IPL, and the left anterior insula (boldface entries in Table 3, the exact p values are shown). We do not find significant incongruency effects in the ROIs located in the premotor cortex at ROIs identified bilaterally at the junctions between BA 6 and BA 44 or in the left IPL.

The same analysis for pairings between meaningful and scrambled stimuli (Table 3) shows that the selective incongruency effects for the two meaningful stimulus categories are not a thresholding artifact. The p values in pSTS (bilaterally), left SMA, IFG, and IPL are in almost all cases a magnitude higher than for the speech/body action comparison.

DISCUSSION

Posterior STS

The pSTS bilaterally is strongly activated by all stimulus types in the localizer experiment, although the extent of activation is larger on the right than on the left for all three stimulus types. We found considerable overlap in the ROIs we identify. The ROIs defined by speech stimuli are more anterior than the ROIs defined by body action and scrambled signals.

The right pSTS shows significant interactions between speech and body action in the ROIs defined by all stimulus types, whereas the interaction in the left pSTS is borderline significant for anterior, speech-defined ROI and just insignificant for the more posterior ROIs.

Szycik et al. (2008) attribute the increased activation in the pSTS for incongruent stimuli to audiovisual integration processes, whereas Hocking and Price (2008) argue for a role in conceptual representation. Our results also show increased activity, but only for incongruent stimuli that consist of two different meaningful stimulus categories, not if a meaningful stimulus is paired with a scrambled signal. If the primary function of the pSTS was to reconcile and to match signals from different categories without recourse to categorical or conceptual representations, one might have expected similar behavior for all stimulus categories. If, on the other hand, the pSTS represents or derives conceptual representations, then interactions between familiar, but not between familiar and novel, stimuli would be expected. Our findings support the view that the pSTS is primarily involved in conceptual knowledge representation rather than as an integration stage that reconciles stimuli in the visual and auditory modality before conceptual representations emerge.

Premotor Areas

We found extensive activation in the premotor areas, bilaterally, for all three stimulus categories compared with the rest condition. A direct comparison of the audiovisual conditions in the localizer experiment did not show significant differences between these activation patterns at the relatively conservative threshold level of $p < .05$ (FWE corrected). The conjunction analysis used to identify ROIs defined by activation to auditory-alone and visual-alone signals, however, showed a left-lateralized response to speech and a more right-lateralized response to body action and scrambled signals.

We found no evidence for interactions between incongruent stimulus components within premotor cortex in the main experiment (Table 3). The rationale for the conjunction analysis was that isolating areas that respond to visual-alone and auditory-alone signals should select bimodal or supramodal representations. If the ROIs defined by this approach represent dissociable and lateralized supramodal representations of the motor actions for speech and body action, then no interactions would be expected.

Schubotz and colleagues considered pattern prediction in very similar visual (Schubotz & von Cramon, 2001) and auditory (Schubotz, von Cramon, & Lohmann, 2003) tasks. Subjects had to localize (where), identify (what), or temporally match abstract geometric visual and artificial auditory stimuli to be able to detect three repeats in a sequence. They identified different foci of activity for three stimulus property conditions. The "what" condition, which is closely related to our task, caused peak activity in the superior ventrolateral premotor areas bilaterally (corresponding to the anterioventral sections of area ❸ and ❹ in Figure 4 and Table 3. We also found extensive activation slightly superior to these areas. They also identified the left pre-SMA (❺, Figure 4 and Table 3) and areas in the left STG as specific to object identification–related tasks.

SMA Involvement

The areas identified by the localizer scans largely match those areas reported in previous studies that identified either temporal (pSTS) areas or action recognition networks (premotor, IFG, and IPL) as involved in multimodal speech and body action processing. We found strong activation for all stimulus types and interactions between incongruent and congruent body action–speech signal pairs in the SMA. Hall, Fussell, and Summerfield (2005) also identified an area in the medial superior frontal gyrus (BA 6) with similar Montreal Neurological Institute coordinates to the area we identified in an audiovisual speech perception task. They argue that this area may be involved in motor planning because their subjects were required to respond manually to auditory and visual speech stimuli. In our experiment, motor planning is not a plausible explanation for activation in this area because motor responses were rare and explicitly modeled. Bidet-Caulet et al. (2005) also report selective SMA activation in an auditory action recognition task and attribute the activation to preparation of the motor response.

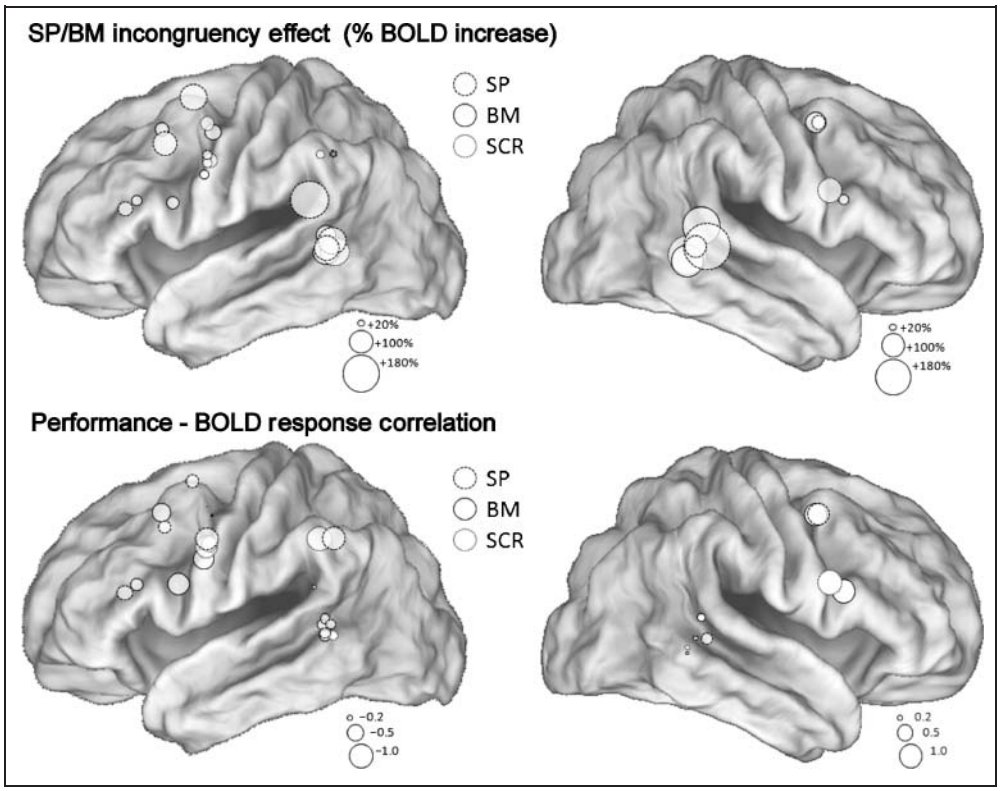

Figure 5. The top row of shows relative BOLD increases in the incongruent AV condition compared with the congruent condition, whereas the bottom row shows correlations between BOLD response and percentage correct responses as a measure of task difficulty. The center of each circle identifies the center of the corresponding cluster; dotted edges mark clusters identified by speech (SP), lines body action (BM), and dotted lines scrambled (SCR) stimuli. The circle size represents the relative BOLD response increase when comparing incongruent versus congruent speech and body action conditions (top) and show the correlation between BOLD response and behavioral performance. We see the largest (and significant) BOLD response increases in bilateral pSTS, left IFG, and left SMA but much smaller increases in premotor and parietal regions. The correlation of BOLD response with performance shows an almost exact opposite picture: Areas that show little incongruency effect (bilateral premotor and left parietal areas) show BOLD responses that are highly (inversely) correlated with performance.

The SMA is associated not only with a range of motor control tasks, such as task sequencing or movement initiation, but also with various other functions such as sensory processing, listening comprehension, speech expression, and working memory (for a review, see Chung, Man, Jeong, & Jack, 2005). Chung et al. (2005) assessed SMA activation for five tasks: finger movement, heat sensation, word generation, listening comprehension, and a two-back digit recognition (working memory) task. They found selective activation for word generation and working memory tasks in the anterior sections, whereas the posterior areas of the SMA responded to motor and sensory tasks. The position of the anterior regions matched the location of the activation clusters we found. The involvement of the SMA in speech generation is also evident in lesion studies where transient speech disorders after left SMA resection are common (Krainik et al., 2003).

Working Memory

The one-back task used in our experiments exercises working memory, and one way to interpret the finding of shared neural circuits is that these shared areas represent parts of a multicomponent working memory buffers (e.g., Baddeley, 2000). Baddeley (2000) proposed three working memory buffers under the control of a central executive system. A visuospatial sketch responsible for visual or spatial tasks, a phonological loop coding word sequences, and an episodic memory buffer that interfaces between the other two buffers and long-term memory. We see activation in the pSTS, which has been proposed as a phonological or a sensory memory buffer (Jacquemot & Scott, 2006; Wise et al., 2001), and Broca's area, which is an obvious candidate area for the generative part of the rehearsal system. Rehearsal might also explain the shared activity and interactions seen when congruent and incongruent speech and body action are presented in the main experiment. This could be achieved either by subvocal rehearsal of the category labels or rehearsal of the action sequences that define the body action and speech stimuli. It is, however, difficult to reconcile the activation of a phonological buffer or subvocal rehearsal for the scrambled stimuli. These stimuli were novel, so that a subvocal rehearsal of category labels seems unlikely and also could not be rehearsed by generating appropriate action sequences as could the speech and body action signals, yet the behavioral performance and activation pattern in the localizer experiment showed broadly similar patterns to those seen in the two meaningful recognition tasks.

A somewhat different view is taken by Postle (2006), who argues that working memory is an emergent property that is caused by sustained activity in those brain areas that are responsible for the representation of information in non–working memory tasks rather than a dedicated neural

system; for our data, this might mean that the pSTS is a brain area responsible for the representation of sensory information that turns into a working memory buffer by virtue of sustained activation, perhaps closely linked to sustained activation in areas responsible for motor pattern generation (BA 44 and BA 6). He proposes a second important principle: We opportunistically and automatically recruit as many mental representations as are afforded by the information that is retained. This is an attractive proposition for the audiovisual signals we employ in our study: The three signal types contain linguistic and nonlinguistic and auditory and visual components and therefore should activate phonological and visuospatial representations to varying degrees. We do not see complementary activation patterns for audiovisual speech, body action, and nonsense patterns as might be predicted but rather subtle differences in common activation patterns for all three tasks.

Postle's (2006) view strengthens the argument that shared underlying representations are used for the processing of speech and body action. There is a substantial body of evidence (reviewed in the Introduction) that the pSTS, posterior IFG, and premotor areas respond preferentially to speech and body actions in tasks that do not explicitly require working memory, which lends support to the view that the activation patterns we see do not represent working memory per se but may well be enhanced by the need to briefly retain information in the one-back task.

Task Difficulty

The interaction analysis compares incongruent with congruent trials, and it is reasonable to argue that the incongruent task is more difficult and may therefore explain the higher BOLD response in incongruent conditions. There are two reasons to doubt this explanation:

The first argument is that our behavioral data show that incongruency between auditory and visual components does not per se affect performance; there is little difference between the congruent speech and the body action conditions (78.4% BM, 67.4% SP; Table 2A) and the incongruent condition where audio speech and visual body action are shown (70.8%). For visual speech and auditory body action, performance is much worse (38.9%). This suggests that rather than incongruence, it is the relative information content in the two modalities that affects performance: Speech is difficult to lip-read because articulators in the oral cavity are not visible, whereas body actions are hard to identify from audio cues because most sounds associated with actions are impact sounds that are highly variable (steps on different materials) or only indirectly linked to the action (a tennis serve produces the same highly distinctive sounds as a backhand).

The second argument is based on a direct comparison of the relative BOLD response increase for the incongruent condition with the correlation between BOLD response increase and performance. The top row of Figure 5 shows the percentage increase in BOLD response for incongruent stimulation relative to congruent stimulation for all areas identified by the localizer. The largest relative BOLD increases are seen in the pSTS bilaterally, left IFG, SMA, and IPL, whereas premotor areas (bilaterally) show the smallest increases. An analysis of the correlation between BOLD response and behavioral performance as a measure of task difficulty shows exactly the opposite pattern: Sites in the premotor cortex bilaterally and left IPL show the most (negative) correlation with behavioral performance, whereas the areas identified as showing the largest incongruency effects show least correlation with performance. We conclude that incongruency effects are not directly caused by differences in task difficulty.

Differences in Activation in pSTS and Premotor Areas

Skipper, van Wassenhove, Howard, Nusbaum, and Small (2007) studied how cortical structures linked to speech production mediate audiovisual speech perception by presenting subjects with consistent and inconsistent (McGurk & MacDonald, 1976) audiovisual speech stimuli. They propose an active perception model where the observed auditory and visual data (represented in pSTS) are interactively compared with predicted auditory and somatosensory outcomes of motor commands that are generated in ventral premotor cortex to achieve motor goals specified in pars opercularis from a specification in pSTS. The model fits an interpretation of the pSTS as a phonetic buffer (Wise et al., 2001) and is compatible with the assumption that (mirror) neurons in premotor cortex are involved in speech perception (e.g., Iacoboni, 2008) via a generative process.

We see significant overlap in the spatial activation patterns and large interactions between speech and body action in the pSTS but relatively little overlap in activation patterns and interactions in the premotor cortex. These activation patterns are not directly correlated with performance or (indirectly) task difficulty. Activation in the premotor cortices, however, shows little interaction between semantic categories but much higher correlation with performance.

A model that assumes that the pSTS acts as a memory buffer would explain the increased activation in incongruent trials where two separate semantic categories have to be represented and maintained. We see low correlations between performance and pSTS activation; this could be because one of the incongruent stimuli (visual body action and auditory speech) is a relatively "easy" stimulus, whereas the complementary pair (visual speech and auditory body action) is hard to recognize. Both stimuli would put roughly equal demands on memory buffer.

Treating the premotor area as a generative recognition system that generates motor representations of speech and body actions to aid in the recognition would predict the correlation between performance and BOLD response in this area because more difficult stimuli, presumably, would require more potential action hypotheses to be generated before a match can be found. One explanation for

not finding interactions between speech and body action in premotor cortex might be that speech and body actions are represented predominantly in different hemispheres, as our localizer results suggest. Another explanation might lie in the operation of generative pattern-matching systems: It is relatively straightforward to implement passive, template-based, pattern-matching systems, such as those proposed for the pSTS in parallel as a bank of signal detectors; a fundamental constraint in active hypothesis generation systems, however, is that a series of hypotheses is generated sequentially until a match is found. Because all candidate hypotheses are generated using the same circuitry, they have to be implemented serially rather than in parallel. We would therefore expect to see significant interactions in the controlling structures that have to arbitrate between the speech and the body actions that need to be synthesized (the IFG according to Skipper, van Wasenhove, et al., 2007), but not in the ventral premotor cortex where action representations are synthesized sequentially.

Conclusions

Our task required the identification and conceptual matching of audiovisual signals. Two of them were based on very familiar (speech and body actions) signals, one class represented highly structured and distinctive novel stimuli (scrambled). A first, very striking, finding is that for our carefully matched stimuli, we found very similar activation patterns for all three stimulus classes, which is consistent with previous studies that identify substantially the same areas as being involved in the perception of speech and biological motion signals.

We presented unimodal and bimodal stimuli in the localizer task but did not observe areas where responses to bimodal stimuli were significantly larger than the summed responses for unimodal stimuli. This is consistent with the findings of Hocking and Price (2008). A conjunction analysis was used to select areas that respond to visual and auditory signals as candidate ROIs and identified a network of brain areas including the pSTS bilaterally, the premotor cortex, bilaterally extending into BA 44, and a left dominant network including the anterior insula, the Broca's area (BA 45/44), the SMA (BA 6), and the IPL (BA 40). These areas are not only an integral part of the dual stream model of speech perception (Hickok & Poeppel, 2007) but also represent the putative extended mirror neuron network (Pineda, 2008). In this context, it is perhaps not surprising that activity in most areas is not stimulus specific.

Within the ROIs, we compared responses for conceptually congruent and incongruent stimuli and found significant interactions between the two meaningful stimuli (speech and body actions) where incongruent stimulation resulted in higher BOLD responses than congruent signals. This supports the view advanced by Hocking and Price (2008), who argue that pSTS is part of a distributed set of regions involved in conceptual matching for speech signals. We extend Hocking and Price's findings by showing that pSTS responds to body action and scrambled stimuli as well as to speech signals and suggest that pSTS is involved in conceptual matching of sensory information beyond speech or biological motion perception.

Saygin (2007), on the basis of lesion studies, argues that the frontal and temporal areas have separate roles in biological motion perception. For speech stimuli, an incomplete split into articulatory functions in frontal areas and a perceptual role for the temporal areas are well supported by the literature (e.g., Hickok & Poeppel, 2007). Our imaging results echo Saygin's findings, not only in showing involvement in premotor areas in biological action recognition (Saygin et al., 2004) but also in showing differences in lateralization, sensitivity to semantic incongruencies, and task difficulty between pSTS and premotor areas, which supports the view that both regions have specific roles.

Jacquemot and Scott (2006) propose that perceptual and articulatory STM systems are very closely linked and that phonological working memory arises from the cycling of information between an articulatory buffer in left inferior frontal areas (Broca's area) and the left posterior temporal areas (Wernicke's area). We show that both areas respond strongly to speech and body action and therefore propose to extend the model from the speech domain into more general action observation.

We find significant spatial overlap between speech and body action–evoked activity as well as interactions between semantically incongruent signal components within a subset of these areas, which is inconsistent with claims that attribute specialized processing circuits for either speech or biological motion stimuli; rather, it suggests that both signal types draw on common neural substrates to solve a common problem, which is to extract underlying invariant categorical representations from complex, overlapping, and highly variable actions. Both tasks appear to draw on action representations in the pFC and on sensory representations in the pSTS to achieve this goal.

Acknowledgments

This work was supported by a Wellcome Trust grant (082831/Z/07/Z). The authors thank Dr. Roland Rutschmann for technical support and two anonymous reviewers for their helpful comments.

Reprint requests should be sent to Georg F. Meyer, School of Psychology, Liverpool University, Eleanor Rathbone Building, Liverpool, United Kingdom, L69 7ZA, or via e-mail: georg@liv.ac.uk.

Notes

1. Activation in response to bimodal stimulation in the most anterior part of the left STS (BA 22) and the most posterior and superior portion of the STS (BA39), bilaterally, was significantly larger than the sum of the two unimodal responses, but this was due to a deactivation, relative to the resting condition, for the two unimodal stimuli. In this case, the contrast is not meaningful.
2. $LI = \frac{\sum A_L - \sum A_R}{\sum A_L + \sum A_R}$ where A is the number of suprathreshold voxels in the left (L) and right (R) hemisphere.

REFERENCES

Aziz-Zadeh, L., Wilson, S. M., Rizzolatti, G., & Iacoboni, M. (2006). Congruent embodied representations of visually presented actions and linguistic phrases describing actions. *Current Biology, 16,* 1818–1823.

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory. *Trends in Cognitive Sciences, 4,* 417–423.

Barraclough, N. E., Xiao, D., Baker, C. I., Oram, M. W., & Perrett, D. I. (2005). Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. *Journal of Cognitive Neuroscience, 17,* 377–391.

Beauchamp, M. S. (2005). See me, hear me, touch me: Multisensorial integration in lateral occipital temporal cortex. *Current Opinion in Neurobiology, 15,* 145–153.

Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H., & Martin, A. (2004). Unraveling multisensory integration: Patchy organization within human STS multisensory cortex. *Nature Neuroscience, 7,* 1190–1192.

Beauchamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron, 41,* 809–823.

Bidet-Caulet, A., Voisin, J., Bertrand, O., & Fonlup, P. (2005). Listening to a walking human activates the temporal biological motion area. *Neuroimage, 28,* 132–139.

Blesser, B. A. (1969). *Perception of spectrally rotated speech*. Ph.D. thesis, Department of Electrical Engineering, Massachusetts Institute of Technology.

Brett, M., Anton, J.-L., Valabregue, R., & Poline, J. P. (2002). *Region of interest analysis using an SPM toolbox* [Abstract]. Presented at the 8th International Conference on Functional Mapping of the Human Brain, June 2–6, 2002, Sendai, Japan. Available on CD-ROM in *Neuroimage, 16*(2).

Brooks, A., van der Zwan, R., Billard, A., Petreska, B., Clarke, S., & Blanke, O. (2007). Auditory motion affects visual biological motion processing. *Neuropsychologia, 45,* 523–530.

Buccino, G., Binkowski, F., & Riggio, L. (2004). The mirror neuron system and action recognition. *Brain and Language, 89,* 370–376.

Bunzeck, N., Wuestenberg, T., Lutz, K., Heinze, H., & Jäncke, J. (2005). Scanning silence: Mental imagery of complex sounds. *Neuroimage, 26,* 1119–1127.

Cabeza, R., & Nyberg, L. (2000). Imaging cognition: II. An empirical review of 275 PET and fMRI studies. *Journal of Cognitive Neuroscience, 12,* 1–47.

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

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.

Chung, G. H., Man, Y. M., Jeong, S. H., & Jack, C. R. (2005). Functional heterogeneity of the supplementary motor area. *AJNR, American Journal of Neuroradiology, 26,* 1819–1823.

Damasio, A. R., & Geschwind, N. (1984). The neural basis of language. *Annual Reviews of Neuroscience, 7,* 127–147.

Doehrmann, O., Weigelt, S., Altmann, C. F., Kaiser, J., & Naumer, M. J. (2010). Audiovisual functional magnetic resonance imaging adaptation reveals multisensory integration effects in object-related sensory cortices. *Journal of Neuroscience, 30,* 3370–3379.

Eickhoff, S., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., et al. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *Neuroimage, 25,* 1325–1335.

Eickhoff, S. B., Heim, S., Zilles, K., & Amunts, K. (2006). Testing anatomically specified hypotheses in functional imaging using cytoarchitectonic maps. *Neuroimage, 32,* 570–582.

Eickhoff, S. B., Paus, T., Caspers, S., Grosbras, M. H., Evans, A., Zilles, K., et al. (2007). Assignment of functional activations to probabilistic cytoarchitectonic areas revisited. *Neuroimage, 36,* 511–521.

Fadiga, L., & Craighero, L. (2006). Hand actions and speech representation in Broca's area. *Cortex, 42,* 486–490.

Friston, K. J., Holmes, A. P., Price, C. J., Buchel, C., & Worsley, K. J. (1999). Multisubject fMRI studies and conjunction analyses. *Neuroimage, 10,* 385–396.

Friston, K. J., Penny, W. D., & Glaser, D. E. (2005). Conjunction revisited. *Neuroimage, 25,* 661–667.

Gaab, N., Gaser, C., Zaehle, T., Jancke, L., & Schlaug, G. (2003). Functional anatomy of pitch memory—An fMRI study with sparse temporal sampling. *Neuroimage, 19,* 1417–1426.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain, 119,* 593–609.

Grezes, J., Armony, J. L., Rowe, J., & Passingham, R. E. (2003). Activations related to "mirror" and "canonical" neurones in the human brain: An fMRI study. *Neuroimage, 18,* 928–937.

Grossman, E. D., Batelli, L., & Pascual-Leone, A. (2005). Repetitive TMS over posterior STS disrupts perception of biological motion. *Vision Research, 45,* 2847–2853.

Grossman, E. D., & Blake, R. (2001). Brain activity evoked by inverted and imagined biological motion. *Vision Research, 41,* 1475–1482.

Grossman, E. D., Donnelly, M., Price, P., Morgan, V., Pickens, D., Neighbor, G., et al. (2000). Brain areas involved in the perception of biological motion. *Journal of Cognitive Neuroscience, 12,* 711–720.

Hall, D. A., Fussell, C., & Summerfield, A. Q. (2005). Reading fluent speech from talking faces: Typical brain networks and individual differences. *Journal of Cognitive Neuroscience, 17,* 939–953.

Hall, D. A., Haggard, M. P., Akeroyd, M. A., Palmer, A. R., Summerfield, A. Q., Elliott, M. R., et al. (1999). "Sparse" temporal sampling in auditory fMRI. *Human Brain Mapping, 7,* 213–223.

Hamzei, F., Rijntjes, M., Dettmers, C., Glauche, V., Weiller, C., & Büchel, C. (2003). The human action recognition system and its relationship to Broca's area: An fMRI study. *Neuroimage, 19,* 637–644.

Hein, G., & Knight, R. T. (2008). Superior temporal sulcus—It's my area: Or is it? *Journal of Cognitive Neuroscience, 20,* 2125–2136.

Henson, R. N. A., Rugg, M. D., & Friston, K. J. (2001). The choice of basis functions in event-related fMRI. *Neuroimage, 13,* 149.

Hickok, G., & Buchsbaum, B. (2003). Temporal lobe speech perception systems are part of the verbal working memory circuit: Evidence from two recent fMRI studies. *Behavioral and Brain Sciences, 26,* 740–741.

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition, 92,* 67–99.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8,* 393–402.

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. *Journal of Neurophysiology, 60,* 1615–1637.

Hocking, J., & Price, C. J. (2008). The role of the posterior superior temporal sulcus in audiovisual processing. *Cerebral Cortex, 18,* 2439–2449.

Iacoboni, M. (2008). The role of premotor cortex in speech perception: Evidence from fMRI and rTMS. *Journal of Physiology, Paris, 102,* 31–34.
Jacquemot, C., & Scott, S. K. (2006). What is the relationship between phonological short-term memory and speech processing? *Trends in Cognitive Sciences, 10,* 480–486.
Keysers, C., Kohler, E., Umiltà, M. A., Nanetti, L., Fogassi, L., & Gallese, V. (2003). Audiovisual mirror neurons and action recognition. *Experimental Brain Research, 153,* 628–636.
Kohler, E., Keysers, C., Umiltà, M. A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: Action representation in mirror neurons. *Science, 297,* 846–848.
Krainik, A., Lehéricy, S., Duffau, H., Capelle, L., Chainay, H., Cornu, P., et al. (2003). Postoperative speech disorder after medial frontal surgery: Role of the supplementary motor area. *Neurology, 60,* 587–594.
Laurienti, P. J., Kraft, R. A., Maldjian, J. A., Burdette, J. H., & Wallace, M. T. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. *Experimental Brain Research, 158,* 405–414.
Lewis, J., Wightman, F., Brefczynski, J., Phinney, R., Binder, J., & DeYoe, E. (2004). Human brain regions involved in recognizing environmental sounds. *Cerebral Cortex, 14,* 1008–1021.
Liberman, A. M. (1996). *Speech: A special code. Learning, development, and conceptual change series*. Cambridge, MA: MIT Press.
McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746–748.
Meister, I. G., Boroojerdi, B., Foltys, H., Sparing, R., Huber, W., & Topper, R. (2003). Motor cortex hand area and speech: Implications for the development of language. *Neuropsychologia, 41,* 401–406.
Meister, I. G., & Iacoboni, M. (2007). No language-specific activation during linguistic processing of observed actions. *PLoS One, 9,* e891.
Meister, I. G., Wilson, S. M., Deblieck, C., Wu, A. D., & Iacoboni, M. (2007). The essential role of the premotor cortex in speech perception. *Current Biology, 17,* 1692–1696.
Meredith, M. A., Nemitz, J. W., & Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. *Journal of Neuroscience, 7,* 3215–3229.
Meredith, M. A., & Stein, B. E. (1996). Spatial determinants of multisensory integration in cat superior colliculus neurons. *Journal of Neurophysiology, 75,* 1843–1857.
Meyer, G., & Morse, R. (2003). The intelligibility of consonants in noisy vowel–consonant–vowel sequences when vowels are selectively enhanced. *Speech Communication, 41,* 429–440.
Meyer, G. F., Wuerger, S. M., Röhrbein, F., & Zetzsche, C. (2005). Low-level integration of auditory and visual motion signals requires spatial co-localisation. *Experimental Brain Research, 166,* 538–547.
Obleser, J., & Eisner, F. (2009). Pre-lexical abstraction of speech in the auditory cortex. *Trends in Cognitive Sciences, 13,* 14–19.
Pelphrey, K. A., Morris, J. P., Michelich, J. R., Allison, T., & McCarthy, G. (2005). Functional anatomy of biological motion perception in posterior temporal cortex: An fMRI study of eye, mouth and hand movements. *Cerebral Cortex, 15,* 1866–1876.
Petrides, M., & Pandya, D. N. (1997). Comparative architectonic analysis of the human and the macaque frontal cortex. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology* (Vol. IX, pp. 17–58). New York: Elsevier.
Pineda, J. A. (2008). Sensorimotor cortex as a critical component of an "extended" mirror neuron system: Does it solve the development, correspondence, and control problems in mirroring? *Behavioral and Brain Functions, 4,* 47.
Pizzamiglio, L., Aprile, T., Spitoni, G., Pitzalis, S., Bates, E., D'Amico, S., et al. (2005). Separate neural systems for processing action- or non-action-related sounds. *Neuroimage, 24,* 852–861.
Postle, B. R. (2006). Working memory as an emergent property of the mind and brain. *Neuroscience, 139,* 23–38.
Puce, A., Allison, T., Bentin, S., Gore, J. C., & McCarthy, G. (1998). Temporal cortex activation in humans viewing eye and mouth movements. *Journal of Neuroscience, 18,* 2188–2199.
Puce, A., & Perrett, D. (2003). Electrophysiology and brain imaging of biological motion. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences, 358,* 435–445.
Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience, 6,* 576–582.
Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences, U.S.A., 103,* 7865–7870.
Rizzolatti, G., & Arbib, M. A. (1998). Language within our grasp. *Trends in Neurosciences, 21,* 188–194.
Sadaghiani, S., Maier, J. X., & Noppeney, U. (2009). Natural, metaphoric and linguistic auditory direction signals have distinct influences on visual motion processing. *Journal of Neuroscience, 29,* 6490–6499.
Santi, A., Servos, P., Vatikiotis-Bateson, E., Kuratate, T., & Munhall, K. (2003). Perceiving biological motion: Dissociating visible speech from walking. *Journal of Cognitive Neuroscience, 15,* 800–809.
Saygin, A. P. (2007). Superior temporal and premotor brain areas necessary for biological motion perception. *Brain, 130,* 2452–2461.
Saygin, A. P., Driver, J., & de Sa, V. R. (2008). In the footsteps of biological motion and multisensory perception: Judgements of audiovisual temporal relations are enhanced for upright walkers. *Psychological Science, 19,* 469–475.
Saygin, A. P., Wilson, S. M., Hagler, D. J., Bates, E., & Sereno, M. I. (2004). Point-light biological motion perception activates human premotor cortex. *Journal of Neuroscience, 24,* 6181–6188.
Schubotz, R. I., & von Cramon, D. Y. (2001). Functional organization of the lateral premotor cortex: fMRI reveals different regions activated by anticipation of object properties, location and speed. *Cognitive Brain Research, 11,* 97–112.
Schubotz, R. I., von Cramon, D. Y., & Lohmann, G. (2003). Auditory what, where, and when: A sensory somatotopy in lateral premotor cortex. *Neuroimage, 20,* 173–185.
Scott, S. K., McGettigan, C., & Eisner, F. (2009). A little more conversation, a little less action—Candidate roles for the motor cortex in speech perception. *Nature Reviews Neuroscience, 10,* 295–302.
Seltzer, B., & Pandya, D. N. (1989). Frontal lobe connections of the superior temporal sulcus in the rhesus monkey. *Journal of Comparative Neurology, 281,* 97–113.
Skipper, J. I., Goldin-Meadow, S., Howard, C., Nusbaum, H. C., & Small, S. L. (2007). Speech-associated gestures, Broca's area, and the human mirror system. *Brain and Language, 101,* 260–277.
Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2005). Listening to talking faces: Motor cortical activation during speech perception. *Neuroimage, 25,* 76–89.
Skipper, J. I., van Wassenhove, V., Howard, C., Nusbaum, H. C., & Small, S. L. (2007). Hearing lips and seeing voices: How

cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex, 17,* 2387–2399.
Soto-Faraco, S., Navarra, J., & Alsius, A. (2004). Assessing automaticity in audiovisual speech integration: Evidence from the speeded classification task. *Cognition, 92,* B13–B23.
Stevenson, R. A., Kim, S., & James, T. W. (2009). An additive-factors design to disambiguate neuronal and areal convergence: Measuring multisensory interactions between audio, visual, and haptic sensory streams using fMRI. *Experimental Brain Research, 198,* 183–194.
Szycik, G. R., Jansma, H., & Münte, T. F. (2009). Audiovisual integration during speech comprehension: An fMRI study comparing ROI-based and hole brain analyses. *Human Brain Mapping, 30,* 1990–1999.
Szycik, G. R., Tausche, P., & Münte, T. F. (2008). A novel approach to study audiovisual integration in speech perception: Localizer fMRI and sparse sampling. *Brain Research, 1220,* 142–149.
Teder-Salejärvi, W. A., Di Russo, F., McDonald, J. J., & Hillyard, S. A. (2005). Effects of spatial congruity on audio-visual multimodal integration. *Journal of Cognitive Neuroscience, 17,* 1396–1409.
Thompson, J. C., Jillian, E., Hardee, J. E., Panayiotou, A., Crewther, D., & Puce, A. (2007). Common and distinct brain activation to viewing dynamic sequences of face and hand movement. *Neuroimage, 37,* 966–973.
Troje, N. F., & Westhoff, C. (2006). The inversion effect in biological motion perception: Evidence for a "life detector". *Current Biology, 16,* 821–824.
Vaina, L. M., Solomon, J., Chowdhury, S., Sinha, P., & Belliveau, J. W. (2001). Functional neuroanatomy of biological motion perception in humans. *Proceedings of the National Academy of Sciences, U.S.A., 98,* 1165–1166.
van Atteveldt, N., Formisano, E., Goebel, R., & Blomert, L. (2004). Integration of letters and speech sounds in the human brain. *Neuron, 43,* 271–282.
Van Essen, D. C., Dickson, J., Harwell, J., Hanlon, D., Anderson, C. H., & Drury, H. A. (2001). An integrated software system for surface-based analyses of cerebral cortex. *Journal of American Medical Informatics Association, 8,* 443–459.
van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences, U.S.A., 102,* 1181–1186.
Watkins, K. E., Strafella, A. P., & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia, 41,* 989–994.
Werner, S., & Noppeney, U. (2009). Superadditive responses in superior temporal sulcus predict audiovisual benefits in object categorization. *Cerebral Cortex, 20,* 1829–1842.
Werner, S., & Noppeney, U. (2010a). Distinct functional contributions of primary sensory and association areas to audiovisual integration in object categorization. *Journal of Neuroscience, 30,* 2662–2675.
Werner, S., & Noppeney, U. (2010b). Superadditive responses in superior temporal sulcus predict audiovisual benefits in object categorization. *Cerebral Cortex, 20,* 1892–1842.
Wilke, M., & Lidzba, K. (2007). LI-tool: A new toolbox to assess lateralization in functional MR-data. *Journal of Neuroscience Methods, 163,* 128–136.
Wilson, M. W., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience, 7,* 701–702.
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.
Zaehle, T., Wüstenberg, T., Meyer, M., & Jäncke, L. (2004). Evidence for rapid auditory perception as the foundation of speech processing: A sparse temporal sampling fMRI study. *European Journal of Neuroscience, 20,* 1460–9568.