doi:10.1093/scan/nsu047 SCAN (2015) 10, 255–261

Eye'm talking to you: speakers' gaze direction modulates co-speech gesture processing in the right MTG

Judith Holler,[1,*] Idil Kokal,[2,*] Ivan Toni,[2] Peter Hagoort,[1,2] Spencer D. Kelly,[3] and Aslı Özyürek[1,4]
[1]Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525XD Nijmegen, The Netherlands, [2]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Geert Grooteplein Noord 21, 6525EZ Nijmegen, The Netherlands, [3]Psychology Department, Center for Language and Brain, Colgate University, 13 Oak Drive, 13346 Hamilton, NY, USA, and [4]Centre for Language Studies, Radboud University, Erasmusplein 1, 6525 HT Nijmegen, The Netherlands

Recipients process information from speech and co-speech gestures, but it is currently unknown how this processing is influenced by the presence of other important social cues, especially gaze direction, a marker of communicative intent. Such cues may modulate neural activity in regions associated either with the processing of ostensive cues, such as eye gaze, or with the processing of semantic information, provided by speech and gesture. Participants were scanned (fMRI) while taking part in triadic communication involving two recipients and a speaker. The speaker uttered sentences that were and were not accompanied by complementary iconic gestures. Crucially, the speaker alternated her gaze direction, thus creating two recipient roles: addressed (direct gaze) *vs* unaddressed (averted gaze) recipient. The comprehension of Speech&Gesture relative to SpeechOnly utterances recruited middle occipital, middle temporal and inferior frontal gyri, bilaterally. The calcarine sulcus and posterior cingulate cortex were sensitive to differences between direct and averted gaze. Most importantly, Speech&Gesture utterances, but not SpeechOnly utterances, produced additional activity in the right middle temporal gyrus when participants were addressed. Marking communicative intent with gaze direction modulates the processing of speech–gesture utterances in cerebral areas typically associated with the semantic processing of multi-modal communicative acts.

Keywords: co-speech gestures; speech–gesture integration; eye gaze; communicative intent; middle temporal gyrus

INTRODUCTION

In face-to-face conversation, the most common form of everyday talk, language is always accompanied by additional communicative signals—it is a multi-modal joint activity. However, traditionally, these various communicative signals have been investigated in isolation. Here, we investigate the interplay of three communicative modalities core to human interaction, speech, gesture and eye gaze. Our aim is to provide a first insight into the neural underpinnings of multi-modal language processing in a multi-party situated communication scenario. More precisely, our focus is on how recipients who are directly looked at by a speaker (addressees) process speech and co-speech gestures compared with recipients who are not looked at (unaddressed recipients) in a social, communicative context.

Co-speech gestures are speakers' spontaneous movements, typically of the hands and arms, which represent meaning that is closely related to the meaning in the speech that they accompany (e.g. depicting a round shape when referring to a full moon). By now there is much evidence, both behavioural and neurophysiological, that our brain processes and semantically integrates information from speech and co-speech gestures (e.g. Holle and Gunter, 2007; Wu and Coulson, 2007; Özyürek *et al.*, 2007; Kelly *et al.* 1999, 2004, 2007, 2010a,b; Holle *et al.*, 2012), primarily in the left inferior frontal gyrus (LIFG) (Skipper *et al.*, 2007; Willems *et al.*, 2007, 2009; Holle *et al.*, 2010; Straube *et al.*, 2011a) and bilateral middle temporal gyrus (MTG)/ the posterior superior temporal sulcus (pSTS) (Holle *et al.*, 2008; Dick *et al.*, 2009, 2012; Green *et al.*, 2009; Willems *et al.*, 2009; Straube *et al.*, 2011a)[1]. However, while aforementioned studies have made an important step in investigating the processing of both the verbal *and* gestural components of utterances (McNeill, 1992; Kendon, 2004), they have typically presented subjects with speech–gesture utterances in isolation of other communicative information that could be gleaned from the speaker's face (e.g. Kelly *et al.*, 2007, 2010a,b; Özyürek *et al.*, 2007; Willems *et al.*, 2007, 2009; Holle *et al.* 2008, 2010, 2012). Studies that have included the face have not manipulated eye gaze direction systematically (e.g. Kelly *et al.*, 2004; Dick *et al.*, 2009, 2012; Green *et al.*, 2009; Skipper *et al.*, 2009; Straube *et al.*, 2012). Thus, it remains unknown to what extent the neural processing of multi-modal speech–gesture messages may be influenced by face-related cues, such as the speaker's gaze direction.

Eye gaze is a powerful communicative cue (Pelphrey and Perlman, 2009; Senju and Johnson, 2009; Vogeley and Bente, 2010; Wilms *et al.*, 2010) and one of the first ones humans attend to (Farroni *et al.*, 2002). It is crucial in initiating and maintaining social interaction (Kendon, 1967; Argyle and Cook, 1976; Goodwin, 1981) and is tightly linked to the perception of communicative intent (Senju and Johnson, 2009). Despite its importance and omnipresence in face-to-face communication, eye gaze, too, has been predominantly investigated in isolation of other communicative modalities, especially language. The neural signature of eye gaze processing as such has been fairly well investigated, initially with paradigms employing static faces/eyes (e.g. Farroni *et al.*, 2002; Kampe *et al.*, 2003) followed by studies with more dynamic scenarios including gaze shifts, where these have been embedded in contexts (primarily Virtual Reality environments) simulating approach

Received 26 June 2013; Revised 20 February 2014; Accepted 17 March 2014
Advance Access publication 19 March 2014

J.H. was supported through a Marie Curie Fellowship (255569) and European Research Council Advanced Grant INTERACT (269484); A.O. was supported through European Research Council Starting Grant (240962). I.T. and I.K. were supported by VICI grant [453-08-002] from the Nederlandse Organisatie voor Wetenschappelijk Onderzoek. The authors would like to thank Manuela Schütze for acting as the confederate speaker, Nick Wood for video editing and Ronald Fischer and Pascal de Water for assistance with the Presentation script. They also thank the Neurobiology of Language department (Max Planck Institute for Psycholinguistics), the Intention & Action research group (Donders Institute for Brain, Cognition & Behaviour) and the Gesture & Sign research group (Max Planck Institute for Psycholinguistics and Centre for Language Studies, Radboud University) for valuable feedback on and discussion of this study.

Correspondence should be addressed to Judith Holler, Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525XD Nijmegen, The Netherlands. E-mail: judith.holler@mpi.nl

*These authors contributed equally to this work.

[1] Note that equivalent areas of neural activation have been found when manual movements are used for communication in the absence of speech, such as with pantomimes or signs (Emmorey *et al.*, 2007; MacSweeney *et al.*, 2008; Schippers *et al.*, 2009, 2010; Xu *et al.*, 2009).

and the initiation of interaction (e.g. Pelphrey *et al.*, 2004; Schilbach *et al.*, 2006). A recent study has taken this line of research even further by exploring eye gaze as an interactively contingent signal (Pfeiffer *et al.*, 2012). While becoming progressively more interactive, semantics are not yet a common feature of paradigms developed for exploring eye gaze processing. Only a handful of studies to date have explored the neural underpinnings of perceiving eye gaze and linguistic cues in conjunction, and this work has focused on infants (Parise *et al.*, 2011) or the specific effect of feeling addressed when hearing one's name (Stoyanova *et al.*, 2010).

Here, we pull these two strands of research together in order to investigate the neural processing of human multi-modal language comprehension in the context of eye gaze during situated social encounters. In our paradigm, participants were made to believe that they were engaging in a live communication task involving one speaker and two recipients. Crucially, the speaker alternated her gaze between the two recipients, thus rendering each of them momentarily addressed or unaddressed (Goffman, 1981; Goodwin, 1981). The continuously shifting recipient roles created a dynamic, situated communication setting and thus an opportunity for exploring the influence of social eye gaze on verbal and gestural communication in a more conversation-like context.

Two recent studies have shown that the neural integration of information from speech and gesture can be modulated by perceived communicative intentions. ERP studies by Kelly *et al.* (2007, 2010a) have demonstrated that our brain integrates speech and gesture less strongly when the two modalities are perceived as not intentionally coupled (i.e. gesture and speech being produced by two different persons) than when they are perceived as forming a composite utterance (i.e. gesture and speech being produced by the same person). This is an interesting finding which begs the question of whether pragmatic cues that provide interlocutors with information about the speaker's intentional stance in face-to-face contexts, such as the speaker's gaze direction, might also modulate the integration of gesture and speech (produced by the same person). A study by Straube *et al.* (2010) showed stronger activation in brain areas traditionally associated with 'mentalising' when participants observed a frontally compared with a laterally oriented speaker-gesturer. However, in their study, gaze direction was not manipulated independently of body or gesture orientation, and speech was always accompanied by gestures, preventing us from drawing conclusions about the influence of gaze direction on the processing of speech and gesture.

Here, we take the next step by asking whether there is neurophysiological evidence that the ostensive cue of social gaze modulates the integration of speech and co-speech gestures, and if so, where in the brain this modulation takes place. Several candidate regions offer themselves in this respect. One possibility is that semantic gesture–speech integration itself remains unaffected, with activity changes being evident mainly in cerebral areas involved in eye gaze processing. Eye gaze direction has been shown to involve a wide range of brain areas, but the right posterior STS and the medial prefrontal and orbitofrontal cortex are of particular interest here as they have been activated in studies using dynamic gaze stimuli (see Senju and Johnson, 2009, for a review), a feature present in our stimuli as well (see Method). Alternatively, the ostensive cue of eye gaze may modulate activity in areas directly involved in the semantic processing of speech and gesture, such as LIFG and MTG (discussed earlier). Yet another possibility is that perceived communicative intentions influence the integration of gesture with information from other modalities, but during early sensory rather than semantic processing stages. This may involve the integration of gestural information with information from ostensive social cues, such as eye gaze. A recent study has revealed that nonverbal, social, self-relevant cues (e.g. being pointed or gazed at) are neurally integrated in pre-motor areas (Conty *et al.*, 2012). This makes the motor system, and the SMA in particular, another candidate region for the modulation of multi-modal integration.

METHOD

Participants

Twenty-eight female native German speakers (age: 19–23 years), all right-handed, participated in the study after giving written consent according to the guidelines of the local ethics committee (Commissie Mensengebonden Onderzoek region Arnhem-Nijmegen, Netherlands). The participants had normal or corrected to normal vision and no history of neurological or psychiatric disorders. The participants received payment or course credits for their contribution. One participant was excluded from the analyses due to skepticism about the presence of other participants during the experiment.

Experimental set-up and procedure

Upon arrival, each participant was informed that they were going to engage in a triadic communication task involving one speaker and two recipients, with them taking on the role of one of the recipients. They were told that a one-way live audio–video connection would be established between her and the two other individuals (all located in separate rooms), and that the speaker had been placed in a room with two cameras in front of her, hooked up to two different computer monitors viewed by the two different recipients. Participants were further told that the speaker (who was actually a confederate) could view a laptop screen on a table in front of her (out of shot) displaying drawings and words that she had been asked to package into short communicative messages in a way that felt natural to her (no explicit mention of gesture was made). The idea behind this cover story was that it would have seemed implausible to participants that the speaker had learned the content of all messages by heart. Further, participants were told that she had been asked to sometimes address one and sometimes the other recipient by directing her gaze toward the respective camera.

In fact, the experiment involved only one real participant (the participant in the MR-scanner) (see Figure 1). The speaker shown to the participant was a pre-recorded video of a confederate producing scripted utterances, and the second recipient was fictive (which all participants included in our analyses believed). We decided to sacrifice the benefits of actual live interaction with spontaneous behaviour for experimental control to ensure that each participant processed identical stimuli under identical circumstances.

Two features of the experimental procedure were introduced to increase the likelihood that participants believed to be engaged in a live communicative scenario. First, shortly after the participant was positioned inside the MR-scanner, an introductory video clip (14 s) was presented in which the speaker introduced herself both to the participant and to a (fictive, unseen) second recipient. This procedure was also instrumental to adjusting the volume of the audio system to each participant's hearing abilities. Second, each participant was told that there could be technical problems with the video link and asked to report any visual disturbance in the quality of the video by pressing a button of an MR-compatible box with their right thumb (fORP, Current Designs, USA). In fact, these disturbances were 16 prearranged fillers (in half of the filler trials speech was accompanied by gestures). During these fillers, the video would turn monochrome after a variable epoch (range: 1–2 s) following video onset. This task feature allowed us to monitor participants' attention during the experiment. To ensure that participants would process the gestural and spoken information, they were instructed to attend to the speaker, regardless of her gaze direction, and that at the end of the experiment they would be quizzed about the content of the communicated messages (this test

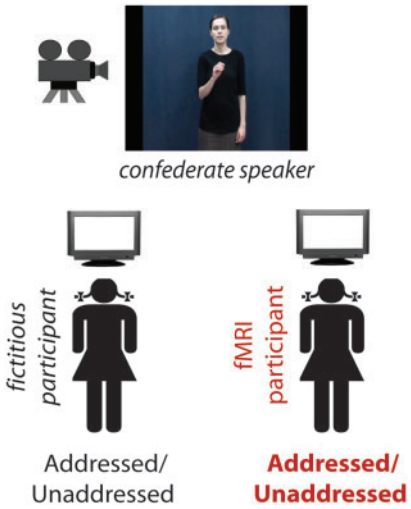

Fig. 1 Illustration of the experimental set-up.

Fig. 2 Example of stills from the four types of video stimuli used.

was announced as a motivational measure only and was not constructed or evaluated for actual analysis; however, see Straube *et al.*, 2010, 2011b, for the influence of recipient status on sentence memory).

Experimental design

There were four experimental conditions (160 trials), and an attentional control condition (16 filler trials), pseudo-randomly varying according to an event-related fMRI design, including four different trial orders counterbalanced across participants. Each trial consisted of a video clip of a spoken sentence performed by a female native German speaker, with or without co-speech gestures, and with the speaker looking either straight at the participant or slightly to the left of the participant to the allegedly second recipient. This resulted in a 2×2 factorial design, with COMMUNICATIVE MODALITY (SpeechOnly, Speech&Gesture) and RECIPIENT STATUS (Addressed Recipient, Unaddressed Recipient) as factors (see Figure 2).

Stimuli

The sentences produced by the speaker consisted of 3–5 words, with the same syntactical structure (subject-verb-object), and were in German. During the SpeechOnly trials, the speaker did not move her hands. During the Speech&Gesture trials, the speaker produced scripted iconic gestures that always complemented the content of speech. For instance, the speaker uttered the sentence 'she trains the horse' ('sie trainiert das Pferd'). The action described by this sentence is underspecified in terms of the aspect of manner, as training a horse can involve a range of things, including riding it, feeding it a treat as reward, etc. The gestures accompanying the sentences always specified the manner of action, in this example by depicting a whipping action. Each video clip started with the speaker looking down (at the laptop, see Experimental set-up and procedure), before raising her head and orienting her eyes towards one of the two cameras. After this orientation phase (average duration: 689 ms), the speaker uttered a sentence and then lowered her gaze again. Each video clip was followed by a baseline condition (a white fixation cross on a black background) with a variable duration of 4–6 s (jittered; average = 5 s).

One of our aims was to avoid confounding the visual angle of the gestures in the addressed and unaddressed conditions with recipient status signaled through eye gaze direction. Thus, rather than showing the same gestures recorded from a lateral and a frontal visual angle, the confederate speaker repeated each stimulus sentence, including the gestures, once with direct and once with averted gaze (order counterbalanced during recording to avoid possible order effects), while the perspective on the gesture was held constant (see Figure 2). Consistency in intonation and gesture execution was checked by one of the experimenters (J.H.). We also carried out a pre-test to make sure the gestures were equally well interpretable in the two conditions. An independent set of participants ($N = 16$) with similar demographic characteristics to the fMRI participants were asked to write down the meaning of the gestures, one group ($N = 8$) for the 160 gestures in the averted gaze videos and the other ($N = 8$) for the 160 gestures in the frontal gaze videos (the speaker's head was masked to avoid gaze direction influencing participants' ratings). Results revealed no difference in the frequency with which the correct interpretation was given for the corresponding video clips in the two gesture-video sets, $t(318) = 0.68$, $P = 0.495$).

Apparatus

The video clips were presented to participants inside the MR-scanner through an LCD projector directed at a mirror positioned on top of the MR-head coil. The clips were shown at a viewing distance of 80 cm. The spoken sentences were presented to the participants through MR-compatible earplugs (Sensometric, Malden, MA, USA) to dampen scanner noise. Stimuli and responses were software-controlled with Presentation 13.0 (Neurobehavioral Systems, Davis, CA, USA). The presentation of each video clip (average duration = 2847 ms) was followed by the presentation of a white fixation cross on a black background (4000–6000 ms).

fMRI data acquisition and analyses

Data acquisition

All functional images were acquired on a 3T MRI-scanner (Trio, Siemens Medical Systems, Erlangen, Germany) equipped with a 32-channel head coil using a multiecho GRAPPA sequence (Poser *et al.*, 2006) [repetition time (TR): 2.35 s, echo times (TEs, 4): 9.4/21.2/33/45 ms, 36 transversal slices, ascending acquisition, distance factor: 17%, effective voxel size $3.5 \times 3.5 \times 3.0$ mm, field of view (FoV): 212 mm]. A T1-weighted structural scan was acquired with TR = 2300 ms, TE = 3.03 ms, 192 sagittal slices, voxel size $1.0 \times 1.0 \times 1.0$ mm, flip angle = 90°.

Data preprocessing

We used SPM8 (www.fil.ion.ucl.ac.uk/spm) implemented in MATLAB 7.11 (Mathworks Inc., Sherborn, MA, USA) for data analyses. The first four volumes of each participant's EPI time series were discarded to allow for T1 equilibration. Head motion parameters for spatial realignment were estimated on the MR images with the shortest echo time (9.4 ms), using a least-squares approach with six parameters (three translations, three rotations). Following spatial realignment, applied to each of the four echo images collected for each excitation, the four echo images were combined into a single EPI volume using an optimised echo weighting method (Poser *et al.*, 2006). The fMRI time series were transformed and resampled at an isotropic voxel size of 2 mm into the standard Montreal Neurological Institute (MNI) space using both linear and nonlinear transformation parameters as determined in a probabilistic generative model that combines image registration, tissue classification, and bias correction (i.e. unified segmentation and normalisation) of the co-registered T1-weighted image (Ashburner and Friston, 2005). The normalised functional images were spatially smoothed using an isotropic 8 mm full-width at half-maximum Gaussian kernel.

Statistical inference

The fMRI time series of each subject were analysed using an event-related approach in the context of the general linear model. Vectors describing the onset and duration of each video clip of the five conditions (four experimental + one filler condition) were convolved with a canonical haemodynamic response function and its temporal derivative, yielding 10 task-related regressors. The potential confounding effects of residual head movement-related effects were modeled using the original, squared, first-order and second-order derivatives of the movement parameters as estimated by the spatial realignment procedure (Lund *et al.*, 2005). Three further regressors, describing the time course of signal intensities averaged over different image compartments (i.e. white matter, cerebrospinal fluid and the portion of the MR-image outside the skull) were also added (Verhagen *et al.*, 2008). Finally, the fMRI time series were high-pass filtered (cut-off 128 s). Temporal autocorrelation was modeled as a first-order autoregressive process.

Consistent effects across subjects were tested using a random effects multiple regression analysis that considered, for each subject, four contrast images relative to the (SpeechOnly, Speech&Gesture) × (Addressed Recipient, Unaddressed Recipient) combinations of the 2 × 2 factorial design used in this study. Anatomical inference is drawn by superimposing the SPMs showing significant signal changes on the structural images of the subjects. Anatomical landmarks were identified using the cytoarchitectonic areas based on the anatomy toolbox (Eickhoff *et al.*, 2005) for SPM.

RESULTS

Behavioral data

One experimenter (I.K.) monitored participants' performance and gaze behavior during the experiment to check that the video clips were attended to during the experiment. Participants successfully detected the filler items (mean: 12.7, s.d.: 6.2 out of 16 fillers). One participant (the same one who doubted our triadic set-up) was excluded because she stopped looking at the video clips in the second half of the experiment.

Functional MRI data

Main effect of COMMUNICATIVE MODALITY

Table 1 reports cerebral regions with stronger responses during the processing of the Speech&Gesture compared with the SpeechOnly utterances. There were spatially widespread effects, largely bilateral, reflecting the differences in both sensory and communicative features between the two conditions. Significantly differential effects were found in the LIFG including BA 44 and BA 45, right inferior frontal gyrus including BA 45, left inferior parietal lobule (PF), left superior parietal lobule (7A), bilateral middle occipital gyri (MOG), MTG, bilateral fusiform gyri, and bilateral hippocampi, amygdale, and thalamus. Differences in BOLD signal between the SpeechOnly and the Speech&Gesture conditions were found in the superior frontal gyri, orbital gyri, bilateral middle frontal gyri (MFG), right fusiform gyrus and right superior parietal lobule.

Table 1 Cerebral regions with significant effect of COMMUNICATIVE MODALITY

Brain region	Hemisphere	Cluster size	Local maxima			*Z*-value
MT+	L	8003	−52	−72	−6	7.6
Middle temporal gyrus	L		−50	−58	−2	6.4
Middle occipital gyrus	L		−38	−62	4	6.3
Middle occipital gyrus	R	5275	46	−64	4	7.5
Fusiform gyrus	R		44	−54	−20	6.1
Superior temporal gyrus	R		58	−36	12	5.8
Inferior temporal gyrus	R		44	−52	−10	5.7
Inferior frontal gyrus	L	963	−52	12	16	4.3
Inferior frontal gyrus	L		−50	30	12	3.8
Inferior frontal gyrus	R	339	56	34	4	5.3
Amygdala	R	602	18	−4	−16	4.9
Hippocampus	R		32	−4	−24	3.8
Inferior parietal lobule	L	231	−44	−48	58	3.8
Superior parietal lobule	L		−34	−56	60	3.5
Thalamus	R	355	20	−26	0	5.7
Thalamus	R		30	−20	−10	3.3
Thalamus	L	211	−16	28	−2	5.1
Fusiform gyrus	L	337	−30	−4	−34	5.2
Amygdala	L		−22	−6	−16	4.1
Hippocampus	L		−22	0	−34	3.7

Cluster-level statistical inferences were corrected for multiple comparisons using family-wise error (FWE) correction (Friston et al. 1996; FWE: $P < 0.05$, on the basis of an intensity threshold of $t > 3.4$). MNI stereotactic coordinates of the local maxima of regions showing stronger responses during processing of speech and gesture videos than speech-only videos. For large clusters spanning several anatomical regions, more than one local maximum are given. Cluster size is given in number of voxels.

Main effect of RECIPIENT STATUS

When comparing cerebral responses during the processing of utterances in which the speaker's gaze was directed at the participant (Addressed) with utterances in which the speaker's gaze was averted from the participant (Unaddressed), there was a single significant cluster in the right inferior occipital gyrus (calcarine gyrus: 20, −92, −6; *z*-value: 4.76, cluster size: 335 voxels). The reverse comparison revealed significant effects in the left posterior cingulate cortex (−12, −50, 38; *z*-value: 4.32, cluster size: 268 voxels).

Interaction between RECIPIENT STATUS and COMMUNICATIVE MODALITY

The focus of this study is on whether recipient status influences the processing of speech and gestures. This effect corresponds to the increase of the cerebral effect of processing Speech&Gesture (as compared with SpeechOnly) as an Addressed Recipient (as compared with an Unaddressed Recipient). A portion of the right MTG showed this significant interaction (48, −64, 12; *z*-value: 4.54, cluster size: 192 voxels, Figure 3). This cluster remained significant when masked ($P < 0.05$ uncorrected) with the simple main effects of Speech&Gesture Addressed Recipient *vs* Speech&Gesture Unaddressed Recipient and Speech&Gesture Addressed Recipient *vs* SpeechOnly Addressed Recipient, indicating that this interaction

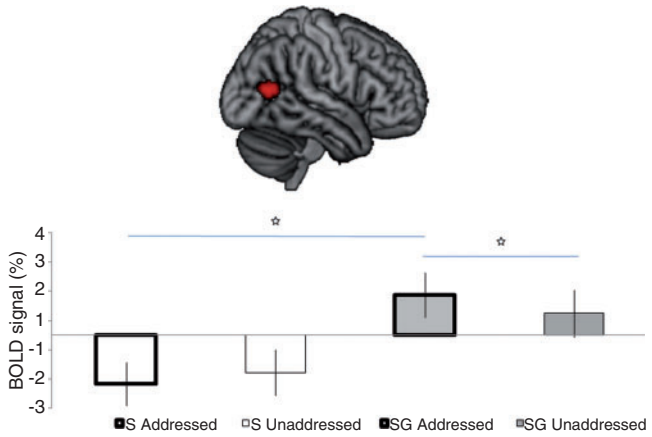

Fig. 3 Anatomical location of a cluster along the right middle temporal gyrus (in red, overlaid on a rendered brain) showing a significant differential response to Speech&Gesture (SG) utterances [as compared with SpeechOnly (S)] when participants were addressed (as compared with unaddressed). *Post hoc* paired *t*-tests indicated significant differences between the conditions marked with a star [SG-Addressed *vs* SG-Unaddressed: $t(27) = 3.238$, $P = 0.003$; SG-Addressed *vs* S-Addressed: $t(27) = 8.706$, $P = 0.001$].

effect was driven by stronger BOLD responses during the Speech&Gesture Addressed Recipient trials (see Figure 3).

DISCUSSION

The neural processing of multi-modal language in situated communication contexts as a domain of scientific enquiry is relatively unexplored. The present study provides an initial glimpse of how different communicative modalities that are core to human face-to-face interaction influence each other on a neural level during comprehension. Specifically, we focused on the influence of the ostensive cue of eye gaze direction on the processing of semantic information from speech and iconic gestures in a situated triadic communication scenario. In this scenario, eye gaze fulfilled the social function of indicating which of two recipients was directly addressed at any given moment, an important function regulating coordination in everyday multi-party conversation (Goodwin, 1981).

The results show that a speaker's gaze towards different recipients modulated co-speech gesture processing. This interaction was driven by a stronger neural response in the right MTG when addressed recipients (direct gaze) were presented with Speech&Gesture utterances than when unaddressed recipients (averted gaze) were presented with Speech&Gesture utterances (while recipient status did not lead to significant differences in the processing of uni-modal SpeechOnly utterances).

Both the LIFG and the MTG have been identified as primary locations for speech–gesture integration (e.g. Skipper *et al.*, 2007; Willems *et al.*, 2007, 2009; Dick *et al.*, 2009, 2012; Green *et al.*, 2009; Holle *et al.*, 2010; Straube *et al.*, 2011a, 2012). Here, we have provided corroborating evidence to this extent, as both neural regions were activated more strongly in our bi-modal compared with our uni-modal conditions. However, LIFG and MTG have also been described as fulfilling different functions with respect to integration during language processing (Hagoort, 2005), including the multi-modal integration of language with action (Willems *et al.*, 2009). While MTG seems to predominantly be involved with integrating information from different input streams when this information maps onto a common, stable conceptual representation (such as the picture of a sheep with the sound it makes or the word 'to write' with a pantomimic depiction of writing), LIFG seems to be the predominant neural region for unification (i.e. the integration of lexical items, or words and gestures, into coherent sentential representations) (Hagoort, 2005; Willems *et al.*, 2009). These different functions may help to explain why, in the present study, we observed eye gaze modulating speech–gesture integration in MTG but not in LIFG. In our study, speech and gesture were always complementary, that is, speech provided global information about an action (e.g. 'to train') while the gesture further specified the manner of action (e.g. 'whipping'). Thus, the action information provided by speech and gesture mapped essentially onto the same concept, with the meaning depicted gesturally being a sub-category of the meaning provided by speech (i.e. whipping as part of training a horse). In our study, the MTG is therefore a likely candidate for computing this sort of bi-modal conceptual matching (see also Dick *et al.*, 2012), which nicely fits with research by Kable *et al.* (2005) pinpointing the MTG as being particularly involved in the conceptual representation of actions.

That our modulation of speech–gesture integration through eye gaze occurred exclusively in the right hemisphere makes sense considering that our paradigm involved a comparatively rich social context for gesture and speech. After all, the right hemisphere is often associated with information processing of a more social and pragmatic nature, especially related to communicative intentions, such as jokes, irony, figurative language, metaphors and indirect requests (Weylman *et al.*, 1989; Bottini *et al.*, 1994; Beeman and Chiarello, 1998; Sabbagh, 1999; Coulson and Wu, 2005; Noordzij *et al.*, 2009; Paz Fonseca *et al.*, 2009; Weed, 2011). Our creation of dynamic, situated and pragmatically complex communication scenarios, in combination with social gaze indicating recipient status, may therefore explain why we found a right-lateralised effect and other neuroimaging studies on co-speech gesture have not. In fact, the MTG cluster activated in the present study might indeed encompass portions of the right posterior superior temporal sulcus that have previously been associated with the processing of both linguistic and non-linguistic intentions (Noordzij *et al.* 2009; Enrici *et al.*, 2011), as well as with the perception of intentions associated with dynamic eye gaze stimuli (Pelphrey *et al.*, 2003; Mosconi *et al.*, 2005; Senju and Johnson, 2009—note that, unlike in the present study, gaze shifts in these studies occurred towards the left and the right side, making it unlikely that our right-lateralised interaction effect is due to the left-lateralised gaze shifts). Furthermore, a recent study has pinpointed the right MTG as one cerebral area involved in the processing of communicative intentions associated with the production of pointing gestures (Cleret de Langavant *et al.*, 2011). The present findings suggest that this link may generalise also to the comprehension of iconic gestures.

Thus, our results fit the notion of LIFG and MTG fulfilling core but differential roles in the multi-modal integration of speech and gesture. However, although MTG is involved in the lower-level integration of different input streams while the higher-order unification processes happen in LIFG (Willems *et al.*, 2009), at least the right MTG might be influenced by higher-order pragmatic processes.

Some readers may wonder what evidence there is that our participants processed speech and gesture indeed semantically. There are several reasons to believe that this was the case. First, in line with past research, our results showed significant activation in the two main areas (LIFG and MTG) of gesture–speech semantic integration in response to our bi-modal compared with our uni-modal stimuli. Further assurance comes from the fact that our participants believed that their understanding of the presented communicative messages would be tested at the end of the study. Finally, a behavioural study employing the same basic paradigm and stimuli showed that both addressed and unaddressed recipients took significantly longer to make content-related judgements for the bi-modal than the uni-modal messages (Holler *et al.*, 2012).

Another question is whether the different activations of MTG in the Speech&Gesture condition for addressed and unaddressed recipients indeed reflect differences in integration of those two modalities. While there is evidence that the MTG is implicated in speech–gesture integration (e.g. Holle *et al.*, 2008; Kircher *et al.*, 2009; Straube *et al.*, 2011a), this brain area also appears to be involved in the semantic processing of gestures in the absence of speech (e.g. Straube *et al.*, 2012). As we did not use a gesture-only condition, we have to remain tentative regarding our integration interpretation. However, in light of the number of previous studies pinpointing the MTG as a speech–gesture integration hub, our favoured interpretation of the present data is that the differences in MTG activation do reflect differences in speech–gesture integration. This interpretation is interesting in light of some recent results from a behavioural study using the same stimuli in a Stroop-like task (Holler *et al.*, 2012). Those findings suggest that unaddressed recipients may be processing gestures more strongly than addressees. Putting the two studies together, it is possible that although unaddressed recipients process information from the gestural modality quite strongly (i.e. they zoom into gesture more than into speech), they fail to successfully integrate it with information conveyed though speech. However, as Holler *et al.*'s (2012) findings do not unequivocally rule out the possibility that their unaddressed recipients may actually have processed gestures less rather than more strongly than addressees, it will be important for future research to determine whether gaze direction modulates gesture processing independently of speech or whether it actually affects the process of gesture–speech integration per se.

Our first step in drawing together two different strands of research has proven fruitful from several perspectives. For one thing, our results show that a social, ostensive cue can impact on the processing of semantic information from two concurrent modalities, speech and co-speech gestures. It thus underlines the remarkable power of eye gaze in human communication. For another, our results reveal that when eye gaze is observed in the context of semantic communication, some brain areas that have often been associated with dynamic eye gaze processing (e.g. pSTS, mPFC, OFC, see Senju and Johnson, 2009) do not distinguish significantly between direct and averted gaze in these more contextualised communication settings. Instead, our main effect of gaze led to activation in the calcarine gyrus and posterior cingulate cortex; the former was also significantly activated in Straube *et al.*'s (2010) study in response to participants observing a frontally as compared with a laterally oriented speaker. The calcarine gyrus may thus reflect sensitivity to whether one is communicatively addressed, or attended to, through the visual cue of eye gaze (and/or other bodily cues) in more situated, multi-modal contexts (because occipital areas are directly associated with the processing of visuo-spatial information, the right-lateralised activation observed in the present study may be due to the actor's gaze having been exclusively directed towards the left side of the screen). The posterior cingulate cortex was activated more strongly when gaze was being averted. The role of this region is rather unclear (Leech *et al.*, 2012), but one potential interpretation is that its activation was associated with perspective taking (Ruby and Decety, 2001), or mind-reading (Fletcher *et al.*, 1995; Brunet *et al.*, 2000), as participants may have engaged more in perspective taking when the other recipient was addressed rather than themselves. However, further scrutiny of this assumption is required before firm conclusions can be drawn. Finally, it is interesting that pre-motor areas involved in the binding of self-relevant visual social cues (including gaze and gesture) (Conty *et al.*, 2012) did not emerge as a primary binding site in the present study. This is a strong indicator that visual signals, including gaze and gesture, may be processed quite differently in the context of speech, especially in joint activities focussing on the exchange of propositional meaning.

To conclude, here, we have shown that the processing of multimodal speech–gesture messages is modulated by recipient status as indicated through social eye gaze. The present findings suggest that the right MTG in particular appears to play a core role in modulating speech–gesture utterance comprehension when it is situated in a pragmatically rich social context simulating face-to-face multi-party communication.

REFERENCES

Argyle, M., Cook, M. (1976). *Gaze and Mutual Gaze*. Cambridge: Cambridge University Press.

Ashburner, J., Friston, K.J. (2005). Unified segmentation. *Neuroimage*, *26*, 839–51.

Beeman, M.J., Chiarello, C. (1998). Complementary right- and left-hemisphere language comprehension. *Current Directions in Psychological Science*, *7*, 2–7.

Bottini, G., Corcoran, R., Sterzi, R., et al. (1994). The role of the right hemisphere in the interpretation of figurative aspects of language: A positron emission tomography activation study. *Brain*, *117*, 1241–53.

Brunet, E., Sarfati, Y., Hardy-Baylé, M.-C., Decety, J. (2000). A PET investigation of the attribution of intentions with a nonverbal task. *Neuroimage*, *11*, 157–66.

Cleret de Langavant, L., Remy, P., Trinkler, I., et al. (2011). Behavioral and neural correlates of communication via pointing. *PLoS One*, *6*, e17719.

Conty, L., Dezecache, G., Hugueville, L., Grèzes, J. (2012). Early binding of gaze, gesture and emotion: neural time course and correlates. *Journal of Neuroscience*, *32*, 4531–9.

Coulson, S., Wu, Y.C. (2005). Right hemisphere activation of joke-related information: an event-related brain potential study. *Journal of Cognitive Neuroscience*, *17*, 494–506.

Dick, A.S., Goldin-Meadow, S., Hasson, U., Skipper, J., Small, S.L. (2009). Co-speech gestures influence neural responses in brain regions associated with semantic processing. *Human Brain Mapping*, *30*, 3509–26.

Dick, A.S., Mok, E.H., Beharelle, A.R., Goldin-Meadow, S., Small, S.L. (2012). Frontal and temporal contributions to understanding the iconic co-speech gestures that accompany speech. *Human Brain Mapping*, *35*, 900–17.

Eickhoff, S.B., Stephan, K.E., Mohlberg, H., et al. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *Neuroimage*, *25*, 1325–35.

Emmorey, K., Mehta, S., Grabowski, T.J. (2007). The neural correlates of sign versus word production. *Neuroimage*, *36*, 202–8.

Enrici, I., Adenzato, M., Cappa, S., Bara, B.G., Tettamanti, M. (2011). Intention processing in communication: a common brain network for language and gestures. *Journal of Cognitive Neuroscience*, *23*, 2415–31.

Farroni, T., Csibra, G., Simion, F., Johnson, M.H. (2002). Eye contact detection in humans from birth. *Proceedings of the National Academy of Sciences USA*, *99*, 9602–5.

Fletcher, P.C., Happé, F., Frith, U., et al. (1995). Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. *Cognition*, *57*, 109–28.

Goffman, E. (1981). *Forms of Talk*. Philadelphia: University of Pennsylvania Press.

Goodwin, C. (1981). *Conversational Organization: Interaction between Speakers and Hearers*. New York: Academic Press.

Green, A., Straube, B., Weis, S., et al. (2009). Neural integration of iconic and unrelated coverbal gestures: a functional MRI study. *Human Brain Mapping*, *30*, 3309–24.

Hagoort, P. (2005). On Broca, brain, and binding: a new framework. *Trends in Cognitive Sciences*, *9*, 416–23.

Holle, H., Gunter, T.C. (2007). The role of iconic gestures in speech disambiguation: ERP evidence. *Journal of Cognitive Neuroscience*, *19*, 1175–92.

Holle, H., Gunter, T.C., Rüschemeyer, S.A., Hennenlotter, A., Iacoboni, M. (2008). Neural correlates of the processing of co-speech gestures. *Neuroimage*, *39*, 2010–24.

Holle, H., Obermeier, C., Schmidt-Kassow, M., Friederici, A.D., Ward, J., Gunter, T.C. (2012). Gesture facilitates the syntactic analysis of speech. *Frontiers in Psychology*, *3*, 74.

Holle, H., Obleser, J., Rüschemeyer, S.A., Gunter, T.C. (2010). Integration of iconic gestures and speech in left superior temporal areas boosts speech comprehension under adverse listening conditions. *Neuroimage*, *49*, 875–84.

Holler, J., Kelly, S., Hagoort, P., Ozyurek, A. (2012). When gestures catch the eye: the influence of gaze direction on co-speech gesture comprehension in triadic communication. In: Miyake, N., Peebles, D., Cooper, R.P., editors. *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Society, pp. 467–72.

Kable, J.W., Kan, I.P., Wilson, A., Thompson-Schill, S.L., Chatterjee, A. (2005). Conceptual representations of action in the lateral temporal cortex. *Journal of Cognitive Neuroscience*, *17*, 1855–70.

Kampe, K., Frith, C.D., Frith, U. (2003). 'Hey John': signals conveying communicative intention towards the self activate brain regions associated with mentalising regardless of modality. *Journal of Neuroscience*, *23*, 5258–63.

Kelly, S.D., Barr, D., Church, R.B., Lynch, K. (1999). Offering a hand to pragmatic understanding: the role of speech and gesture in comprehension and memory. *Journal of Memory and Language*, *40*, 577–92.

Kelly, S.D., Creigh, P., Bartolotti, J. (2010a). Integrating speech and iconic gestures in a Stroop-like task: evidence for automatic processing. *Journal of Cognitive Neuroscience*, *22*, 683–94.
Kelly, S.D., Kravitz, C., Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language*, *89*, 253–60.
Kelly, S.D., Özyürek, A., Maris, E. (2010b). Two sides of the same coin: speech and gesture mutually interact to enhance comprehension. *Psychological Science*, *21*, 260–7.
Kelly, S.D., Ward, S., Creigh, P., Bartolotti, J. (2007). An intentional stance modulates the integration of gesture and speech during comprehension. *Brain and Language*, *101*, 222–33.
Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica*, *26*, 22–63.
Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.
Kircher, T., Straube, B., Leube, D., et al. (2009). Neural interaction of speech and gesture: differential activations of metaphoric co-verbal gestures. *Neuropsychologia*, *47*, 169–79.
Leech, R., Braga, R., Sharp, D.J. (2012). Echoes of the brain within the posterior cingulate cortex. *The Journal of Neuroscience*, *32*, 215–22.
Lund, T.E., Nbrgaard, M.D., Rostrup, E., Rowe, J.B., Paulson, O.B. (2005). Motion or activity: their role in intra- and inter-subject variation in fMRI. *Neuroimage*, *26*, 960–4.
MacSweeney, M., Capek, C.M., Campbell, R., Woll, B. (2008). The signing brain: the neurobiology of sign language. *Trends in Cognitive Sciences*, *12*, 432–40.
McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. Chicago: University of Chicago Press.
Mosconi, M.W., Mack, P.B., McCarthy, G., Pelphrey, K.A. (2005). Taking an "intentional stance" on eye-gaze shifts: A functional neuroimaging study of social perception in children. *Neuroimage*, *27*, 247–52.
Noordzij, M.L., Newman-Norlund, S.E., Ruiter, J.P.A.de, Hagoort, P., Levinson, S.C., Toni, I. (2009). Brain mechanisms underlying human communication. *Frontiers in Human Neuroscience*, *3*, art. 14.
Özyürek, A., Willems, R.M., Kita, S., Hagoort, P. (2007). On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. *Journal of Cognitive Neuroscience*, *19*, 605–16.
Parise, E., Handl, A., Palumbo, L., Friederici, A.D. (2011). Influence of eye gaze on word processing: an ERP study with infants. *Child Development*, *82*, 842–53.
Paz Fonseca, R., Scherer, L.C., de Oliveira, C.R., de Mattos Pimenta Parente, M.A. (2009). Hemispheric specialization for communicative processing: neuroimaging data on the role of the right hemisphere. *Psychology & Neuroscience*, *2*, 25–33.
Pelphrey, K.A., Perlman, S.B. (2009). Charting brain mechanisms for the development of social cognition. In: Rumsey, J.M., Ernst, M., editors. *Neuroimaging in Developmental Clinical Neuroscience*. Cambridge: Cambridge University Press, pp. 73–90.
Pelphrey, K.A., Singerman, J.D., Allison, T., McCarthy, G. (2003). Brain activation evoked by the perception of gaze shifts: the influence of context. *Neuropsychologia*, *41*, 156–70.
Pelphrey, K.A., Viola, R.J., McCarthy, G. (2004). When strangers pass: processing of mutual and averted gaze in the superior temporal sulcus. *Psychological Science*, *15*, 598–603.
Pfeiffer, U.J., Schilbach, L., Jording, M., Timmermans, B., Bente, G., Vogeley, K. (2012). Eyes on the mind: investigating the influence of gaze dynamics on the perception of others in real-time social interaction. *Frontiers in Cognitive Science*, *3*, 537.
Poser, B.A., Versluis, M.J., Hoogduin, J.M., Norris, D.G. (2006). BOLD contrast sensitivity enhancement and artifact reduction with multiecho EPI: Parallel-acquired inhomogeneity-desensitized fMRI. *Magnetic Resonance in Medicine*, *55*, 1227–35.
Ruby, P., Decety, J. (2001). Effect of subjective perspective taking during simulation of action: a PET investigation of agency. *Nature Neuroscience*, *4*, 546–50.
Sabbagh, M. (1999). Communicative intentions and language: evidence from right-hemisphere damage and autism. *Brain and Language*, *70*, 29–69.
Schilbach, L., Wohlschläger, A.M., Newen, A., et al. (2006). Being with others: neural correlates of social interaction. *Neuropsychologia*, *44*, 718–30.
Schippers, M.B., Gazzola, V., Goebel, R., Keysers, C. (2009). Playing Charades in the fMRI: Are mirror and/or mentalizing areas involved in gestural communication? *PLoS One*, *4*, e6801.
Schippers, M.B., Roebroeck, A., Renken, R.J., Nanetti, L., Keysers, C. (2010). Mapping the information flow from one brain to another during gestural communication. *Proceedings of the National Academy of Sciences USA*, *107*, 9388–93.
Senju, A., Johnson, M.H. (2009). The eye contact effect: Mechanisms and development. *Trends in Cognitive Sciences*, *13*, 127–34.
Skipper, J.I., Goldin-Meadow, S., Nusbaum, H.C., Small, S.L. (2007). Speech associated gestures, Broca's area, and the human mirror system. *Brain and Language*, *101*, 260–77.
Skipper, J.I., Goldin-Meadow, S., Nusbaum, H.C., Small, S.L. (2009). Gestures orchestrate brain networks for language understanding. *Current Biology*, *19*, 661–7.
Stoyanova, R.S., Ewbank, M.P., Calder, A.J. (2010). You talkin' to me? Self-relevant auditory signals influence perception of gaze direction. *Psychological Science*, *21*, 1765–9.
Straube, B., Green, A., Bromberger, B., Kircher, T. (2011a). The differentiation of iconic and metaphoric gestures: common and unique integration processes. *Human Brain Mapping*, *32*, 520–33.
Straube, B., Green, A., Chatterjee, A., Kircher, T. (2011b). Encoding social interactions: the neural correlates of true and false memories. *Journal of Cognitive Neuroscience*, *23*, 306–24.
Straube, B., Green, A., Jansen, A., Chatterjee, A., Kircher, T. (2010). Social cues, mentalizing and the neural processing of speech accompanied by gestures. *Neuropsychologia*, *48*, 382–93.
Straube, B., Green, A., Weis, S., Kircher, T. (2012). A supramodal neural network for speech and gesture semantics: an fMRI study. *PLoS One*, *7*, e51207.
Verhagen, L., Dijkerman, H.C., Grol, M.J., Toni, I. (2008). Perceptuo-motor interactions during prehension movements. *The Journal of Neuroscience*, *28*, 4726–35.
Vogeley, K., Bente, G. (2010). 'Artificial humans': psychology and neuroscience perspectives on embodiment and nonverbal communication. *Neural Networks*, *23*, 1077–90.
Weed, E. (2011). What's left to learn about right hemisphere damage and pragmatic impairment? *Aphasiology*, *25*, 872–89.
Weylman, S.T., Brownell, H.H., Roman, M., Gardner, H. (1989). Appreciation of indirect requests by left- and right-brain-damaged patients: the effects of context and conventionality of wording. *Brain and Language*, *36*, 580–91.
Willems, R.M., Özyürek, A., Hagoort, P. (2007). When language meets action: The neural integration of gesture and speech. *Cerebral Cortex*, *17*, 2322–33.
Willems, R.M., Özyürek, A., Hagoort, P. (2009). Differential roles for left inferior frontal and superior temporal cortex in multimodal integration of action and language. *Neuroimage*, *47*, 1992–2004.
Wilms, M., Schilbach, L., Pfeiffer, U., Bente, G., Fink, G.R., Vogeley, K. (2010). It's in your eyes—using gaze-contingent stimuli to create truly interactive paradigms for social cognitive and affective neuroscience. *Social Cognitive and Affective Neuroscience*, *5*, 98–107.
Wu, Y.C., Coulson, S. (2007). How iconic gestures enhance communication: An ERP study. *Brain and Language*, *101*, 234–45.
Xu, J., Gannon, P., Emmorey, K., Smith, J.F., Braun, A.R. (2009). Symbolic gestures and spoken language are processed by a common neural system. *Proceedings of the National Academy of Sciences USA*, *106*, 20664–9.