The multisensory perception of co-speech gestures – A review and meta-analysis of neuroimaging studies

Lars Marstaller\textsuperscript{1,2} and Hana Burianová\textsuperscript{2,3}

\textsuperscript{1}Department of Cognitive Science, Macquarie University, Sydney, Australia

\textsuperscript{2}ARC Centre of Excellence in Cognition and its Disorders, Macquarie University, Sydney, Australia

\textsuperscript{3}Centre for Advanced Imaging, University of Queensland, Brisbane, Australia

Correspondence should be addressed to:
Lars Marstaller
Centre for Advanced Imaging
University of Queensland
QLD 4071, AUSTRALIA
l.marstaller@uq.edu.au

Tables: 2
Figures: 1
Abstract

Co-speech gestures constitute a unique form of multimodal communication because here the hand movements are temporally synchronized and semantically integrated with speech. Recent neuroimaging studies indicate that the perception of co-speech gestures might engage a core set of frontal, temporal, and parietal areas. However, no study has compared the neural processes during perception of different types of co-speech gestures, such as beat, deictic, iconic, and metaphoric co-speech gestures. The purpose of this study was to review the existing literature on the neural correlates of co-speech gesture perception and to test whether different types of co-speech gestures elicit a common pattern of brain activity in the listener. To this purpose, we conducted a meta-analysis of neuroimaging studies, which used different types of co-speech gestures to investigate the perception of multimodal (co-speech gestures) in contrast to unimodal (speech or gestures) stimuli. The results show that co-speech gesture perception consistently engages temporal regions related to auditory and movement perception as well as frontal-parietal regions associated with action understanding. The results of this study suggest that brain regions involved in multisensory processing and action understanding constitute the general core of co-speech gesture perception.

Keywords: co-speech gestures, multisensory perception, meta-analysis, action understanding
**Introduction**

When we speak, we typically gesture. Co-speech gestures (CSGs) are hand movements that accompany speech and allow the speaker to effectively communicate thoughts and ideas in two separate modalities (for a review, see Hostetter, 2011), i.e., linguistic content in the auditory domain and imagistic content in the visual domain. Neuroimaging studies of the neural correlates of CSG perception have found that during the observation of CSG the brain shows increased activity in areas that are involved in auditory and semantic processing of language. Several studies found that the observation of a person who produces gestures while speaking engages a core set of frontal, parietal, and temporal areas including superior temporal gyrus (STG), middle temporal gyrus (MTG), intraparietal sulcus (IPS), as well as inferior frontal gyrus (IFG) when compared to watching a video of a speaker who does not gesture or a speaker producing gestures without sound (e.g., Green et al., 2009; Holle et al., 2008, 2010; Hubbard et al., 2009; Willems et al., 2009).

Evidence shows that the perception of CSGs activates STG when compared to unimodal speech or gesture (Hubbard et al., 2009) especially under increased noise conditions (Holle et al., 2010). In addition, there is evidence that the reduced ability to understand CSGs in younger children is related to the level of activity in left STG (Dick et al., 2012a). These findings suggest that STG forms an important area for matching speech sounds and gesture movements. Furthermore, studies show that left MTG is activated when CSGs relate to the linguistic meaning metaphorically (Kircher et al., 2009; Straube et al., 2011) or when gestures provide additional information that is not present in speech, e.g., when the manner of the movement referred to as “falling” in speech is visually expressed in gesture (Dick et al., 2012b). In contrast, activation in left IPS is found for CSGs that are not related to speech, i.e., CSGs where speech audio and
gesture video from different trials have been combined (Dick et al., 2009; Green et al., 2009), for gestures whose meaning does not match the meaning of the accompanying speech (Willems et al., 2007, 2009), and for CSGs in contrast to self-grooming movements (Holle et al., 2008). Together, these findings suggest that MTG and IPS are crucial for processing the semantic-communicative dimension of CSGs. Finally, evidence shows that activity in left IFG is highly sensitive to the relationship between speech and gestures (Skipper et al., 2007, 2009) and that IFG is active when the relationship between speech and gesture is semantically complex (Dick et al., 2012b; Kircher et al., 2009; Straube et al., 2011; Willems et al., 2007, 2009) or when the communicative intention of hand movements is ambiguous (Dick et al., 2009; Green et al., 2009). All in all, the findings from these studies strongly suggest that during the observation of CSGs, frontal and temporal regions are engaged in semantic processing, whereas frontal and parietal areas are associated with assessing whether the movements are communicatively intended or accidentally produced together with speech.

In addition to fMRI studies, electroencephalography has also been used to investigate CSG perception, especially whether viewing of CSGs has modulatory effects on the N400 (Cornejo et al., 2009; Gunter & Bach, 2004; Habets et al., 2011; Holle & Gunter, 2007; Kelly et al., 2007; Obermaier et al., 2011; Wu & Coulson, 2005, 2007a, 2007b). The N400 is an event-related component with increased negativity that peaks approximately 400 ms after a semantic violation is perceived and whose strength is related to contextual expectations, such that a reduction of the N400 indicates a reduction in semantic violation (Lau et al., 2008). Studies using this approach have found an increased N400 for meaningless vs. meaningful emblem gestures (Gunter & Bach, 2004), incongruent vs. congruent word-gesture pairs (Holle & Gunter, 2007), incongruent vs. congruent metaphorical expression-gesture pairs (Cornejo et al., 2009),
incongruent vs. congruent gesture-cartoon pairs (Wu & Coulson, 2005), unimodal (speech) vs. cross-modal (speech and gesture) related word-probe pairs (Wu & Coulson, 2007a), and unrelated vs. related word-gesture pairs (Wu & Coulson, 2007b). The strength of the N400 for CSGs depends on the temporal coordination of speech and gesture (Habets et al., 2011; Obermaier et al., 2011) and on the listener’s expectations that the gesture is communicatively intended (Kelly et al., 2007). Taken together, these findings suggest that co-speech gestures shape the semantic context in which the meaning of speech is interpreted.

While these studies suggest a common pattern of brain activity related to CSG perception, none of the mentioned studies tested different types of CSGs, which differ considerably from each other. Based on their kinetic characteristics and relationship with concurrent speech, four different subtypes of CSGs have been proposed (McNeill, 1992). First, simple repetitive movements, such as moving the hands briefly up and down without specific hand shapes, have been termed ‘beat gestures’. They are commonly used during public speaking and it has been suggested that they play a role in ordering discourse information (McNeill, 1992). Second, ‘deictic gestures’ are pointing movements that can be performed with a variety of body parts (Kita, 2000). They are the only type of co-speech gestures that can be used without speech and can even replace speech during interaction. Third, ‘iconic gestures’ depict an action, object, movement, or space. They are closely related to the meaning of the concurrent speech and often complement the meaning expressed linguistically, e.g., when the expression “playing ball” is accompanied by a throwing gesture with rounded hand shape. Fourth, ‘metaphoric gestures’ are co-expressive with speech but differ from iconic gestures in that they also express a second related and more abstract meaning, e.g., when the utterance “this idea will take off very soon” is accompanied by an upward moving
gesture. The different CSG types occur with different frequencies in spoken discourse. McNeill (1992) shows that iconic and beat gestures are the two most common types of gestures (of all spoken clauses in his sample, 33% are accompanied by iconic gestures and 34% are accompanied by beat gestures) and that metaphoric and deictic gestures are much less frequent (5% and 4% respectively). Interestingly, McNeill’s (1992) data also show that one is more likely to produce an iconic or beat gesture than no gesture at all (24%).

The differences between the CSG subtypes cast doubt on the assumption that the perception of all CSGs is correlated with the same underlying neural activity. Hence, the objective of this study was to investigate whether the perception of other types of CSG shows a general neural pattern of CSG perception. A meta-analysis of neuroimaging studies was conducted to investigate the neural processes underlying the perception of different CSG types. Contrasts between multimodal co-speech gestures and unimodal speech and unimodal gesture under normal listening conditions were investigated for beat, iconic, and metaphoric CSGs and it was expected that all types of CSGs would show evidence of multisensory perception as indicated by increased activation in posterior superior temporal gyri and auditory cortex as well as evidence of action understanding as indicated by increased activation in posterior parietal and frontal regions.

**Methods**

**Literature selection**

A search of the Pubmed database (http://www.ncbi.nlm.nih.gov/pubmed/) was performed using the keywords ‘gesture’ and ‘fMRI’ and ‘PET’, which yielded 246 published, peer-reviewed articles as of October 8, 2012. Two criteria were used to select
studies for the meta-analysis. First, we only selected studies that used CSGs as defined by McNeill (1992) as stimuli. Second, we only selected studies that reported the results of whole brain analysis and contrasts of CSGs with a baseline condition in a readily available form and that did not involve an experimental manipulation of syntactic, semantic, or phonetic/prosodic aspects of CSGs or the speech they accompany. A large number of the initial 246 studies investigated phenomena that are related to CSGs, but differed in important ways, such as sign language, or emblematic or social gestures without speech. In addition, some studies did not use spontaneous gesture production but a small set of predefined and learned gestures. As a consequence, only 22 out of 246 studies met the first criterion and used CSGs as defined by McNeill (1992) as stimuli. Further, our intention was to compare simple contrasts, such as CSG, speech, and gesture with each other, but several studies did not report basic contrasts as their results. Also, a number of studies used region-of-interest analysis instead of whole brain analysis. Those studies are unsuitable for meta-analysis. Thus, only 6 of the 22 studies met the second criterion and were included in further analysis. While this number is relatively low compared to other meta-analyses, we are confident that the selected studies investigate co-speech gesture perception as opposed to a related phenomenon (see table 1).

In the selected studies, CSGs of three types were used: iconic (3 studies), metaphoric (1), metaphoric and iconic (1), and beat (1). No study investigated deictic gestures. In half of the studies, the stimuli also showed the face of the person performing the gestures. From these studies, peak activations showing simple contrasts of CSGs with control conditions such as fixation, speech with a non-moving body, or silent grooming behaviors were extracted. The six studies included 102 participants and yielded 65 peaks, which were used in the subsequent analysis. The peak voxel
coordinates of studies that were reported in MNI space were converted to Talairach space using the icbm2tal algorithm (Lancaster et al., 2007) included in the Ginger ALE software (Eickhoff et al., 2009).

(INSERT TABLE 1 HERE)

**Activation likelihood estimation**

In order to identify relevant regions of activation, which would show how different types of CSGs converge on a core set of brain areas, an activation likelihood estimation (ALE) analysis was performed (Eickhoff et al., 2009). In the analysis, the 65 peaks reported in the studies are assumed to reflect probability distributions because of the spatial uncertainty of the results, as well as the variance between subjects and between templates used in each study. For each study, all peaks were therefore converted into a three-dimensional Gaussian probability distribution. Its full-width half maximum was calculated for each study based on an estimate of the between-subjects variance scaled by the number of participants (Eickhoff et al., 2009). For each study, a three dimensional activation likelihood estimation (ALE) map was calculated by combining the Gaussians of each peak. Then, the ALE maps of each study were combined voxelwise to form an ALE map that reflects the activity reported in all studies (Eickhoff et al., 2009). To assess whether the combined ALE map reflects true convergence of findings, a null-distribution of ALE maps was generated by selecting random voxels as foci and repeating the procedure used for real foci. The combined ALE map was then compared to the null-distribution to yield the final p-value that a certain voxel is part of one of the reported clusters (Eickhoff et al., 2009). The significance of results was corrected for multiple comparisons at p < 0.01 using false
discovery rate correction (Laird et al., 2005). In addition, only clusters with a minimum size of 100 mm$^3$ were considered for further investigation. All results were manually labelled for their anatomical location based on the MNI coordinates and the MNI labelled atlas included in the Mango software package (http://ric.uthscsa.edu/mango/).

Results

The meta-analysis yielded eight significant clusters for CSG perception showing activity in right auditory cortex, left posterior superior temporal gyrus, left inferior parietal lobule, right fusiform gyrus, left geniculate thalamic nucleus, and bilateral ventral pre-motor cortex (see Fig 1 and Table 2).

(DISPLAY FIGURE 1 HERE)

(DISPLAY TABLE 2 HERE)

Discussion

The purpose of this study was to investigate the multisensory perception of speech accompanying gestures, i.e., whether all types of such co-speech gestures (CSG) show a common pattern of neural activity. The results of the ALE meta-analysis of three different CSG types showed a general activity pattern related to multisensory integration. Significant activation was found in right primary and secondary auditory cortex during CSG perception. It has been proposed that the right auditory cortex is sampling spectral auditory information with a higher acuity but at the expense of a slower sampling rate (Griffiths & Warren, 2002; Schönwiesner et al., 2006; Poeppel, 2003; Zatorre & Gandour, 2008). Certain acoustic properties of the speech stream
(especially pitch) that form prosody and contribute to pragmatic and para-linguistic functions occur on a similar time-scale. Hence, it is plausible that the pitch contour of an utterance is processed by right auditory cortex. Consequently, we interpret the increased activation of right auditory cortex as evidence that CSGs are processed in conjunction with prosodic features of speech. Evidence for the engagement of right auditory cortex during CSG perception comes from one neuroimaging study, which investigates the perception of speech and beat gestures (Hubbard et al., 2009). Hubbard et al. (2009) found that speech accompanied by beat gestures shows increased activation in bilateral auditory cortex and the right planum temporale when compared to speech without beat gestures. The results of Hubbard et al.’s study (2009) show that beat gestures are processed in a multisensory manner; however, beat gestures constitute only one type of CSG and the relation between beat gesture movements and rhythmic aspects of speech has been seen as one of their defining features (McClave, 1994; McNeill, 1992). Evidence further suggests that the production of beat gestures affects the acoustic properties of pitch accent in speech (Krahmer & Swerts, 2007), which might account at least in part for the findings of Hubbard et al. (2009). However, additional behavioral research supporting the view of a strong integration of gestures with speech prosody has shown that CSGs are engaged in disambiguation of syntactic structures and reference resolution, i.e., that listeners use visual cues from the speaker’s gestures to infer to whom or what the speaker is referring (Holle et al., 2012). In addition, recent evidence shows that under listening conditions where prosody is intelligible but individual speech sounds are not, the alignment of prosody with a speaker’s movement of potential referents is used by listeners to detect the speaker’s object of reference (Jesse & Johnson, 2012). Also, evidence from studies measuring event-related potentials show
that semantic processing of speech is sensitive to the temporal alignment of speech and gestures (Habets et al., 2011; Obermaier et al., 2011).

Together, these findings suggest that CSGs might be used by listeners to process language in a multisensory manner akin to visual speech. Visual speech describes the listener’s use of visual information from the speaker’s lip movements during the perception of speech sounds. It has been shown that the visual presentation of lip movements of a speaker producing /ba/ can alter the auditory perception of a concurrently presented speech sound so that /pa/ is perceived as /ta/ (McGurk & McDonald, 1976). However, in contrast to visual speech, the movements of CSGs do not relate to individual sounds but span one or more words and are aligned with speech prosody, which is known to affect how speech is interpreted on a discourse level (Dahan et al., 2002). As such, they might not affect the perception of individual speech sounds but the interpretation of whole words and phrases.

Our results further showed that CSGs engage left posterior superior temporal gyrus, which is essential for biological motion processing (Pelphrey et al., 2005; Pavlova, 2012) and speech perception (Buchsbaum et al., 2001; Hickok et al., 2000), forming a part of the dorsal speech processing pathway responsible for auditory-motor integration (Friederici, 2011). In addition, left posterior superior temporal gyrus is engaged in combining auditory and visual information during visual speech (Calvert & Campbell, 2003; van Wassenhove et al., 2005). The left STG is generally considered the highest cortical region in the perceptual hierarchy engaged in multisensory perception, where information from unimodal areas, such as primary visual, auditory, and somatosensory cortices, is integrated along cortical and thalamocortical pathways and further modulated by fronto-parietal attention networks (Man et al., 2013; Talsma et al., 2010; for an extended review, see Klemen & Chambers, 2012). In addition, increased
activity in right fusiform gyrus has been found during body movement perception (Peelen & Downing, 2005; Downing & Peelen, 2011) and the integration of body parts into whole body percepts (Taylor, Wiggett, & Downing, 2007), which suggests that this region might be engaged in hand movement perception. Together, these findings seem to suggest that during the perception of CSGs visual information about hand movements from right fusiform gyrus and auditory information about speech prosody from right auditory cortex might be integrated in left STG. Consequently, our results show that multisensory perception is a general feature of multimodal communication involving CSGs, and provide further evidence for the view that CSGs are not perceived as a body language separate from speech but are closely linked to spoken language (see McNeill, 1985).

In addition to audio-visual processing, the ALE meta-analysis also showed that CSG perception activates areas associated with action understanding and imitation, such as left inferior parietal lobule and bilateral ventral premotor cortex (deLange et al., 2008; Iacobini et al., 1999). The activity of regions associated with action understanding is not surprising given that CSGs add hand movement to speech. However, the kinetic integration of speech and gesture is known to be part of a more general behavioral integration that also includes facial expression, eye gaze, and posture (Birdwhistell, 1970; Condon & Ogston, 1967; for a recent review, see Wagner, 2014). Also, when people are engaged in communicative interactions, they tend to imitate each other’s behaviors. These imitations are an integration of language (lexical choice as well as speech intonation), paraverbal communication (gestures, posture, facial expression), and non-verbal behaviors (eye gaze, self grooming, or, object-directed actions; Kimbara, 2006, 2008; Louwerse et al., 2012), which has been associated with social perception (Pentland, 2008). This integration of speech and gestures on the kinetic level might
therefore have a wider social meaning and extends beyond a contribution to speech perception alone.

Conclusions

In sum, the results showed that during perception of different types of CSGs, areas related to multimodal processing are activated. Activity in right auditory cortex and left posterior superior temporal reflects the multisensory integration of speech sounds and gesture movements. The engagement of right auditory cortex in particular suggests that this integration might occur on the utterance rather than word level. In addition, activity in areas, such as ventral premotor and inferior parietal cortices, shows the engagement of action understanding and reflects the perception of CSGs as intentional communicative movements. In conclusion, our results provide evidence that brain regions involved in audiovisual integration and action understanding constitute the general core of CSG perception.
References


Figure Captions

Figure 1: Results of ALE meta-analysis show significant activation in right auditory cortex, left posterior superior temporal gyrus, left inferior parietal lobule, right fusiform gyrus, left geniculate thalamic nucleus, and bilateral ventral pre-motor cortex for multimodal co-speech gesture perception over unimodal speech or gesture perception.