Current Biology 19, 661–667, April 28, 2009 ©2009 Elsevier Ltd All rights reserved DOI 10.1016/j.cub.2009.02.051

Report

Gestures Orchestrate Brain Networks for Language Understanding

Jeremy I. Skipper,[1,2,*] Susan Goldin-Meadow,[1] Howard C. Nusbaum,[1] and Steven L. Small[1,2]
[1]Department of Psychology
[2]Department of Neurology
The University of Chicago
Chicago, IL 60637
USA

Summary

Although the linguistic structure of speech provides valuable communicative information, nonverbal behaviors can offer additional, often disambiguating cues. In particular, being able to see the face and hand movements of a speaker facilitates language comprehension [1]. But how does the brain derive meaningful information from these movements? Mouth movements provide information about phonological aspects of speech [2–3]. In contrast, cospeech gestures display semantic information relevant to the intended message [4–6]. We show that when language comprehension is accompanied by observable face movements, there is strong functional connectivity between areas of cortex involved in motor planning and production and posterior areas thought to mediate phonological aspects of speech perception. In contrast, language comprehension accompanied by cospeech gestures is associated with tuning of and strong functional connectivity between motor planning and production areas and anterior areas thought to mediate semantic aspects of language comprehension. These areas are not tuned to hand and arm movements that are not meaningful. Results suggest that when gestures accompany speech, the motor system works with language comprehension areas to determine the meaning of those gestures. Results also suggest that the cortical networks underlying language comprehension, rather than being fixed, are dynamically organized by the type of contextual information available to listeners during face-to-face communication.

Results and Discussion

What brain mechanisms account for how the brain extracts phonological information from observed mouth movements and semantic information from cospeech gestures? In prior research, we have shown that brain areas involved in the production of speech sounds are active when listeners observe the mouth movements used to produce those speech sounds [7, 8]. The pattern of activity between these areas, involved in the preparation for and production of speech, and posterior superior temporal areas, involved in phonological aspects of speech perception, led us to suggest that when listening to speech, we actively use our knowledge about how to produce speech to extract phonemic information from the face [1, 8]. Here we extrapolate from these findings to cospeech gestures. Specifically, we hypothesize that when listening to speech accompanied by gestures, we use our knowledge about how to produce hand and arm movements to extract semantic information from the hands. Thus, we hypothesize that, just as motor plans for observed mouth movements have an impact on areas involved in speech perception, motor plans for cospeech gestures should have an impact on areas involved in semantic aspects of language comprehension.

During functional magnetic resonance imaging (fMRI), participants listened to spoken stories without visual input ("No Visual Input" condition) or with a video of the storyteller whose face and arms were visible. In the "Face" condition, the storyteller kept her arms in her lap and produced no hand movements. In the "Gesture" condition, she produced normal communicative deictic, metaphoric, and iconic cospeech gestures known to have a semantic relation to the speech they accompany [5]. These gestures were not codified emblems (e.g., "thumbs-up"), pantomime, or sign language [5, 9]. Finally, in the "Self-Adaptor" condition, the actress produced self-grooming movements (e.g., touching hair, adjusting glasses) with no clear semantic relation to the story. The self-adaptive movements had a similar temporal relation to the stories as the meaningful cospeech gestures and were matched to gestures for overall amount of movement (Movies S1 and S2 available online).

We focused analysis on five regions of interest (ROIs) based on prior research (Figure S1): (1) the superior temporal cortex posterior to primary auditory cortex (STp); (2) the supramarginal gyrus of the inferior parietal lobule (SMG); (3) ventral premotor and primary motor cortex (PMv); (4) dorsal pre- and primary motor cortex (PMd); and (5) superior temporal cortex anterior to primary auditory cortex, extending to the temporal pole (STa) [1, 10]. The first four of these areas have been found to be active not only during action production, but also during action perception [11]. With respect to spoken language, STp and PMv form a "dorsal stream" involved in phonological perception/production and mapping heard and seen mouth movements to articulatory based representations (see above) [1, 7, 8, 12]. Whereas STp is involved in perceiving face movements, SMG is involved in perceiving hand and arm movements [13] and, along with PMv and PMd [11, 14], forms a (another) "dorsal stream" involved in the perception/production of hand and arm movements [15]. In contrast, STa is part of a "ventral stream" involved in mapping sounds to conceptual representations ([1], see [12] for more discussion) during spoken language comprehension; that is, STa is involved in comprehending the meaning of spoken words, sentences, and discourse [16–18]. To confirm that STa was involved in spoken word, sentence, and discourse comprehension in our data, we intersected the activity from all four conditions for each participant. Our rationale was that speech perception and language comprehension are common to all four conditions; activation shared across the conditions should therefore reflect these processes. We found that a large segment of STa and a small segment of STp were bilaterally active for at least 9 of 12 participants (Figure S2).

Just as research using single-cell electrophysiology in visual cortex examines which neurons prefer or are "tuned" to

*Correspondence: jis2013@med.cornell.edu

(i.e., show a maximal firing rate) particular stimulus properties over others [19], we investigated hemodynamic "tuning" to the meaningfulness of hand and arm movements with respect to the spoken stories by using a peak and valley analysis method ([20, 21], [22] for a similar method). In each ROI for each condition, we averaged the entire time course of the hemodynamic response (henceforth signal) for all voxels that were active in at least one of the four conditions, both within and across participants. We delayed the response by 4 s to align the brain's hemodynamic response over each entire story to the coded features of that story (i.e., the cospeech gestures in the Gesture condition and the self-adaptive movements in the Self-Adaptor condition). Next, we found peaks in the resulting averaged signal for each condition by using the second derivative of that signal. Gamma functions (with similarity to the hemodynamic response) of variable centers, widths, and amplitudes were placed at each peak and allowed to vary so that the best fit between the actual signal and the summation of the gamma functions was achieved (R^2s > .97). Half of the full width half maximum (FWHM/2) of a resulting gamma function at a peak determined the search region used to decide whether, for example, an aligned cospeech gesture elicited that peak. Specifically, a particular hand and arm movement (i.e., either a cospeech gesture or self-adaptor movement) was counted as evoking a peak if 2/3rds of that peak contained a hand and arm movement, and was counted as not evoking a peak if less than 1/3rd of that peak contained a hand and arm movement. The distance between the FWHM/2 of two temporally adjacent gamma functions determined which aspects of the stimuli caused a decay or valley in the response. Specifically, a particular hand and arm movement was counted as resulting in a valley if 2/3rds of that valley contained a hand and arm movement, and was counted as not resulting in a valley if less than 1/3rd of that valley contained a hand and arm movement. Regions were considered tuned to cospeech gestures or self-adaptor movements that were represented in peaks but not valleys. Significance was determined by two-way chi-square contingency tables (e.g., gestures versus no-gestures at peaks; gestures versus no-gestures at valleys).

Figure 1 illustrates the peak and valley analysis. Frames from the stories associated with peaks are on the top; frames associated with valleys on the bottom. Note that the speaker's cospeech gestures are found only on top and thus are associated with peaks, not valleys (Figure 1A); in contrast, her self-adaptor movements are found on top and bottom and thus are associated with both peaks and valleys (Figure 1B). For the Gesture condition, this pattern held in regions PMv, PMd, SMG, and STa where peaks in the response corresponded to gesture movements and valleys corresponded to times when the hands were not moving (Figure 2, gray bars; PMv $\chi^2 = 11.1$, $p < .001$, $\Phi = .51$; PMd $\chi^2 = 8.9$, $p < .003$, $\Phi = .40$; SMG $\chi^2 = 4.46$, $p < .035$, $\Phi = .28$; STa $\chi^2 = 19.5$, $p < .0001$, $\Phi = .62$). Peaks in these regions' responses were not simply due to movement of the hands—the meaningless hand movements in the Self-Adaptor condition were as likely to result in valleys as in peaks in the PMv, PMd, SMG, and STa response (Figure 2, striped bars; PMv $\chi^2 = 0.06$, $p = .81$, $\Phi = .04$; PMd $\chi^2 = 0.0$, $p = .96$, $\Phi = .01$; SMG $\chi^2 = .42$, $p < .52$, $\Phi = .10$; STa $\chi^2 = 0.56$, $p = .45$, $\Phi = .12$). STp showed no preference for hand movements, either meaningful cospeech gestures or meaningless self-adaptor movements.

To understand how PMv, PMd, SMG, and STa work together in language comprehension, we analyzed the network interactions among all five regions. The fMRI signals corresponding to the brain's response to each condition from each of the five regions were used in structural equation modeling (SEM) to find the functional connectivity between regions. The output of these models reflects between-area connection weights that represent the statistically significant influence of one area on another, controlling for the influence of other areas in the network [10, 23]. SEMs are subject to uncertainty because a large number of models could potentially fit the data well, and there is no a priori reason to choose one model over another. To address this concern, we solved all of the possible models for each condition [24] and averaged the best-fitting models (i.e., models with a nonsignificant chi-square) by using a Bayesian model averaging approach [10, 25, 26]. Bayesian model averaging has been shown to produce more reliable and stable results and provide better predictive ability than choosing any single model [27].

Bayesian averaging of SEMs resulted in one model for each condition with nine physiologically plausible connections. Results show that the brain dynamically organizes activity patterns in response to the demands and available information of the immediate language task (Figure S3). Specifically, during the Gesture condition, connection weights were stronger between SMG and PMd (bidirectionally), from SMG to STp, from STa to PMv, and from PMd to STa than in any other condition. In contrast, in the Face condition, connection weights were stronger between STp and PMv (bidirectionally) and from PMv to STa. In the Self-Adaptor condition, connection weights were stronger only from STa to PMd, and, in the No Visual Input condition, no connection weights were stronger than any others. Although this is a complex set of results, a pattern emerges that can be more easily understood by considering the mean of the connection weights between STa and pre- and primary motor cortex (i.e., PMv and PMd) and between STp and pre- and primary motor cortex (Figure 3). This representation of the results shows that the statistically strongest connection weights associated with STa and pre- and primary motor cortex correspond to the Gesture condition. In contrast, the statistically strongest connection weights associated with STp and pre- and primary motor cortex correspond to the Face condition. This result is consistent with the peak and valley analysis showing that STp does not respond to hand movements (see Figure 2). Thus, when cospeech gestures are visible, PMd, STa, and SMG work together more strongly to interpret the meanings conveyed by the hand movements, which could enhance interpretation of the speech accompanying the gestures. In contrast, when only facial movements are visible, PMv and STp work together more strongly to simulate face movements relevant to speech production, a simulation that could aid phonological interpretation; this finding is consistent with previous work in our lab showing that PMv and STp are sensitive to the correlation between observable mouth movements and speech sounds [7, 8].

Results suggest that cospeech manual gestures provide information germane to the semantic goal of communication, whereas oral gestures support phonemic disambiguation. This interpretation is supported by two additional pieces of evidence. First, percent of items recalled correctly was 100%, 94%, 88%, and 84% for Gesture, Self-Adaptor, Face, and No Visual Input conditions, respectively, as probed by three true/false questions 20 min after scanning. Participants were significantly more accurate at recalling information in the Gesture condition, which displayed face movements *and*