Integrating Speech and Iconic Gestures in a Stroop-like Task: Evidence for Automatic Processing

Spencer D. Kelly[1], Peter Creigh[2], and James Bartolotti[1]

Abstract

■ Previous research has demonstrated a link between language and action in the brain. The present study investigates the strength of this neural relationship by focusing on a potential interface between the two systems: cospeech iconic gesture. Participants performed a Stroop-like task in which they watched videos of a man and a woman speaking and gesturing about common actions. The videos differed as to whether the gender of the speaker and gesturer was the same or different and whether the content of the speech and gesture was congruent or incongruent. The task was to identify whether a man or a woman produced the spoken portion of the videos while accuracy rates, RTs, and ERPs were recorded to the words. Although not relevant to the task, participants paid attention to the semantic relationship between the speech and the gesture, producing a larger N400 to words accompanied by incongruent versus congruent gestures. In addition, RTs were slower to incongruent versus congruent gesture–speech stimuli, but this effect was greater when the gender of the gesturer and speaker was the same versus different. These results suggest that the integration of gesture and speech during language comprehension is automatic but also under some degree of neurocognitive control. ■

INTRODUCTION

Over the past decade, cognitive neuroscientists have become increasingly interested in the relationship between language and action in the human brain (for a recent review, see Willems & Hagoort, 2007). This interest is fueled by the discovery that language shares neural substrates that are involved in more basic processes such as the production and the perception of action (Nishitani, Schurmann, Amunts, & Hari, 2005; Rizzolatti & Craighero, 2004; Rizzolatti & Arbib, 1998). These findings have led many researchers to view language as an ability that grew out of action systems in our evolutionary past (Armstrong & Wilcox, 2007; Corballis, 2003; Kelly et al., 2002), or as Elizabeth Bates so eloquently put it, "[as] a new machine that nature has constructed out of old parts" (MacWhinney & Bates, 1989; Bates, Benigni, Bretherton, Camaioni, & Volterra, 1979). Taking this embodied view that language is inextricably tied to action in present day communication, the current article investigates the strength of this relationship by focusing on a potential interface between the two systems: cospeech iconic gesture.

Cospeech iconic gestures are spontaneous hand movements that naturally, pervasively, and unconsciously accompany speech and convey visual information about object attributes, spatial relationships, and movements (McNeill, 2005; Kendon, 2004; Goldin-Meadow, 2003; Clark, 1996). These gestures are so tightly tied to speech that theorists have argued that the two together constitute a single integrated system (McNeill, 1992, 2005; Kita & Özyürek, 2003; Özyürek & Kelly, 2007). For example, McNeill (2005) argues that because speech and iconic gesture temporally overlap but convey information in two very different ways—speech is conventionalized and arbitrary, whereas iconic gestures are idiosyncratic and imagistic—the two together capture and reflect different aspects of a unitary underlying cognitive process. For example, imagine someone saying, "They were up late last night," while making a drinking gesture. From this example, it should be clear that the iconic gesture and the accompanying speech combine to reveal meaning about an activity that is not fully captured in one modality alone.

Exploring this relationship between speech and iconic gesture may provide a unique opportunity to better understand the extent to which language and action are connected in the brain. Indeed, if language is inextricably linked to action, as many have argued, one might expect iconic gesture to be inextricably linked to speech. The present study investigates the strength of this neural link by exploring the extent to which the two channels are automatically integrated during language comprehension.

The Integration of Gesture and Speech

The first attempts to address this question of whether gesture and speech form an integrated system came from researchers in psychology, linguistics, and education using behavioral methods. Although some researchers have recently attempted to support this claim by investigating

[1]Colgate University, [2]University of Pittsburgh

processes involved in language production (Goldin-Meadow, 2003; Kita & Özyürek, 2003), the majority of experimental work has focused on language comprehension, showing that people typically pay attention to gesture when processing speech (Beattie & Shovelton, 1999; Cassell, McNeill, & McCullough, 1999; Kelly, Barr, Church, & Lynch, 1999; Kelly & Church, 1998; Goldin-Meadow, Wein, & Chang, 1992). However, these studies have been criticized for not conclusively demonstrating that gesture and speech are truly integrated during language comprehension. For example, researchers such as Robert Krauss have argued that this work on gesture comprehension demonstrates, at best, a trivial relationship between speech and gesture at comprehension (Krauss, 1998; Krauss, Morrel-Samuels, & Colasante, 1991). That is, gesture may be used as "add-on" information to comprehend a communicator's meaning only after the speech has been processed. In other words, in cases where gesture does influence comprehension, the attention to gesture is post hoc and occurs well after the semantic processing of speech. So according to this view, when people do pay attention to gesture during language comprehension, it is an afterthought.

Recently, researchers have turned to techniques in cognitive neuroscience to better understand the relationship between speech and gesture during language comprehension. Many of the initial studies used ERPs to investigate the time course of gesture–speech integration (Bernardis, Salillas, & Caramelli, 2008; Holle & Gunter, 2007; Özyürek, Willems, Kita, & Hagoort, 2007; Wu & Coulson, 2007a, 2007b; Kelly, Kravitz, & Hopkins, 2004). For example, Wu and Coulson (2007b) presented gesture–speech utterances followed by pictures that were related either to gesture and speech or to just the speech alone. When pictures were related to gesture and speech, participants produced a smaller N400 component (the traditional semantic integration effect, as in Kutas & Hillyard, 1980) than when the pictures were related to just the speech. This suggests that the visuospatial aspects of gestures combined with speech to build stronger and more vivid expectations of the pictures than just speech alone. This finding makes sense in light of other ERP research demonstrating that the semantic processing of a word is affected by the presence of cospeech gestures that convey either congruent or incongruent visual information (Kelly et al., 2004).

So where are gesture and speech semantically integrated in the brain? Using functional neuroimaging techniques (mostly fMRI), researchers have discovered that this integration occurs in brain regions implicated in the production and comprehension of human action (Hubbard, Wilson, Callan, & Dapretto, 2009; Holle, Gunter, Rüschemeyer, Hennenlotter, & Iacoboni, 2008; Wilson, Molnar-Szakacs, & Iacoboni, 2008; Montgomery, Isenberg, & Haxby, 2007; Skipper, Goldin-Meadow, Nusbaum, & Small, 2007; Willems, Özyürek, & Hagoort, 2007). For example, Skipper et al. (2007) found that inferior frontal regions processed speech and gesture differently when the gestures had meaningful versus nonmeaningful relationships to speech. Moreover, Willems et al. (2007) showed that gestural information and spoken information are both processed in the same brain regions—inferior frontal areas—during language comprehension. This inferior frontal region is in an interesting site for the integration of gesture and speech because it is traditionally seen as a language region (as in Broca's area), but more recently it has been described as a key component to the human mirror neuron system (Rizzolatti & Craighero, 2004). Other researchers have identified more posterior regions, such as the superior temporal sulcus and the inferior parietal lobule, as sites for the integration of gesture and speech (Holle et al., 2008; Wilson et al., 2008). Again, these areas are noteworthy because they, like the more anterior regions, are independently implicated in language processing and action processing.

This research has made good progress toward showing that gesture and speech are indeed integrated during language comprehension and that this integration occurs in brain regions involved with human action. Still, it does not fully address the critique by Krauss et al. (1991) regarding the extent to which gesture and speech are integrated during comprehension. That is, gesture may be linked to speech, but—to borrow an earlier description—not inextricably linked. One way to address this concern is to determine whether people cannot help but integrate gesture and speech.

Automatic and Controlled Processes

To determine the strength of the gesture–speech relationship, it may be highly revealing to test whether the integration of the two modalities is automatic or controlled. This distinction was first introduced within the field of cognitive psychology (Schneider & Shiffrin, 1977; Posner & Snyder, 1975; but for a more recent perspective, see Bargh & Morsella, 2008). The traditional distinction is that automatic processes are low level, fast, and obligatory, whereas controlled processes are high level, slow, and under intentional guidance. For example, recognizing that someone who reaches for an object must want that object is an automatic process—understanding why they want that object, a controlled process. So when someone is presented with gesture and speech, do they automatically integrate the two modalities or do they more consciously and strategically choose to integrate them?

Researchers have found evidence in other domains that certain nonverbal or paralinguistic behaviors (such as facial expression, body posture, and tone of voice) are automatically processed in the brain (de Gelder, 2006; Winston, Strange, O'Doherty, & Dolan, 2002; Pourtois, de Gelder, Vroomen, Rossion, & Crommelinck, 2000). For example, de Gelder (2006) provides evidence for a subcortical neurobiological detection system (which primarily includes the amygdala) that obligatorily and unconsciously processes emotional information conveyed through facial expressions and body posture. Focusing specifically

on manual actions, there is evidence that noncommunicative hand movements, such as reaches and physical manipulations of objects, also elicit automatic processing (Iacoboni et al., 2005; Blakemore & Decety, 2001). For example, Iacoboni et al. (2005) found that the inferior frontal gyrus and the premotor cortex (regions traditionally associated with the human mirror neuron system) differentiated actions that were intentionally versus unintentionally produced. Interestingly, this effect held even when people were not explicitly instructed to attend to the goals of the actions, a finding that led the authors to conclude that people automatically process the meaning of human action.

There have been a handful of studies that have specifically explored the automatic versus controlled nature of gesture processing (Holle & Gunter, 2007; Kelly, Ward, Creigh, & Bartolotti, 2007; Langton & Bruce, 2000). On the one hand, Langton and Bruce (2000) demonstrated that people cannot help but pay attention to pointing and emblematic gestures (in the absence of speech) even when their task was to attend to other information. On the other hand, there is some evidence that when it comes to processing cospeech gestures, attention to gesture may have elements of a controlled process. For example, Holle and Gunter (2007) showed that the presence of nonmeaningful gestures (e.g., grooming behaviors) modulates the extent to which meaningful gestures set up semantic expectations for speech later in a sentence. In addition, Kelly et al. (2007) found that the integration of gesture and speech was influenced by explicit instructions about the intended relationship between speech and gesture. Specifically, when processing speech in the presence of an incongruent versus congruent gesture, there was a larger N400 effect in bilateral frontal sites when people were told that the gesture and speech belonged together, but this N400 effect was not preset in the left hemisphere region when people were told that gesture and speech did not belong together. In other words, people appeared to have some control over whether they integrated gesture and speech, at least under circumstances of explicit instructions about whether to integrate the two modalities.

This finding from Kelly et al. (2007) is provocative because it suggests that high-level information such as knowledge about communicative intent modulates the neural integration of gesture and speech during language comprehension. However, the study leaves open a number of unresolved issues. First, because it used very direct and heavy-handed (as it were) instructions not to integrate gesture and speech, one wonders what might happen in the absence of explicit task demands. Indeed, if people strategically choose (on their own) not to integrate gesture and speech, it would allow for much stronger claims about the inherent controlled nature of gesture–speech integration. Second, the task required people to consciously attend to the semantics of the gesture–speech utterances, which could have inherently encouraged controlled processing. A much more powerful test would be not requiring any overt or conscious attention to the semantics of the stimuli.

The third point is more general and requires some explanation. Previous ERP research on the time course of integrating speech and cospeech gestures has conflated iconicity and indexicality (Kelly et al., 2004, 2007).[1] These previous studies used stimuli of people gesturing to objects—for example, gesturing the tallness of a glass and verbally describing either the tallness of that glass (a congruent pair) or the shortness of an adjacent dish (an incongruent pair). The main finding was that incongruent pairs produced a larger N400 component than congruent pairs. However, it is impossible to determine whether the N400 effect arose from the incongruent iconic information of the gesture (i.e., tall vs. short) or whether it arose from the incongruent indexical information (i.e., glass vs. dish). It is one thing to say that indicating one object with a gesture makes it difficult to simultaneously process speech referring to another object, but it is another to say that processing the iconic meaning of a gesture interacts with the simultaneous semantic processing of the accompanying speech. If one wants to make strong and definitive claims about the automatic or controlled integration of speech and cospeech gesture in the brain, it is very important to precisely identify what aspects of gesture—its indexicality or its iconicity—are part of that integrated system.

The Present Study

The present study used improved methods to explore the automatic integration of gesture and speech during language comprehension. First, to address the issue of "iconicity versus indexicality," we used gestures that conveyed no indexical information. The gestures in the present study conveyed iconic information about various actions in the absence of actual objects. These types of iconic gestures are very common in natural communication (McNeill, 1992), but their meaning is usually unclear or ambiguous in the absence of accompanying speech (unlike indexical gestures). Therefore, better understanding how they are temporally integrated with co-occurring speech is important for theories of gesture and speech, specifically, and action and language, more generally.

Second, to address the problem of heavy-handed task instructions that require conscious analysis of gesture and speech, we used a Stroop-like paradigm to test the integration of the two modalities. The classic Stroop technique presents people with color words that are written in different colored fonts, and the "Stroop effect" arises when the meaning of the written word influences how quickly and accurately people can name the color of the font (Stroop, 1935). This effect has traditionally been viewed as a hallmark for automatic processing (Logan, 1978; Schneider & Shiffrin, 1977; Posner & Snyder, 1975), but more recent research has shown that certain attentional and contextual factors can modulate the effect (Bessner

& Stolz, 1999; MacLeod, 1991). By using a variant of this Stroop procedure, we avoided the problems of explicitly drawing attention to gesture and speech—attention that may unintentionally encourage conscious and strategic processing of the two modalities.

In our modified version of the classic Stroop paradigm, participants watched videos of a man and a woman gesturing and speaking about common actions. The videos differed as to whether the gender of the gesturer and speaker was the same or different and whether the content of the gesture and speech was congruent or incongruent. The task was to simply identify whether the man or the woman produced the spoken portion of the videos while accuracy rates and RTs (behavioral measures) and ERPs (electrophysiological measure) were recorded to the spoken targets. In this way, participants were required to make very superficial judgments about the acoustic properties of the spoken message, that is, whether a word was produced by a man or a woman. Importantly, the actual content of the gesture and speech was irrelevant.

Using this Stroop-like paradigm, we made two predictions. The rationale of the predictions is that although the relationship between gesture and speech was not relevant to the participant's task (i.e., to identify the gender of the spoken portion of the videos), it nevertheless should affect task performance if the integration of gesture and speech is automatic. On the basis of this logic, the first prediction is that participants should be slower to identify the gender of a speaker when gesture and speech have an incongruent versus congruent relationship. The second prediction is that there will be a larger N400 component (indexing semantic integration difficulties) to words accompanied by incongruent versus congruent gestures.

METHODS

Participants

Thirty-two right-handed (measured by the Edinburgh Handedness Inventory) Caucasian college undergraduates (12 males, 20 females; mean age = 20 years) participated for course credit. All participants signed an informed consent approved by the institutional review board. A total of three participants were not analyzed due to excessive artifacts in brain wave data, and one was removed for not following task instructions.

Materials

Participants watched digitized videos of a man and a woman (only the torso was visible) uttering common action verbs while producing iconic gestures also conveying actions. To test our two predictions, the videos were designed in a Stroop-like fashion. Specifically, there was a gender manipulation that varied the relationship between the gesturer and the speaker. In the *gender-same* condition, the gesturer and the speaker were the same person (a man or a woman), but in the *gender-different* condition, the gesturer and the speaker were two different people (e.g., a man gesturing but a woman speaking, and vice versa). The gender-same and gender-different conditions were created by digitally inserting the speech of the man or the woman into the video containing the gestures of the woman or the man.

In addition to the gender relationship, the stimuli varied as to whether the gesture and the speech had a congruent or an incongruent relationship. In the *gesture-congruent* condition, the gesture and the speech were semantically related (e.g., gesturing cut and saying "cut"), and in the *gesture-incongruent* condition, they were semantically unrelated (e.g., gesturing stir and saying "cut"). The congruent gesture stimuli were normed before the experiment ($n = 5$) to ensure that the words and the gestures were indeed semantically related. The norming procedure presented video clips to participants who had to rate (on a scale of 1 to 7) whether the speech and the gesture were related. There were initially 26 gesture–speech pairs, but because two of them had ambiguous gesture–speech relationships, we dropped these two and were left with 24 pairs (that includes one practice pair). The mean congruence rating of the remaining 24 stimuli was 6.61 ($SD = 0.63$). We created the incongruent pairs by inserting the audio portion into gesture videos that conveyed semantically unrelated information. See Appendix A for all the gesture–speech pairs used as stimuli, and refer to Figure 1 for an example of the gender

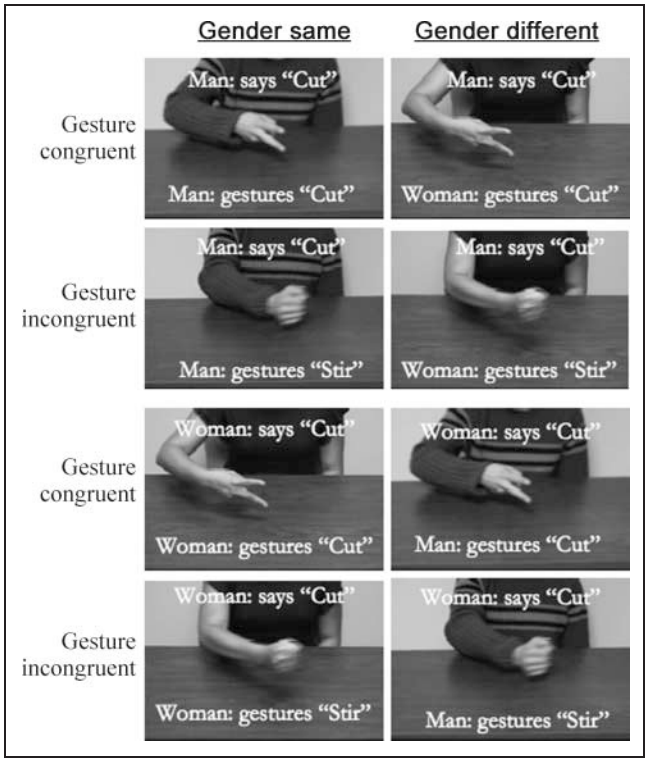

Figure 1. Still frame examples of the congruent and incongruent stimuli for the gender-same and gender-different videos.

and gesture conditions for the "cut" item. The stimuli were designed to create a Stroop-like test that required explicit attention to the gender of the speaker, with the ultimate goal of determining whether people automatically process the semantic relationship between speech and gesture (more below).

Each audiovisual segment was 1 sec in duration. Different from previous research on the integration of speech and cospeech gesture (cf. Kelly et al., 2007), the video began immediately with the stroke portion of the action gesture (i.e., there was no preparation phase as there was in previous research). Following 200 msec after gesture onset, the spoken portion of the video began and lasted for 800 msec. Throughout this 800 msec, the gesture continued. In this way, gesture preceded speech for 200 msec and overlapped with speech for 800 msec. All of the ERP data were time locked to speech onset.[2] The variable ISI was 1.5 to 2.5 sec. The video was 25 min in length.

Procedure

After participants were fitted with the ERP net (see below), the experimenter explained that participants would view videos of a man and a woman making short, one-word spoken utterances. The task was to identify, as quickly as possible, the gender of the verbal portion of the videos (i.e., whether the word was produced in a male or a female voice). A computer recorded these responses, and their latencies were used in the RT analyses. In addition, the computer simultaneously recorded brain wave data to the same responses. Note that this task does not require participants to respond to the semantic content of the speech, nor does it require any attention to gesture. Moreover, the gender of the gesturer is also irrelevant to the task. In this way, the task is a modified Stroop test, with one piece of information relevant to the task—the gender of the speaker—and other pieces of information irrelevant to the task—the semantic relationship between speech and gesture and the physical relationship between the speaker and the gesturer.

This Stroop-like design allowed us to test our two predictions: if gesture and speech are automatically integrated, participants should (1) be slower to identify the gender of a speaker when gesture and speech have an incongruent versus congruent relationship and (2) produce a larger N400 component to words accompanied by incongruent versus congruent gestures.

ERP Setup and Analysis

Participants were fitted with a 128-electrode Geodesic ERP net. The EEG was sampled at 250 Hz using a band-pass filter of 0.1–30 Hz, and impedances were kept below 40 kΩ (the Geonet system uses high-impedance amplifiers). The ERPs were vertex referenced for recording and linked-mastoid referenced for presentation. Following rereferencing, the brain waves were baseline corrected to a 100-msec prestimulus window. Eye artifacts during data collection were monitored with 4 EOG electrodes, with voltage shifts above 70 μV marked as bad (for more on the EOG algorithm, see Miller, Gratton, & Yee, 1988; Gratton, Coles, & Donchin, 1983). Non-EOG channels were marked as bad if there were shifts within the electrode of greater than 200 μV for any single trial. If over 20% of the channels were bad for a trial, the whole trial was rejected. In all, 16% (*SD* = 22%) of the trials were rejected.

The behavioral data were analyzed with a 2 (Gesture, congruent and incongruent) × 2 (Gender, same and different) repeated measures ANOVA with accuracy scores and response latencies as the dependent measures. These behavioral responses were taken to the onset of the verbal potion of the videos.

The ERP data were analyzed with a 2 (Gesture, congruent and incongruent) × 2 (Gender, same and different) × 5 (central, frontal, occipital, parietal, and temporal electrode region) × 2 (left and right) repeated measures ANOVA. The present article focuses on one electrophysiological measure—the N400 component, which indexes semantic integration (Kutas & Hillyard, 1980). The N400 window was created by calculating the average amplitude from 250 to 550 msec (N400) postspeech onset.[3]

The electrode manipulation refers to the location of clusters of electrodes on the scalp. On the basis of previous studies, the 128 electrodes were broken up into five clusters of channels that corresponded roughly to basic anatomical structures of the brain. Refer to Figure 2 for a diagram of the clusters for the 128 electrodes (for more on the clusters, see Kelly et al., 2004). The purpose of these electrode clusters was to test whether the N400 effect was stronger or weaker in particular scalp regions. Specifically, the N400 effect is typically largest in central regions (Kutas & Hillyard, 1980) but is more anterior for stimuli that are comprised of gesture and speech (Kelly et al., 2007; Wu & Coulson, 2007a).

All repeated measures analyses were adjusted for sphericity by using a Greenhouse–Geisser correction (Howell, 2002). Because all contrasts were planned and orthogonal, Student's *t* tests followed up on significant interaction effects.

RESULTS

Behavioral Data

The ANOVA on the accuracy data revealed no significant main effects of Gesture, $F(1,27) = 0.48$, *ns*, or Gender, $F(1,27) = 2.03$, *ns*. In addition, there was not a significant interaction of Gesture × Gender, $F(1,27) = 1.56$, *ns*. These null results are likely due to a ceiling effect in the accuracy scores (refer to Table 1).

The ANOVA on the RT data did not reveal a significant main effect of Gender, $F(1,27) = 0.01$ *ns*, but it did uncover a significant main effect of Gesture, $F(1,27) = 32.60$, $p < .001$. In addition, there was a significant interaction of

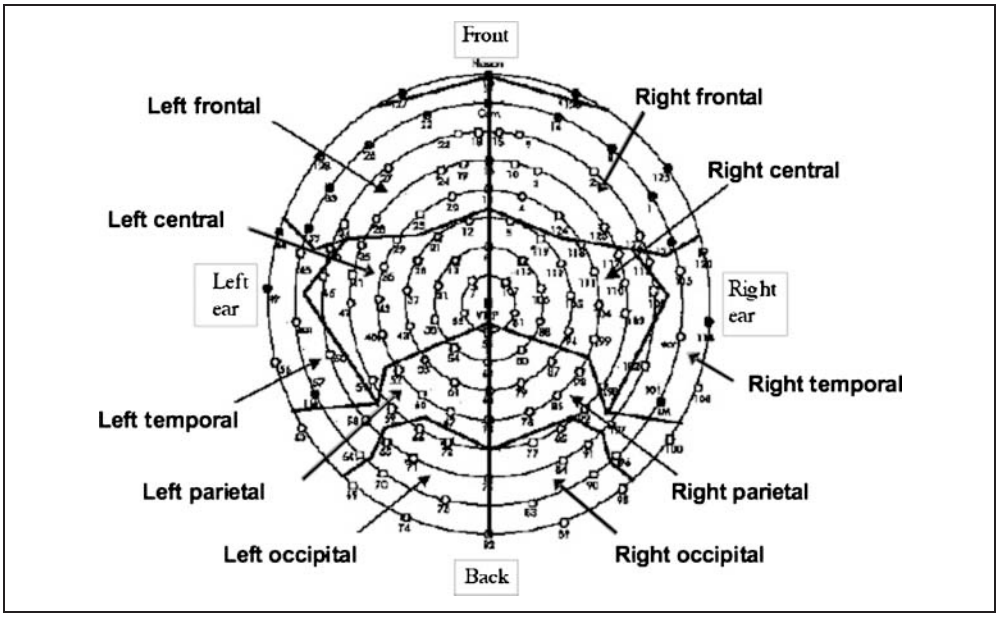

Figure 2. Ten electrode clusters for the 128-electrode geodesic net. For more on the rationale for the clusters, see Kelly et al. (2004).

Gesture × Gender, $F(1,27) = 9.04$, $p = .006$. Simple effects analyses demonstrated that within the gender-same condition, participants were slower when gestures conveyed incongruent information to speech ($M = 818$ msec, $SD = 198$ msec) compared with when it conveyed congruent information ($M = 772$ msec, $SD = 207$ msec), $t(27) = 5.34$, $p < .001$; and within the gender-different condition, participants were also slower—but to a lesser degree—when gesture and speech were incongruent ($M = 805$ msec, $SD = 198$ msec) versus congruent ($M = 786$ msec, $SD = 208$ msec), $t(27) = 3.39$, $p < .001$ (refer to Table 1). Note that although the gesture-incongruent condition was slower than the gesture-congruent condition in both gender conditions, the significant interaction reveals that this difference was greater when the gender of the gesturer and speaker was the same versus different.

These results confirm our first prediction: Participants were slower to identify the gender of the speaker when gesture and speech had an incongruent versus congruent relationship, even when the relationship between the two modalities was not relevant to the task.

Electrophysiological Data

The ANOVA on the N400 time window revealed two main effects. Although not part of the predictions of the present study, there was a significant main effect of Gender, $F(1,27) = 4.76$, $p = .038$, with gender-different stimuli producing a greater negativity than gender-same stimuli.[4] In addition, in accordance with the predictions, there was a significant main effect of Gesture, $F(1,27) = 6.95$, $p = .013$, with gesture-incongruent stimuli producing a more negative going N400 component than gesture-congruent stimuli. Refer to Figure 3 for the main effect of Gesture, Figure 4 for the effects at different electrode sites, and Figure 5 for a topographical map of the N400 effect (peak = 408 msec) across all electrode sites on the scalp. Beyond these main effects, there were no significant two-way interactions for Gesture × Electrode Region, $F(4,108) = 1.23$, *ns*, Gender × Electrode Region, $F(4,108) = 2.62$, *ns*, Gesture × Hemisphere, $F(1,27) = 1.02$, *ns*, Gender × Hemisphere, $F(1,27) = 0.12$, *ns*, or Gesture × Gender, $F(1,27) = 0.01$, *ns*. Nor were there any three- or four-way interactions involving Gesture and Gender variables.

From observation of Figures 3 and 4, it appears that the difference between the gesture-congruent and the gesture-incongruent conditions may be the result of a carry over from differences earlier at the P2 peak.[5] To address this possibility, we subtracted the P2 window (150–250 msec) from the N400 window and reran the ANOVA. Even after using this new baseline, the gesture-incongruent

Table 1. Accuracy Scores and RTs across the Gesture and Gender Conditions

	Mean Proportion Correct (SD)		*Mean RT (SD)*	
	Gesture Congruent	*Gesture Incongruent*	*Gesture Congruent (msec)*	*Gesture Incongruent (msec)*
Gender same	0.99 (0.03)	0.98 (0.03)	772 (207)	818 (208)
Gender different	0.97 (0.03)	0.98 (0.03)	786 (198)	805 (199)

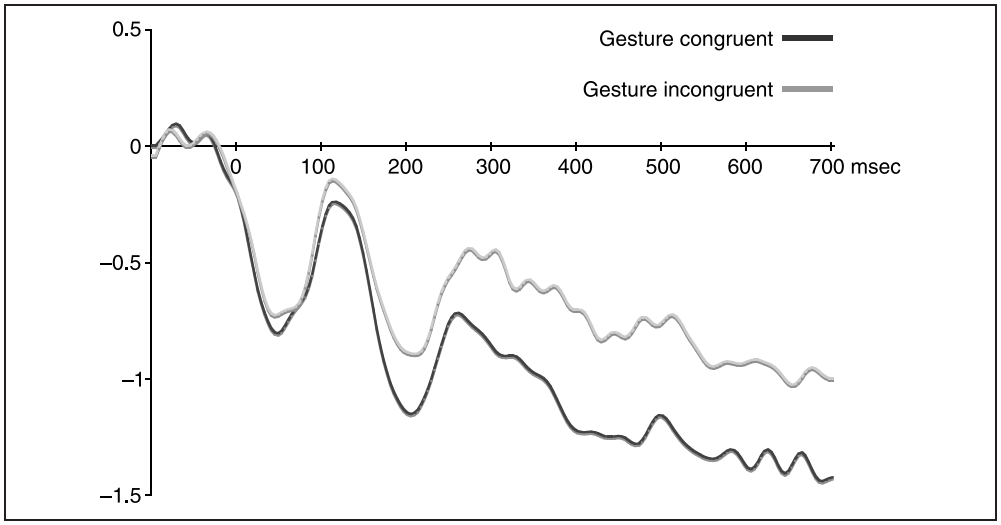

Figure 3. Grand averaged brain waves (collapsed across all electrode regions) for the gesture main effect. The gesture-incongruent condition produced a larger N400 component than the gesture-congruent condition. Microvolts are plotted negative up.

stimuli produced a more negative going N400 component than the gesture-congruent stimuli, $F(1,27) = 4.99$, $p = .034$. Interestingly, this new analysis also yielded a marginally significant interaction of Gesture × Electrode region, $F(4,108) = 2.49$, $p = .103$. This effect was driven by the gesture-incongruent stimuli producing a larger negativity than gesture-congruent stimuli in bilateral Central, $F(1,27) = 5.88$, $p = .022$, and Parietal regions, $F(1,27) = 8.40$, $p = .007$. However, there was not a significant effect in bilateral frontal regions, $F(1,27) = 1.26$, *ns*.

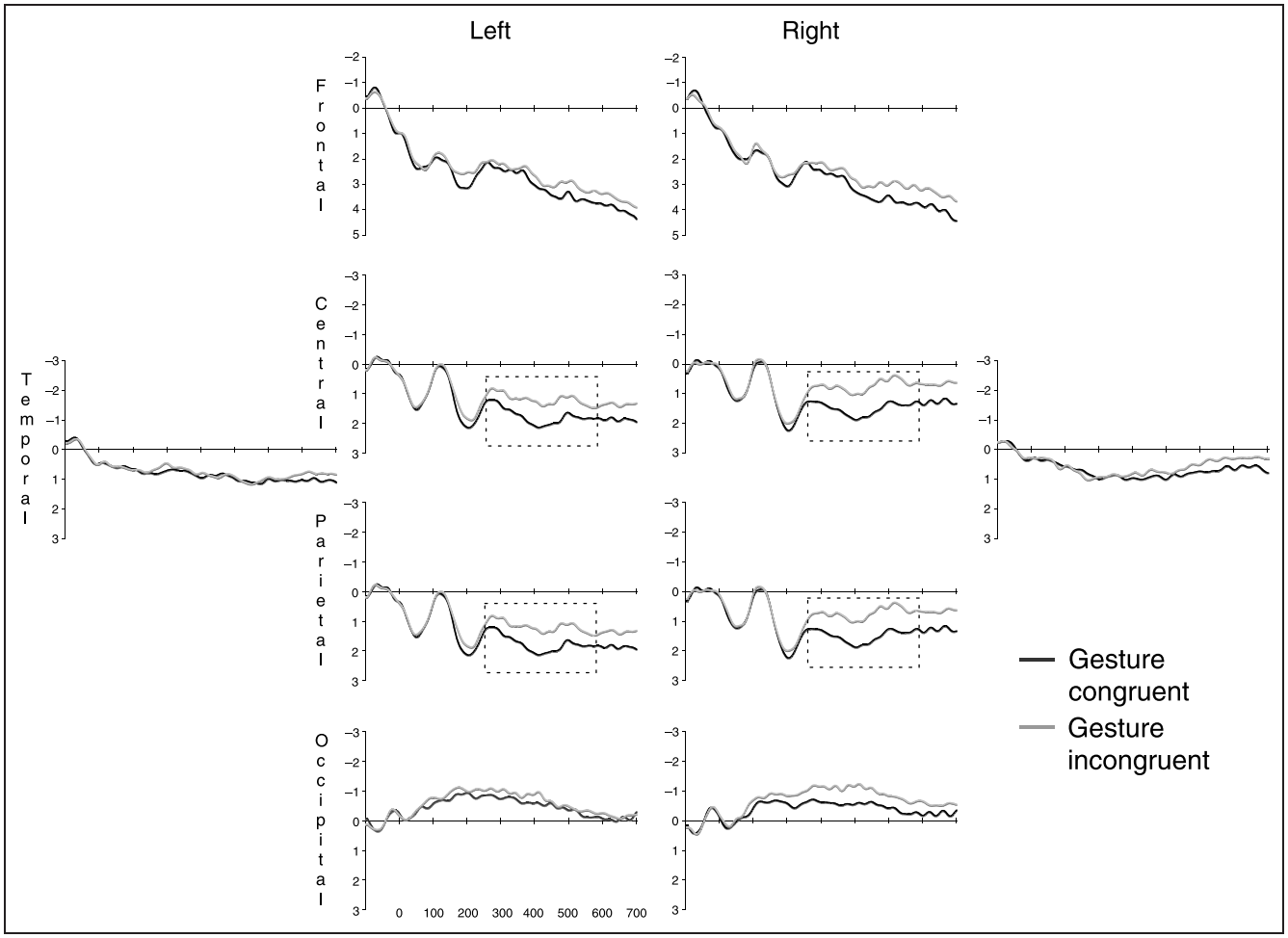

Figure 4. Topographical plots for the 10 electrode regions for the gesture condition. The significant N400 effects are boxed in bilateral central and parietal electrode regions. Microvolts are plotted negative up.

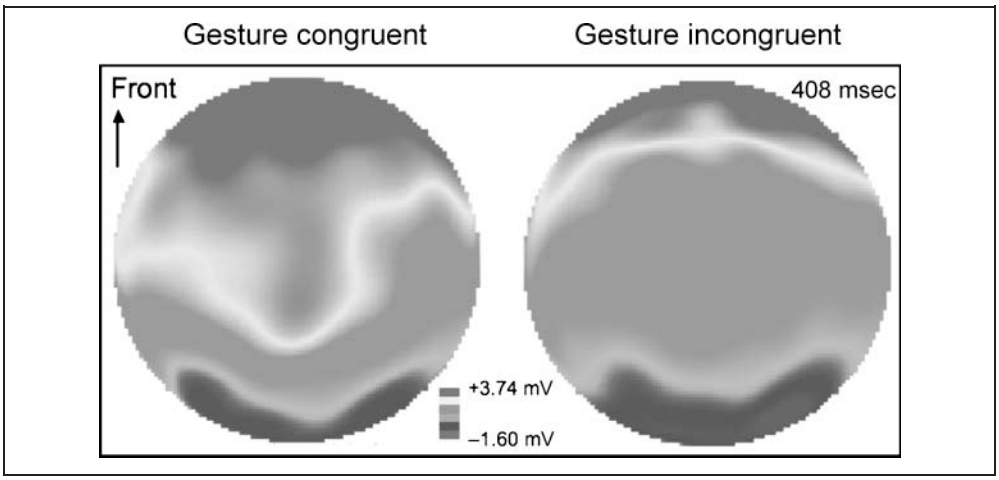

Figure 5. Topographical maps for the N400 peaks (408 msec) for the two gesture conditions. The top of each map is toward the front of the head. Hot colors are positive, and cool colors are negative. The figure can be viewed in color on line.

These results from the electrophysiological measure confirm our second prediction: Although the relationship between gesture and speech was irrelevant to the participants' task, gestures that were semantically incongruent to the content of the speaker's words produced a larger N400 component than gestures that were semantically congruent.

DISCUSSION

The results of the experiment support our two predictions: Participants were slower to process the gender of the speaker's voice when the content of gesture and speech—information not relevant to the task—was incongruent versus congruent (Prediction 1), and participants produced a larger N400 component when gesture and speech were incongruent versus congruent (Prediction 2). These findings are consistent with the claim that gesture and speech are automatically integrated during language comprehension.

Gesture–Speech Integration as an Automatic Process

The traditional hallmarks of automatic processing are that it is low level, fast, and obligatory (Logan, 1978; Schneider & Shiffrin, 1977; Posner & Snyder, 1975). Using a modified Stroop paradigm, we provided definitive evidence for the third aspect of automaticity—the obligatory nature of gesture–speech processing. Our procedure did not require any overt attention to the semantic content of speech and gesture, yet the content of gesture influenced the superficial acoustic analysis of speech in both the behavioral and the electrophysiological measures. That is, when participants simply had to process the auditory differences between a male and a female voice—differences that were extremely salient—they could not do so without also processing irrelevant information conveyed through gesture and speech. This interference resulted in slower RTs and a larger N400 component to incongruent versus congruent gesture–speech stimuli, even when the semantic relationship of gesture and speech had nothing to do with the task at hand.

Regarding the other two hallmarks—low level and fast—our results are less conclusive, but they are at least suggestive of automatic processing. In some recent neuroimaging research on the neural mechanisms of gesture–speech integration, Hubbard et al. (2009) argued that the planum temporale integrates the prosodic elements of speech and gesture (specifically beat gestures that convey no iconic content), suggesting that these types of gesture influence low-level acoustic analysis of spoken input. More generally, Iacoboni et al. (2005) have argued that low-level and phylogenetically old neural mechanisms (e.g., the mirror neuron system, which overlaps with the planum temporale) are designed to automatically process manual hand actions. Considered in this context, one can speculate that iconic hand gestures may interact with low-level, perceptual aspects of speech processing in a similarly automatic fashion.

When it comes to the speed of gesture–speech integration, the present results are consistent with other ERP studies on gesture–speech processing that have identified early semantic components (Bernardis et al., 2008; Holle & Gunter, 2007; Özyürek et al., 2007; Wu & Coulson, 2007a, 2007b). Although we did not report ERP effects earlier than the N400 effect, previous research has uncovered evidence for even earlier integration of verbal and nonverbal information (Willems, Özyürek, & Hagoort, 2008; Kelly et al., 2004). In the Kelly study, the presence of hand gesture influenced the P2 component and even earlier sensory stages of speech processing (within the first 150 msec). Although the present article did not explore such early effects, the two studies together suggest that gesture and speech may be automatically integrated very quickly in the brain.

In general, this claim about the automaticity of gesture processing is consistent with theories that the human brain has evolved specialized neural mechanisms to automatically process nonverbal behaviors, such as bodily gestures, facial expressions, and manual actions (de Gelder,

2006; Iacoboni et al., 2005; Blakemore & Decety, 2001). Building on this, if hand gestures served as a foundation for the evolution of spoken language, as some have argued (Corballis, 2003; Bates & Dick, 2002; Kelly et al., 2002; Rizzolatti & Arbib, 1998), it makes sense that these hand movements would be automatically linked to speech processes. Thus, in contrast to views that gesture is a mere afterthought in language processing, gesture appears to be a real contender in communication, one that has an inextricable link with the speech that it naturally and ubiquitously accompanies.

Importantly, the present study extends previous research on the automatic integration of gesture and speech (Kelly et al., 2007). In this previous work, participants were explicitly instructed to attend to the meaning of speech. In contrast, by using a Stroop-like paradigm in the present study, participants were not required to attend to the meaning of speech—or gesture—but did so anyway. More importantly, the present results shed light on the role of iconicity in the integration of gesture and speech. Unlike previous research that conflated indexicality and iconicity, the present results demonstrated that the representational content (i.e., the iconic meaning) of gesture is automatically connected to the representational content of speech. This finding strongly supports theories that speech and iconic gesture—at least iconic action gestures—are integrated on a deep conceptual level (McNeill, 1992, 2005; Kita & Özyürek, 2003). To further explore this relationship, it would be interesting for future research to investigate the automatic integration of speech and iconic nonaction gestures (e.g., gestures about object attributes or spatial relationships) that have an even more ambiguous meaning in the absence of accompanying speech.

Gesture–Speech Integration Is Not an Exclusively Automatic Process

Although we found strong support for the obligatory nature of gesture–speech integration, the RT data suggest that it may not be an exclusively automatic process. Recall that there were larger RT differences between the two gesture–speech pairs in the gender-same versus gender-different condition. This suggests that participants were sensitive to the context within which they processed gesture and speech. That is, when the context encouraged integration of gesture and speech (i.e., when the same person produced the gesture and word), participants were more likely to combine the two modalities than when the context did not encourage integration (i.e., when different people produced the gesture and word). This finding fits with claims that the Stroop effect can be modulated by higher level contextual variables (Bessner & Stolz, 1999).

Moreover, this modulation is consistent with previous research on the controlled nature of gesture–speech integration (Holle & Gunter, 2007; Kelly et al., 2007). For example, Holle and Gunter (2007) showed that the presence of nonmeaningful gestures influences the extent to which meaningful gestures are integrated with speech during sentence processing. Moreover, Kelly et al. (2007) argued that explicit knowledge about a communicator's intent modulates the neural integration of gesture and speech during word comprehension. Similarly, it is possible that on some level, participants in the present study were aware that gesture and speech produced by different people did not "belong" together, and consequently, they integrated the two modalities differently than when gesture and speech did belong together.

This connection to previous research bears further discussion. In the study by Kelly et al. (2007), there were very similar patterns of behavioral data—in both studies, participants integrated gesture and speech to a greater extent when the two modalities were meant to go together. However, unlike that previous research, the present study did not find a comparable pattern with the electrophysiological data. Specifically, in the previous work, the N400 effect for congruent and incongruent gesture–speech pairs was eliminated in left hemisphere regions when participants believed that the gesture and speech were not intentionally coupled (the no-intent condition). However, in the present study, there was a strong bilateral N400 effect for congruent and incongruent gesture–speech pairs even when gesture and speech did not belong together (the gender-different condition). One possible explanation for this difference is that in previous research, participants were explicitly told that gesture and speech did not belong together, whereas in the present study, they had to determine this for themselves. In fact, participants may have assumed that gesture and speech were meant to go together in the present study. After all, people have experience with video media that artificially combine language and action (e.g., as in watching a dubbed movie or documentary footage with a narrator). Perhaps participants applied this sort of previous experience to the stimuli used in the present study, and this is why the ERP data did not correspond to the previous research. In any event, it will be important for future research to determine the conditions under which people naturally integrate gesture and speech and when this integration is modulated by higher level contextual information.

Focusing just on the present study, it is important to address another inconsistency—why did the gender context modulate the behavioral but not the electrophysiological integration of gesture and speech? This sort of inconsistency between the behavioral and the electrophysiological findings is actually consistent with previous ERP research on the N400 effect (Ruz, Madrid, Lupiánez, & Tudela, 2003; Brown & Hagoort, 1993; Holcomb, 1993). For example, Brown and Hagoort (1993) found that incongruent masked primes produced slower RTs to targets than congruent masked primes, but this effect was not carried over in the electrophysiological data—there was no corresponding N400 effect for masked incongruent items. This suggests that behavioral measures can occasionally reveal differences that electrophysiological measures cannot.

Although this inconsistency tempers strong claims that context modulates the integration of gesture and speech, it encourages future work to explore the circumstances under which controlled mechanisms play a definitive role in the processing of gesture and speech during language comprehension.

Local versus Global Integration

Previous research has made a distinction between "local" and "global" integration of gesture and speech (Özyürek et al., 2007). Local integration concerns how a gesture is integrated with a temporally overlapping word (i.e., when a gesture and a word are presented at roughly the same time). In contrast, global integration concerns how a gesture is integrated with speech over larger spans of discourse (e.g., when a gesture precedes a word in a sentence or vice versa). The majority of ERP research on gesture–speech integration has found evidence for global integration (Bernardis et al., 2008; Holle & Gunter, 2007; Özyürek et al., 2007; Wu & Coulson, 2007a, 2007b), but the present study has provided evidence for local integration.

On the surface, this finding is inconsistent with the work by Özyürek et al. (2007) who, using gestures in larger sentence contexts, found evidence for global—but not local—integration. One explanation for this inconsistency is that there were different SOAs (the timing difference between the onset of speech and gesture) when measuring the local integration of gesture and speech in the two studies. Whereas the present study found an N400 effect for incongruent gesture–speech pairs using SOAs of 200 msec, the work by Özyürek et al. had a simultaneous onset of gesture and speech and they found no N400 effect for local gesture–speech incongruencies. One possible explanation for this inconsistency is that local integration occurs only for gesture–speech pairs that are removed from natural discourse (as in the present study). An equally plausible possibility is that gesture and speech are indeed locally integrated in natural discourse, but not when they are simultaneously presented (as in the study by Özyürek et al., 2007). Indeed, gesture slightly precedes speech onset in natural discourse (McNeill, 1992), and presenting them simultaneously may disrupt their natural integration. It will be important for future ERP research to directly compare different gesture–speech SOAs to more thoroughly understand how the timing of gesture and speech affects the neural integration of the two modalities.

Conclusion

By using a variation on the Stroop procedure, we demonstrated that people obligatorily integrate speech and iconic gestures even when that integration is not an explicit requirement. This suggests that the neural processing of speech and iconic gesture is to some extent automatic. However, there is also evidence (at least with the behavioral measure) that people are sensitive to whether speech and gesture "belong" together, suggesting a potential controlled mechanism as well. Although it will be important to build on these findings using more natural stimuli in larger discourse contexts, they serve as a preliminary foundation for the claim that gesture and speech are inextricably linked during language comprehension. And more generally, by viewing gesture as a unique interface between actions and words, this work will hopefully lend a hand to the growing research on the relationship between body and language in the human brain.

APPENDIX A

List of the 23 Congruent and Incongruent Gesture–Speech Pairs

Congruent		*Incongruent*	
Speech	*Gesture*	*Speech*	*Gesture*
Beat	Beat	Beat	Wipe
Brush	Brush	Brush	Shake
Chop	Chop	Chop	Mow
Close	Close	Close	Screw
Cut	Cut	Cut	Stir
Hammer	Hammer	Hammer	Twist
Knock	Knock	Knock	Write
Lift	Lift	Lift	Stab
Mow	Mow	Mow	Turn
Roll	Roll	Roll	Hammer
Saw	Saw	Saw	Type
Screw	Screw	Screw	Cut
Shake	Shake	Shake	Close
Squeeze	Squeeze	Squeeze	Lift
Stab	Stab	Stab	Sweep
Stir	Stir	Stir	Break
Sweep	Sweep	Sweep	Tear
Tear	Tear	Tear	Roll
Turn	Turn	Turn	Chop
Twist	Twist	Twist	Knock
Type	Type	Type	Brush
Wipe	Wipe	Wipe	Squeeze
Write	Write	Write	Saw
		Practice	
Wring	Wring	Wring	Scrub

Acknowledgments

We would like to thank the anonymous reviewers for their helpful suggestions on previous versions of this manuscript. In addition, we are grateful to Ron Crans for technical assistance in software creation. Finally, we thank the Colgate University for providing generous support for undergraduate research.

Reprint requests should be sent to Spencer Kelly, Department of Psychology, Neuroscience Program, Colgate University, 13 Oak Dr., Hamilton, NY 13346, or via e-mail: skelly@colgate.edu.

Notes

1. This, of course, is not to say that these are the only two ERP studies to find an N400 effect using speech and gesture stimuli (see Bernardis et al., 2008; Holle & Gunter, 2007; Özyürek et al., 2007; Wu & Coulson, 2007a, 2007b). However, they are the only two studies to find an N400 effect to speech stimuli that were simultaneously coupled with incongruent versus congruent gestures.
2. Because the gestures preceded speech by such a short interval (200 msec) and because the ERPs were time locked to speech that co-occurred with gesture, the brainwaves had a positive skew, especially in frontal regions. This skew is common in priming studies that employ very short SOAs (e.g., Kiefer & Brendel, 2006), and the positive slow wave most likely reflects the electrophysiological response to speech riding on top of an electrophysiological response to the gesture. Importantly, this positive skew is uniform for all stimuli, so it should not confound any of our treatment conditions.
3. Note that 250 msec is relatively early for the start of the N400 component, but there is precedence in the literature for this early onset (e.g., Van Berkum, van den Brink, Tesink, Kos, & Hagoort, 2008). A post hoc analysis showed that the early negativity was not a separate N300 component (McPherson & Holcomb, 1999).
4. Although not central to the present study, this gender effect is noteworthy because it is related to a recent ERP study demonstrating a large N400 component when people hear utterances that are unusual for a speaker (e.g., a man saying that he wished he looked like Britney Spears; Van Berkum et al., 2008). The present findings suggest that simply seeing the gender of speaker sets up a "semantic" expectation of the gender of the voice it accompanies.
5. We do not explore the P2 effect in the present manuscript. For coverage of the early integration of gesture and speech, see Kelly et al., 2004.

REFERENCES

Armstrong, D. F., & Wilcox, S. E. (2007). *The gestural origin of language*. New York: Oxford University Press.

Bargh, J. A., & Morsella, E. (2008). The unconscious mind. *Perspectives on Psychological Science*, *3*, 73–79.

Bates, E., Benigni, L., Bretherton, I., Camaioni, L., & Volterra, V. (1979). *The emergence of symbols: Cognition and communication in infancy*. New York: Academic Press.

Bates, E., & Dick, F. (2002). Language, gesture, and the developing brain. *Developmental Psychobiology, 40,* 293–310.

Beattie, G., & Shovelton, H. (1999). Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? An experimental investigation. *Semiotica, 123,* 1–30.

Bernardis, P., Salillas, E., & Caramelli, N. (2008). Behavioural and neurophysiological evidence of semantic interaction between iconic gestures and words. *Cognitive Neuropsychology, 25,* 1114–1128.

Bessner, D., & Stolz, J. A. (1999). Unconsciously controlled processing: The Stroop effect reconsidered. *Psychonomic Bulletin & Review, 6,* 99–104.

Blakemore, S. J., & Decety, J. (2001). From the perception of action to the understanding of intention. *Nature Reviews Neuroscience, 2,* 561–567.

Brown, C., & Hagoort, P. (1993). The processing nature of the N400: Evidence from masked priming. *Journal of Cognitive Neuroscience, 5,* 34–44.

Cassell, J., McNeill, D., & McCullough, K. E. (1999). Speech–gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics & Cognition, 7,* 1–34.

Clark, H. H. (1996). *Using language*. Cambridge, England: Cambridge University Press.

Corballis, M. C. (2003). *From hand to mouth: The origins of language*. Princeton, NJ: Princeton University Press.

de Gelder, B. (2006). Towards the neurobiology of emotional body language. *Nature Reviews Neuroscience, 7,* 242–249.

Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think*. Cambridge, MA: Belknap Press.

Goldin-Meadow, S., Wein, D., & Chang, C. (1992). Assessing knowledge through gesture: Using children's hands to read their minds. *Cognition and Instruction, 9,* 201–219.

Gratton, G., Coles, M. G. H., & Donchin, E. (1983). A new method for off-line removal of ocular artifacts. *Electroencephalography and Clinical Neurophysiology, 55,* 468–484.

Holcomb, P. (1993). Semantic priming and stimulus degradation: Implications for the role of the N400 in language processing. *Psychophysiology, 30,* 47–61.

Holle, H., & Gunter, T. C. (2007). The role of iconic gestures in speech disambiguation: ERP evidence. *Journal of Cognitive Neuroscience, 19,* 1175–1192.

Holle, H., Gunter, T., Rüschemeyer, S. A., Hennenlotter, A., & Iacoboni, M. (2008). Neural correlates of the processing of co-speech gestures. *Neuroimage, 39,* 2010–2024.

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Belmont, CA: Duxberry Press.

Hubbard, A. L., Wilson, S. M., Callan, D. E., & Dapretto, M. (2009). Giving speech a hand: Gesture modulates activity in auditory cortex during speech perception. *Human Brain Mapping, 30,* 1028–1037.

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., & Rizzolatti, G. (2005). Grasping intentions of others with one's own mirror neuron system. *PloS Biology, 3,* 529–535.

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,* 577–592.

Kelly, S. D., & Church, R. B. (1998). A comparison between children's and adults' ability to detect children's representational gestures. *Child Development, 69,* 85–93.

Kelly, S. D., Iverson, J., Terranova, J., Niego, J., Hopkins, M., & Goldsmith, L. (2002). Putting language back in the body: Speech and gesture on three timeframes. *Developmental Neuropsychology, 22,* 323–349.

Kelly, S. D., Kravitz, C., & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language, 89,* 253–260.

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,* 222–233.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press.
Kiefer, M., & Brendel, D. (2006). Attentional modulation of unconscious "automatic" processes: Evidence from event-related potentials in a masked priming paradigm. *Journal of Cognitive Neuroscience, 18,* 184–198.
Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language, 48,* 16–32.
Krauss, R. M. (1998). Why do we gesture when we speak? *Current Directions in Psychological Science, 7,* 54–59.
Krauss, R. M., Morrel-Samuels, P., & Colasante, C. (1991). Do conversational hand gestures communicate? *Journal of Personality and Social Psychology, 61,* 743–754.
Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207,* 203–204.
Langton, S. R. H., & Bruce, V. (2000). You must see the point: Automatic processing of cues to the direction of social attention. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 747–757.
Logan, G. D. (1978). Attention in character classification: Evidence for the automaticity of component stages. *Journal of Experimental Psychology: General, 107,* 32–63.
MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109,* 163–203.
MacWhinney, B., & Bates, E. (1989). *The crosslinguistic study of sentence processing*. New York: Cambridge University Press.
McNeill, D. (1992). *Hand and mind: What gesture reveals about thoughts*. Chicago: University of Chicago Press.
McNeill, D. (2005). *Gesture and thought*. Chicago: University of Chicago Press.
McPherson, W. B., & Holcomb, P. J. (1999). An electrophysiological investigation of semantic priming with pictures of real objects. *Psychophysiology, 36,* 53–65.
Miller, G. A., Gratton, G., & Yee, C. M. (1988). Generalized implementation of an eye movement correction procedure. *Psychophysiology, 25,* 241–243.
Montgomery, K. J., Isenberg, N., & Haxby, J. V. (2007). Communicative hand gestures and object-directed hand movements activated the mirror neuron system. *Social Cognitive & Affective Neuroscience, 2,* 114–122.
Nishitani, N., Schurmann, M., Amunts, K., & Hari, R. (2005). Broca's region: From action to language. *Physiology, 20,* 60–69.
Özyürek, A., & Kelly, S. D. (2007). Gesture, brain, and language. *Brain and Language, 101,* 181–184.
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,* 605–616.
Posner, M. I., & Snyder, C. R. R. (1975). Attention and cognitive control. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola symposium* (pp. 55–85). Hillsdale, NJ: Lawrence Erlbaum.
Pourtois, G., de Gelder, B., Vroomen, J., Rossion, B., & Crommelinck, M. (2000). The time-course of intermodal binding between hearing and seeing affective information. *NeuroReport, 11,* 1329–1333.
Rizzolatti, G., & Arbib, M. A. (1998). Language within our grasp. *Trends in Neurosciences, 21,* 188–194.
Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27,* 169–192.
Ruz, M., Madrid, E., Lupiánez, J., & Tudela, P. (2003). High density ERP indices of conscious and unconscious semantic priming. *Cognitive Brain Research, 17,* 719–731.
Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: Detection, search, and attention. *Psychological Review, 84,* 1–66.
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,* 260–277.
Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18,* 643–662.
Van Berkum, J. J. J., van den Brink, D., Tesink, C., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience, 20,* 580–591.
Willems, R. M., & Hagoort, P. (2007). Neural evidence for the interplay between language, gesture, and action: A review. *Brain and Language, 101,* 278–289.
Willems, R. M., Özyürek, A., & Hagoort, P. (2007). When language meets action: The neural integration of gesture and speech. *Cerebral Cortex, 17,* 2322–2333.
Willems, R. M., Özyürek, A., & Hagoort, P. (2008). Seeing and hearing meaning: ERP and fMRI evidence of word versus picture integration into a sentence context. *Journal of Cognitive Neuroscience, 20,* 1235–1249.
Wilson, S. M., Molnar-Szakacs, I., & Iacoboni, M. (2008). Beyond superior temporal cortex: Intersubject correlations in narrative speech comprehension. *Cerebral Cortex, 18,* 230–242.
Winston, J. S., Strange, B. A., O'Doherty, J., & Dolan, R. J. (2002). Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nature Neuroscience, 5,* 277–283.
Wu, Y. C., & Coulson, S. (2007a). Iconic gestures prime related concepts: An ERP study. *Psychonomic Bulletin & Review, 14,* 57–63.
Wu, Y. C., & Coulson, S. (2007b). How iconic gestures enhance communication: An ERP study. *Brain and Language, 101,* 234–245.