Published in final edited form as:

Neuroreport. 2010 May 12; 21(7): 522–526. doi:10.1097/WNR.0b013e32833904bb.

Gestures modulate speech processing early in utterances

Ying Choon Wu^{a,b} and Seana Coulson^c

^a Center for Research in Language, UC San Diego

^b Swartz Center for Computational Neuroscience, UC San Diego

^c Department of Cognitive Science, UC San Diego, La Jolla, California, USA

Abstract

Electroencephalogram was recorded as healthy adults viewed short videos of spontaneous discourse in which a speaker used depictive gestures to complement information expressed through speech. Event-related potentials were computed time locked to content words in the speech stream and to subsequent related and unrelated picture probes. Gestures modulated event-related potentials to content words co-timed with the first gesture in a discourse segment, relative to the same words presented with static freeze frames of the speaker. Effects were observed 200–550 ms after speech onset, a time interval associated with semantic processing. Gestures also increased sensitivity to picture probe relatedness. Effects of gestures on picture probe and spoken word analysis were inversely correlated, suggesting that gestures differentially impact verbal and image-based processes.

Keywords

auditory comprehension; audio-visual processing; discourse; gesture; multisensory integration; semantics; speech

Introduction

Recent research has shown that spontaneous gestures during conversation can affect language processing in listeners. Hubbard *et al.* [1] found greater blood oxygen level-dependent activation in brain regions implicated in speech perception, such as higher-order auditory cortex (left STG/S) and right planum temporale when listeners were presented with speech accompanied by rhythmic, nonsemantic gestures (so-called beats) versus nonsense movements. Semantic gestures (iconics) have also been shown to affect the comprehension of speech [2–4]. Less is known, however, about the time course of speech–gesture integration in the real-time processing of multimodal discourse.

This issue was addressed in this study by recording EEG as healthy adults viewed short video segments of spontaneous discourse with iconic gestures. Each discourse segment was followed by either a related or unrelated picture probe. Event-related potentials (ERPs) were computed time-locked to the onset of all content words throughout the audio stream and to picture probes.

Research on written language comprehension suggests that as a sentence progresses, and additional contextual information becomes available, the meaning of subsequent content words is easier to understand [5]. To assess whether a similar facilitative effect occurs for words in multimodal discourse, ERP effects were compared between content words occurring during

the first gesture produced in an utterance and those occurring during subsequent gestures or after gesturing was completed. It is possible that words co-timed with the first gesture in an utterance may be impacted differently from words that follow because more information is available to the listener after gesturing is fully completed. Alternatively, progressive gesturing may not yield incremental benefit.

To explore the relationship between the online and downstream effects of gestures on comprehension, ERPs were also measured to picture probes. Given the extensive behavioral research showing that co-speech gestures benefit comprehension, one might expect a positive correlation between these effects – that is, individuals who exhibit the greatest sensitivity to gestures during discourse processing also exhibit the greatest sensitivity to the semantic relationship between subsequent pictures and preceding discourse. Alternatively, gestures may modulate online and downstream comprehension in different ways, with individuals varying in their degree of responsivity to words and pictures.

Method

Twenty healthy, right-handed adults (Edinburgh Inventory laterality quotient=0.78 [6]; mean age=20.51m, SD=3.5; 10 male) received academic credit for their participation.

Materials and procedure

One hundred and seventy-five discourse primes (168 experimental, 7 practice) were derived from segments of continuous video-recorded conversation in which a speaker described everyday experiences to an off-camera interlocutor (see Supplemental Materials, Supplemental Digital Contents 1, 2 and 3, <http://links.lww.com/WNR/A40>, <http://links.lww.com/WNR/A41>, and <http://links.lww.com/WNR/A42>). On an average, the speaker produced 1.7 (SD=0.77) gestures per trial. Roughly half of the trials contained only one gesture; additional gestures were used primarily to elaborate information established by the first one. The speaker did not know the experimental motivation for the filming and was not explicitly instructed to gesture. Counterpart controls were constructed by extracting a freeze-frame from a gesture-free portion of each video prime. In the speech component of these statics, control primes remained identical to the speech in the video-based primes.

The onset time of each content word was determined by visually inspecting the speech signal's spectrogram. Furthermore, content words were classified by two independent coders according to their temporal alignment with the gesture stream. First-gesture words were uttered simultaneously with any phase of the first gesture in a discourse segment. Subsequent-gesture words were produced during subsequent gestures, or rarely, immediately after gesturing was finished. Classification of first and subsequent-gesture words was generally consistent between raters, as indicated by the high degree of correlation between their responses [$r(326)=0.8$, $P<0.0001$].

On an average, participants were presented with 211 first-gesture words and 114 subsequent-gesture words. For analysis purposes, a second dataset containing approximately 100 items of each word type was created. As measured by the London-Lund Corpus of English Conversation [7], the mean frequency was 41 (SD=76) for first-gesture words, and 48 (SD=98) for subsequent-gesture words. Mean length was 5.4 (SD=1.7) and 5.5 (SD=1.8) letters for first and subsequent-gesture words, respectively.

Both gesture and static primes were followed by related or unrelated picture probes, yielding a 2x2 factorial design with two levels of relatedness and two levels of dynamicity (Fig. 1). Related picture probes matched information in both the speech and all gestures in a preceding prime. Each of the four randomized lists contained 176 trials. No video or probe was repeated

within a list. Across lists, however, each picture appeared once as a related and once as an unrelated stimulus, in this way serving as its own control.

Each trial began with a topic title (e.g. pet bird). Next, a discourse prime was presented, lasting from 2.6 to 7.6s. After a 250-ms pause, a picture probe appeared for 500ms, followed by the recognition item (also 500ms). Subsequently, a question mark appeared, prompting participants to respond with a button press (See Ref. [2] for additional methods information).

Participants were asked to read each title silently and to fixate on a small cross in the center of the screen while watching and listening to the speaker. At the prompt, they classified as old or new one of these three types of recognition items: a written word, a video frame, or a picture. Old items were those that had just been encountered in the preceding prime–probe pair. This task encouraged participants to attend to both visual and spoken aspects of stimuli. Responses were registered with a button box, and response hand was counterbalanced across participants.

EEG recording and analysis

EEG was recorded at 26 standard International 10–20 sites [8] in an electromagnetically shielded chamber (see Ref. [2] for electrode montage). EEG was amplified and digitized online at 250Hz (band pass filtered, 0.01–40Hz), referenced to the left mastoid.

Trials contaminated by artifacts such as blinks, eye movements, and drift were rejected off-line. Before averaging, each trial was baseline corrected by subtracting the mean amplitude of the prestimulus interval (–100ms to 0) from each point in the epoch (–100 to 920ms) at each electrode site. On an average, critical bins contained 35 trials (37 median). The mean artifact rejection rate was 14% (SD=12%).

The mean amplitude of ERPs for each participant was measured from 200 to 550ms after the onset of content words, and from 350 to 550ms after the onset of picture probes. (Analyses of mean ERPs from 550 to 850 are described in Supplemental results, Supplemental digital content 4, <http://links.lww.com/WNR/A43>). Measurements underwent a 2×2 repeated-measures analysis of variance (ANOVA) with the factors of dynamicity (full videos with speech and gesture versus static controls with speech only) and gesture sequence (words spoken during the first gesture of an utterance versus concurrently with subsequent gestures). To rule out the possibility that differences in ERPs to first-gesture and subsequent-gesture words could be driven by the imbalanced quantities of trials in these two conditions, all analyses were performed on the balanced dataset, which contained approximately 100 trials of each type. To investigate the scalp distribution of ERP effects, an additional factor of electrode site (26 levels – corresponding to the 26 electrode channels) was included in the omnibus ANOVA. Original degrees of freedom are reported; however, where appropriate, *P* values reflect Geisser–Greenhouse correction [9].

Results

Behavior

On an average, irrespective of the item type, participants accurately classified 93% (SD=0.08) of task items. A one-way repeated-measures ANOVA did not reveal any reliable differences between accuracy rates in classifying the three types of task items [$F(<1.5)$, NS]. Overall, the high accuracy rate suggests that participants consistently attended to both the spoken and the visual aspects of discourse primes, as well as to the picture probes.

Content words

Figure 2 shows ERPs time-locked to the onset of spoken content words accompanied either by first or by subsequent gestures or by static video frames. For first-gesture words, both static and dynamic trials yielded a negative-going onset potential peaking around 75ms poststimulus onset. Subsequently, from approximately 200ms poststimulus to the epoch end, content words elicited less negative ERPs when gestures were visible than when the same discourse occurred without gestures. Over the back of the head, a negativity peaking around 500ms poststimulus can be seen. The otherwise relatively homogenous morphology of the data may be because of the fact that the time-locking events (i.e. content words) were embedded in continuous discourse.

As indicated by the gesture stream×dynamicity×electrode site interactions [full data set: $F(25,475)=3$, $P<0.05$; balanced data set: $F(25,475)=3.47$, $P<0.01$], the presence or absence of gestures affected speech comprehension in different ways over regions of the scalp, depending on whether the spoken item occurred with a first or subsequent gesture. Follow-up tests revealed that on average, content words co-occurring with the first gesture in an utterance reliably elicited less negative ERPs than the same words presented without accompanying gestures [full dataset: dynamicity×electrode site interaction, $F(25,475)=5$, $P<0.005$; balanced data set: dynamicity×electrode site interaction, $F(25,475)=5.34$, $P<0.001$]. The effect of gestures was largest over central and centroparietal midline and medial channels. In contrast, brain response to content words produced with later gestures was not modulated by the presence or absence of gesture information (dynamicity main effect, and dynamicity×electrode site interaction, F values <1 , NS).

Picture probes

The magnitude of the picture probe relatedness effect differed depending on the presence or absence of gesture in the prime {relatedness×dynamicity×electrode site interaction [$F(25,475)=8$, $P<0.001$]}. When preceded by either a gesture or a static prime, related picture probes elicited reliably less N400 than unrelated ones, with maximal effects over frontocentral midline and medial electrode sites. This effect was larger over right hemisphere electrodes relative to left ones for both types of primes {relatedness×dynamicity×hemisphere interaction [$F(1,19)=26$, $P<0.0001$]}. However, in the case of gesture primes, the relatedness effect was larger than that obtained in response to static ones, especially over right hemisphere electrodes (Fig. 3a).

The effects of gesture dynamicity on content word processing and picture probe comprehension were inversely correlated [$r(18)=-0.433$, $P<0.05$]. In individuals for whom picture probes tended to elicit larger relatedness effects when gestures were present than absent (reflecting priming by gestures), ERP effects in response to speech concurrent with first gestures tended to be smaller (Fig. 3b), and *vice versa*.

Discussion

Iconic gestures modulated how accompanying speech was processed. When gesture information was available, content words elicited less negative ERPs relative to when gestures were absent. The time course and scalp distribution of this effect are consistent with those of the classic N400 effect, in which the amplitude of a negative component peaking around 400ms poststimulus is inversely correlated with the degree of fit between a meaningful stimulus and its context. In keeping with earlier brain wave studies reporting N400-like effects in response to spoken words [10,11], the effect of gestures on content word processing was detected by 200ms poststimulus onset in this study. As the speaker's iconic gestures presumably provided contextually relevant information, it is likely that words accompanied by gestures were comprehended and integrated more readily than words presented with static frames.

To our knowledge, this study is the first to report the impact of spontaneously produced iconic gestures on real-time processing of speech in videos of naturally produced discourse. The present finding replicates and extends several recent reports by investigators using less naturalistic materials that spoken words exhibit reduced N400 when coupled with congruent iconic gestures relative to incongruent ones [3,4,12–14].

A key finding in our study was that the presence of iconic gestures impacted the brain response to words occurring early in a discourse segment, but had no detectable effect on ERPs to words co-timed with later gestures. The latter (null) finding suggests that it is not the mere presence of motion in the videos that drives the effect on words produced early in the discourse stream. In fact, the difference between words occurring early versus late may derive from the information content of the accompanying gestures, as the later gestures often simply repeated or elaborated the information presented in first gestures.

Another novel result involves the relationship between speech–gesture integration and downstream comprehension as assessed by ERPs to picture probes. N400 relatedness effects for pictures were larger following discourse primes with gestures than without, indicating that iconic gestures impact the construction of visually specific aspects of a comprehender's situation model. Moreover, the magnitude of this sensitivity was inversely correlated with the effect size of gestures on early content words, suggesting that, depending on the individual, gestures can affect either local speech–gesture integration or higher-order image-based aspects of downstream comprehension.

Conclusion

Iconic cospeech gestures modulated brain response to content words 200–550ms postonset, suggesting that these gestures made it easier to understand the meaning of the accompanying speech; they also affected ERPs to picture probes presented afterwards, indicating that the presence of gestures affected visual processing of the topic of discourse. Results suggest that individuals vary considerably in their sensitivity to iconic gestures, as well as in the extent to which gestures impact verbal or visuospatial processing mechanisms.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by grants to S.C. from the NSF (#BCS-0843946), an Innovative Research Award from the Kavli Institute for Brain and Mind, and an NIH Training Grant (T32 MH20002-10) award to Y.W.

References

1. Hubbard AL, Wilson SM, Callan DE, Dapretto M. Giving speech a hand: Gesture modulates activity in auditory cortex during speech perception. *Hum Brain Mapp* 2009;30:1028–1037. [PubMed: 18412134]
2. Wu YC, Coulson S. How iconic gestures enhance communication: an ERP study. *Brain Lang* 2007a; 101:234–245. [PubMed: 17222897]
3. Holle H, Gunter TC. The role of iconic gestures in speech disambiguation: ERP evidence. *J Cogn Neurosci* 2007;19:1175–1192. [PubMed: 17583993]
4. Ozyurek A, Willems RM, Kita S, Hagoort P. On-line integration of semantic information from speech and gesture: insights from event-related brain potentials. *J Cogn Neurosci* 2007;19:605–616. [PubMed: 17381252]

5. Van Petten C, Kutas M. Interactions between sentence context and word frequency in event related brain potentials. *Mem Cogn* 1990;18:380–393.
6. Oldfield RC. The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia* 1971;9:97–113. [PubMed: 5146491]
7. Brown G. A frequency count of 190,000 words in the London-Lund corpus of English conversation. *Behav Res Meth, Instrum Comput* 1984;16:502–532.
8. Nuwer MR, Comi G, Emerson R, Fuglsang-Frederiksen A, Guerit JM, Hinrichs H, et al. IFCN standards for digital recording of clinical EEG. *Electroencephalogr Clin Neurophysiol Suppl* 1999;52:11–14. [PubMed: 10590972]
9. Geisser S, Greenhouse S. On methods in the analysis of profile data. *Psychometrika* 1959;24:95–112.
10. Hagoort P, Brown CM. ERP effects of listening to speech: semantic ERP effects. *Neuropsychologia* 2000;38:1518–1530. [PubMed: 10906377]
11. Van Petten C, Coulson S, Rubin S, Plante E, Parks M. Time course of word identification and semantic integration in spoken language. *J Exp Psychol: Learn Mem Cogn* 1999;25:394–417. [PubMed: 10093207]
12. Kelly SD, Creigh P, Bartolotti J. Integrating speech and iconic gestures in a stroop-like task: evidence for automatic processing. *J Cogn Neurosci* 2009;M:M–M.
13. Kelly SD, Kravitz C, Hopkins M. Neural correlates of bimodal speech and gesture comprehension. *Brain Lang* 2004;2004:253–260. [PubMed: 15010257]
14. Kelly SD, Ward S, Creigh P, Bartolotti J. An intentional stance modulates the integration of gesture and speech during comprehension. *Brain Lang* 2007;101:222–233. [PubMed: 16997367]

Fig. 1.

Experimental design. Speech was accompanied either by a video clip (with gestures) or by static freeze frame of the speaker. Discourse primes were followed by related or unrelated picture probes.

Fig. 2.

Words co-timed with either the first gesture of an utterance (black line) or a static freeze frame (gray line). (a) Event-related potential (ERPs) elicited by first-gesture content words exhibited more positive waveforms within the time window of the N400 than words without gestures (measured in microvolts). (b) The distribution of the ERP effect at 400 ms (ERPs to words with first gestures subtracted from ERPs to the same words without gestures).

Fig. 3.

Brain response to picture probes is modulated by the presence of iconic gestures. (a) The relatedness effect (unrelated minus related) is larger over right hemisphere electrodes (F4/FC4/C4/CP4/P4) than left ones (F3/FC3/C3/CP3/C3) when gestures are present in the preceding discourse relative to when they are absent. (b) The effect size of gestures on content words (*x*-axis) is inversely correlated with sensitivity to prime-probe relatedness (*y*-axis). Along the *x*-axis, negative values indicate that when gestures were present, words elicited less negative ERPs than when absent. Along the *y*-axis, negative values indicate that the N400 relatedness effect elicited by pictures was larger when gestures were present than when absent.