Title

Verbal working memory and co-speech gesture processing.

Permalink

https://escholarship.org/uc/item/4qp6r095

Authors

Momsen, Jacob
Gordon, Jared
Wu, Ying
et al.

Publication Date

2020-12-01

DOI

10.1016/j.bandc.2020.105640

Peer reviewed

Published in final edited form as:
Brain Cogn. 2020 December ; 146: 105640. doi:10.1016/j.bandc.2020.105640.

Verbal working memory and co-speech gesture processing

Jacob Momsen[1], **Jared Gordon**[2], **Ying Choon Wu**[3], **Seana Coulson**[1,2]

[1]Joint Doctoral Program Language and Communicative Disorders, San Diego State University and UC San Diego

[2]Cognitive Science Department, UC San Diego

[3]Swartz Center for Computational Neuroscience, UC San Diego

Abstract

Multimodal discourse requires an assembly of cognitive processes that are uniquely recruited for language comprehension in social contexts. In this study, we investigated the role of verbal working memory for the online integration of speech and iconic gestures. Participants memorized and rehearsed a series of auditorily presented digits in low (one digit) or high (four digits) memory load conditions. To observe how verbal working memory load impacts online discourse comprehension, ERPs were recorded while participants watched discourse videos containing either congruent or incongruent speech-gesture combinations during the maintenance portion of the memory task. While expected speech-gesture congruity effects were found in the low memory load condition, high memory load trials elicited enhanced frontal positivities that indicated a unique interaction between online speech-gesture integration and the availability of verbal working memory resources. This work contributes to an understanding of discourse comprehension by demonstrating that language processing in a multimodal context is subject to the relationship between cognitive resource availability and the degree of controlled processing required for task performance. We suggest that verbal working memory is less important for speech-gesture integration than it is for mediating speech processing under high task demands.

Keywords

working memory; iconic gestures; representational gestures; speech-gesture integration; multisensory integration

Corresponding author: Seana Coulson, Cognitive Science, Mail Code 0515, 9500 Gilman Drive, La Jolla, CA 92093 scoulson@ucsd.edu.

Author Contributions
Jacob Momsen: Investigation, Data curation, Software, Visualization, Project administration, Writing – original draft, Writing – review & editing
Jared Gordon: Investigation, Data curation, Writing – review & editing
Ying Choon Wu: Conceptualization, Methodology, Software, Writing – review & editing
Seana Coulson: Conceptualization, Funding acquisition, Software, Writing – original draft, Writing – review & editing

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1 Introduction

Communication in natural environments is often accompanied by paralinguistic information, such as prosody or gesture, which can immediately impact the meaning of a speaker's message (Hagoort & Van Berkum, 2007). Processing gestures during communication has been shown to facilitate the comprehension of spatial information (Austin & Sweller, 2014; So et al, 2015) and help construct visually refined representations of speaker meaning (Wu and Coulson, 2007). Although considerable attention has been given to the impact of gestures on speakers and comprehenders, here we examine the cognitive mechanisms that underlie real-time speech and gesture integration.

This study centers on iconic gestures, which typically represent their referent in a perceptually analog way, e.g., demonstrating a bowl by cupping one's hands together to mimic the contours of the actual dish (Habets et al, 2011). The speaker can convey the general shape of the bowl, i.e. how shallow or steep the edges are, as well as information about its relative size. Because these gestures directly represent visuospatial information, it is possible that perceptual and cognitive mechanisms in the brain directly map contour and motion features of the gesture onto image-based representations of their meaning—similar to how one might perceive and process a cartoon or photograph (Wu and Coulson, 2011). Concepts related to iconic gestures would then presumably rely on visuospatial representations, which would establish visuospatial encoding mechanisms and working memory (WM) resources as necessary for successful speech and gesture processing.

Alternatively, gestural information might be processed and mapped onto language-based semantic representations during discourse comprehension. One might assume that this would be the case for emblematic gestures, i.e., conventionalized body forms that, unlike iconic gestures, can be readily understood in the absence of speech (McNeill, 2005; Fabbri-Destro et al, 2015). Indeed, neuroimaging data suggests these gestures can influence semantic memory activation even when presented in isolation, supporting hypotheses about their status as lexicalized gestural forms (McNeill, 1992; Gunter & Bach, 2006). This possibility is also consistent with a number of studies that have identified a close relationship between the development of gesture and language as well as evidence for overlap in neural networks involved in processing both linguistic and representational gesture information (Goldin-Meadow, 1998; Iverson & Goldin-Meadow, 2005; Willems & Hagoort, 2007; Yang et al, 2015; Proverbio et al, 2015).

Presenting supplementary iconic gestures in tandem with speech often promotes activation in left inferior frontal gyrus (IFG), as well as in medial and superior temporal gyri (MTG, STG) (Özyürek, 2014). Similar sensitivity in the IFG as well as posterior temporal regions to semantic content delivered by representational gestures or speech information has led to the hypothesis that these areas process representational content in a manner that is modality independent (Xu et al, 2009; Straube et al, 2012). The inferior prefrontal cortex also supports online maintenance of verbal information alongside a distributed network including parietal, premotor, and supplementary motor areas (Marvel & Desmond, 2012; Woodward et al, 2006). Thus, verbal encoding and WM resources may be necessary for the comprehension of iconic gestures during multimodal discourse.

1.1 The Cognitive Neuroscience of Working Memory

In recent years, cognitive neuroscientists' understanding of working memory has undergone considerable evolution (see Christophel, 2017 for a review). Early accounts describe a buffer for the short-term maintenance of information stored in a unitary code (Atkinson & Shiffrin, 1968), a model amenable to localization in a single brain region such as pre-frontal cortex (e.g., Goldman-Rakic, 1991). By contrast, the multi-component model suggested WM involved both the maintenance and manipulation of information utilizing at least two different representational formats (Baddeley, 2000; Baddely & Hitch, 1974). The suggestion that WM utilizes dissociable verbal and visuospatial codes was supported by neuroimaging studies implicating fronto-temporal language networks in memory for verbal information and frontal-parietal networks in memory for visual (object) and spatial information (e.g., Smith & Jonides, 1998). Subsequent research has supported the construal of WM as a particularly active state of long-term memory so that there is no single locus of the WM system (Cowan, 2017). Rather, the ability to maintain and manipulate information relies on the neural substrate of the attentional control system for reactivating memory stored in a distributed fashion across the cortex while preventing task-irrelevant information from interfering with current behavioral goals (D'Esposito & Postle, 2015; Oberauer, 2019).

Even as investigators have learned that limitations on WM are often due to constraints of the attentional control system (Unsworth & Engle, 2007), the notion of different verbal and visuospatial WM systems has been maintained as a heuristic distinction, as functional networks recruited for verbal versus visuospatial tasks are partially orthogonal (c.f. Buschbaum & D'Esposito, 2019; Jerde & Curtis, 2013). When correlations are found between measures of verbal versus visuospatial processing capacity, they likely index the modality-independent executive control processes important for performing the types of complex span tasks often used to assess those skills (Kane & Engel, 2002). However, when testing a population with uniformly high executive skills, verbal and visuospatial WM can be dissociated based on differences in experience and aptitude with verbal versus visuospatial material (Conway, et al., 2005; Schwering & MacDonald, 2020).

1.2 Verbal Working Memory and Multimodal Discourse Comprehension

The present study focuses on verbal WM and its relationship to multimodal discourse comprehension. By determining if increased loads on verbal WM processes interfere with online speech processing during multimodal discourse comprehension, it will be possible to draw conclusions about the importance of verbal WM in message-level comprehension processes that involve combining analog information from iconic gestures with symbolic language representations. Dual-task paradigms have often been used to identify potential overlap in the cognitive resources recruited for different tasks (see e.g., Luck & Vogel, 2001). These studies typically involve a memory component, where active performance on a primary target task is embedded within the maintenance portion of a secondary memory task (Logan, 1979; Baddeley, 1986). If performance on one task suffers because of the demands imposed by a simultaneous concurrent task, the tasks are thought to involve overlapping cognitive resources, which results in interference that diminishes successful task performance. The current study uses this logic to investigate a potential cognitive intersection between verbal WM and the ability to integrate speech and iconic gestures as they unfold in real time.

Event-related potentials time-locked to speech can provide information about how contextual information activates relevant semantic features of upcoming words. To index the facilitative impact that gestures have on accompanying speech, here we utilize the N400, a well-studied neural response to words and other meaningful stimuli. The N400 is thought to index the retrieval of information from semantic memory and its amplitude is a function of the fit between a stimulus and its preceding and concurrent context, as unexpected stimuli elicit large N400 whereas contextually primed stimuli elicit reduced (i.e., less negative) N400 amplitudes (see Kutas & Federmeier, 2011 for review).

Prior work on speech and gesture processing has shown that the presence of gestures during communication reduce N400 amplitudes, suggesting semantically related gestures can make it easier to understand the accompanying speech (Holle & Gunter, 2007; Wu & Coulson, 2010; Fabbri-Destro et al, 2015). Moreover, the size of N400 effects can serve as an indicator of the availability of verbal WM operations. Previous investigators have shown that N400 effects can be reduced or eliminated by experimental manipulations that serve to occupy WM resources. For example, Gunter, Jackson, and Mulder (1995) compared the size of N400 congruity effects in sentences whose structure posed either high or low demands on WM. In young adults, N400 congruity effects were smaller for sentences that posed high demands on WM than those that posed low demands; in older adults, N400 congruity effects were absent in sentences that posed high demands on WM, and present in sentences that posed low demands on WM (Gunter, et al., 1995). Similarly, D’Arcy and colleagues varied the load of a concurrent verbal WM task and found N400 congruency effects were smaller when memory loads were high than when they were low (D’Arcy, et al., 2005). If verbal WM availability similarly mediates the use of gestural information to aid in lexical processing, then demands on verbal WM systems should reduce N400 sensitivity to the semantic relationship between speech and gestures in a similar fashion.

1.3 The Present Study

Here we used electrophysiological measures of speech comprehension to assess healthy adults’ sensitivity to co-speech iconic gestures while under varying levels of verbal WM load. College-aged adults performed a dual-task paradigm involving multimodal discourse processing embedded within the retention period of a verbal WM task. ERPs provided a real-time neural index of discourse comprehension to investigate the impact of increased verbal WM load on speech-gesture integration. Under the hypothesis that verbal WM is recruited for the use of gesture information to aid online speech comprehension, we expected to observe larger N400 effects of gesture congruity for words in the low memory load condition compared to those in the high memory load condition. By contrast additive effects of WM load and gesture congruity would point to the relative independence of the two processes.

2 Methods

2.1 Participants

18 English speaking, right-handed college-aged adults participated in the experiment.[1] None reported any history of neurological or learning disorders (14 females; mean age = 21). All participants gave informed consent and received academic credit for their participation.

2.2 Working Memory Tasks

Before the EEG recording, two computerized tasks were given to each participant to provide an independent measure of their WM abilities. Spatial skills were tested using the Corsi block task, which required that participants memorize and recall the order of squares that flashed in random sequences on a computer screen (Corsi scores: mean = 20.4, SD = 3.2). One participant's Corsi score information was missing from all analyses due to lost data. The Sentence Span task was used as a measurement of verbal WM capacity (Daneman & Carpenter,1980). This task required participants to listen to a series of unrelated sentences while memorizing the final word in each sentence. Questions probing sentence understanding were administered to ensure that participants processed sentence meaning. Final scores were calculated based on the number of final words the participant could recall correctly (Conway et al, 2005) (Verbal scores: mean= 34.1, SD = 3.5). See Wu and Coulson (2014) for further description of the WM assessments.

2.3 Materials

140 videos were used as discourse primes for the experiment. Discourse primes were constructed from video footage of a speaker describing various objects and actions using both speech and iconic gestures, (e.g., the shape of furniture or swinging a golf club). Incongruent videos were created by digitally recombining the audio (speech) and video (gesture) portions of the original video recordings that were used in the congruent condition. The final collection of stimuli included 140 congruent and 140 incongruent videos, with speech and video content fully matched across conditions. The actor's face in each video was blurred to obscure any mismatch between orofacial movements and the accompanying speech. Video onset always corresponded to the stroke of the first gesture, and video offset was at the end of the utterance unit. Variability in the length of video stimuli (between 2 and 8 seconds) was a result of ensuring that all clips contained a coherent utterance. The onset of the first content word in the speech files, (i.e., an open-class noun or verb), occurred at various timepoints across the discourse stimuli (mean = 743ms post video onset; SD = 466ms). This marked the first time point in the clip that participants could determine the presence of a semantic mismatch between speech and gestures.

Picture probes were photographic images of the objects described in each video and were intended to match the speech content of the discourse primes. Iconic gesture information was thus unrelated to the picture probes in half of the trials. However, because ERPs to picture probes were not analyzed, they will not be discussed further.

[1] Our target sample size was between 16 and 20 participants and was based on samples enrolled in similar studies (Willems, Özyürek, & Hagoort, 2007; Wu & Coulson, 2011). Power analysis with the WebPower package in R suggested n=16 was sufficient, assuming Cohen's d of 0.87 (Zhang & Yuan, 2018).

Each participant viewed 140 videos, 70 congruent and 70 incongruent. Two stimulus lists were created to counterbalance whether a given audio file appeared with congruent or incongruent gestures. Subjects were distributed evenly across the two lists (List 1: n=9; List 2: n=9).

2.4 Procedure

Subjects performed the task in a dimly lit, sound-attenuated room, and viewed stimuli on a 19” color monitor. Behavioral responses were recorded from a mouse that participants were instructed to keep their dominant hand on for the entire experiment. Each trial began with a fixation cross that appeared in the center of the screen. 500ms after the cross, the WM portion of the experiment began with a series of spoken digits presented at a rate of 1 digit per second. In low memory load trials, only one digit was presented; in high load trials, a sequence of four randomly ordered digits were presented. Participants were instructed to memorize the sequence of digits in the order of their presentation to recall them at the end of the trial. After the encoding portion of the WM task, 500ms passed before the discourse primes (videos) were played in the center of the screen.

Discourse primes were videos of a speaker describing objects or actions while producing co-speech iconic gestures. 500ms after the completion of the video, a picture probe appeared in the center of the screen. Participants were instructed simply to attend to both the primes and probes. 500ms after the picture probe disappeared, the free recall portion of the WM task began. Participants were presented with a randomized display of 10 single-digit numerals across the screen and were instructed to use the mouse to select the digits they had heard in the order that they heard them. Feedback on the ordered recall task was given 500ms after the response on each trial. 14 trials comprised a single block during the experiment, and each participant completed 10 blocks (140 trials total). Blocks were separated by self-paced breaks.

2.5 EEG Recording and Analysis

EEG was recorded in a soundproof, electromagnetically shielded chamber with 29 scalp electrodes placed at standard International 10–20 sites. Two additional mastoid electrodes and three facial electrodes were used for referencing and artifact detection, respectively. Signals referenced to the left mastoid were bandpass filtered (0.01–40Hz) and digitized online at 250Hz. After recording, EEG data were re-referenced to the mean of left and right mastoid sites. After segments of continuous data containing drift or muscle artifacts were trimmed, ICA was applied to the continuous EEG data in order to correct data for eye-related artifacts. After the removal of ICs representing non-brain related activity, back-projected EEG data were epoched and remaining trials that contained artifacts related to eye movements, drift, or muscle noise were eliminated before statistical analysis (Mean epochs rejected = 5%; SD = 2.9%). EEG epochs spanned from 200ms before stimulus onset to 800ms afterward, and a 200ms baseline correction was applied. Eighteen electrode sites were used in the ERP analysis based on our expectations about the central distribution of N400 effects (See Figure 3). Due to a priori motivations to focus on how the availability of verbal WM resources impacts real-time speech-gesture integration, analysis focused on ERPs elicited by the first content word in each discourse video. Trials in which participants responded incorrectly on the memory task were not included in the analysis. A 15Hz low-pass filter was applied to all averaged ERPs presented in figures.

Our analysis focused on the neural response to words to examine the impact of increased verbal WM load on the ability to integrate speech and gesture information during discourse comprehension. Omnibus ANOVAs for first content words included factors of Load (High and Low), Video (Congruent and Incongruent), and Anteriority (6 levels, see Figure 3). In addition, the omnibus model included a between-subjects factor representing experimental List (1 or 2) to ensure that discrepancies related to the pairing of discourse stimuli with each WM Load condition did not influence effects of Video congruity. Reported p-values represent statistics after performing Greenhouse-Geisser corrections. Effects of discourse congruity were calculated as the difference between averaged ERPs time-locked to initial content words in videos containing congruent speech and gesture combinations from incongruent ones.

3 Results

3.1 Memory Task Performance

A generalized linear mixed effects model with a logit link function was used to observe the influence of memory load and video congruity manipulations on performance of the verbal recall task. These models are well-suited for the binomial distribution of the performance data and allow for generalization to both novel participants and items (Baayen et al, 2008). Accordingly, random intercepts for subjects were included to eliminate variance associated with by-subject effects, and random intercepts for discourse prime were included to eliminate variance associated with by-item effects. The model revealed that the Load manipulation significantly influenced performance ($\beta = -1.85$; SE = 0.38; $p<0.0001$), such that recall was less likely to be accurate in the high memory load condition. By contrast, the video congruity manipulation did not significantly influence performance ($p=0.99$) (Figure 2).

Although task performance was relatively high across all condition types (total proportion of correct trials = 95.8%), individual differences were nonetheless present. In a separate regression analysis, performance on the verbal recall task across both load conditions was modeled in a linear regression model using standardized Corsi and Sentence Span (total) scores as predictors. This linear model revealed a positive relationship between Sentence Span scores and performance on the verbal WM task ($\beta = 0.02$; SE = 0.007; $p<0.01$), which accounted for approximately 41% of unique variance in verbal recall scores (Figure 2B). There was no significant influence of visuospatial WM abilities on task performance ($p=0.9$), suggesting that our secondary WM task indeed targeted verbal WM.

3.2 ERPs to Words

ERPs time-locked to the onset of the first content word in the discourse primes were measured between 200 and 500ms and these mean amplitude measurements were compared across Load and Video conditions. Repeated measures ANOVA indicated a significant Load by Video by Anteriority effect ($F(5,85)=9.63$; $p<0.001$, $\varepsilon = 0.4$). The omnibus model also returned interactions between Load and Anteriority (F(5, 80) = 4.87, $p < 0.05$, ε = .30) and Load by List by Anteriority (F(5,80)=14.49; p<0.05, ε = .30). These interactions result because words in High Load trials elicit slightly more positive (~1.2 uV) ERPs over frontal and frontocentral sites, and post hoc analyses indicated this load effect was larger and more robust in one of our stimulus lists (Load x Anteriority: $p < 0.05$) than the other (Load x Anteriority: p=0.27). Importantly, no interactions between Video and experimental List were identified (List by Video: p=0.28; List by Video by Anteriority: p=0.26; List by Video by Load: p=0.27; List by Video by Load by Anteriority: p=0.53).

Separate analyses conducted in each memory load condition revealed a significant Video x Anteriority effect in the Low Load condition (F(5,85)=3.25, $p < 0.05$, ε = 0.4), reflecting a more negative response to words in incongruent than congruent videos over frontal and frontocentral electrode sites, and a Video x Anteriority effect in the High Load condition (F(5,85) = 4.4, $p < 0.05$, ε = 0.3), reflecting a relative positivity to words in incongruent videos over frontal and frontocentral sites. Figure 3 displays the ERPs to words encountered in the congruent versus incongruent videos in the low memory load condition. Relative to congruent trials, words in incongruent videos elicited a negativity (N400) that was more pronounced over anterior channels. This expected Video effect suggests words in low Load trials were comprehended more easily in the congruent than incongruent multimodal discourse. By contrast, in the high Load trials, words in incongruent videos elicited a positivity with a similar frontal distribution (Figure 4B). The impact of co-speech gestures on word processing was thus qualitatively different under conditions of high and low verbal WM Load (see Figure 4).

4 Discussion

The present study examined the relationship between verbal working memory (WM) and speech-gesture integration by measuring neural indices of discourse comprehension under varying levels of verbal WM load. Performance on the verbal WM task administered in conjunction with video presentation was significantly worse in the high WM load condition, and our offline measure of participants' verbal WM capacity was positively associated with overall task performance. By contrast, our measure of visuospatial WM capacity was not related to task performance, suggesting that the digit recall task successfully diverted verbal WM resources during the presentation of discourse videos. If similar cognitive resources are involved both in maintaining verbal information in WM and the integration of speech and gesture, then we would expect increased WM loads to interact with neural indices of speech-gesture congruity.

As predicted, the ERP effects of speech-gesture congruity *did* differ as a function of memory load–however, the nature of the interaction was unexpected. Under conditions of low memory load, words accompanied by incongruent gestures elicited a larger N400 than they did for congruent trials. When verbal resources were less available, words accompanied by incongruent gestures elicited a larger frontal positivity. Below, we relate this positivity to the P3a component, and suggest the availability of verbal WM resources impacts the allocation of attention to multimodal discourse. The current study supports the claim that successful multimodal discourse comprehension is supported by multisensory integration mechanisms that operate automatically under optimal processing conditions, and that verbal WM is important for comprehension when this mechanism fails rather than for online speech-gesture integration per se.

4.1 Speech-gesture integration and the influence of task demands

The speech-gesture incongruity effect observed in low WM load trials is similar to the N400 effects reported in previous ERP studies exploring the benefit that meaningful gestures can provide for semantic retrieval (Wu & Coulson, 2010; Holle & Gunter, 2007). While the classical N400 component is largest over parietal sites (Kutas & Federmeier, 2011), the fronto-central distribution of the N400 incongruity effect in the present study is very common in studies of action and gesture comprehension (see Amoruso, et al., 2013 for a review). Considering the literature on the neural generators of the N400 component, Amoruso and colleagues (2013) suggest the frontally distributed N400 elicited by gestural stimuli reflects the increased contribution of motor and pre-motor cortices to neural generators of the classical N400 in the frontal, temporal, and parietal lobes (see Lau, Phillips, & Poeppel, 2009 for a review).

Recent MEG work similarly comparing activity between matching and mismatching speech and iconic gestures provides some indirect support for this hypothesis. Namely, Drijvers et al (2018) localized beta band suppression effects in response to incongruent compared to congruent speech-gesture combinations in left inferior frontal, precentral, supplementary motor, and somatosensory cortices. Although they did not examine event related potentials to speech directly, a number of studies point to correspondence between modulations of neural oscillations in the beta band range (~13–30Hz) and N400 generation (Dave et al, 2020; He et al, 2020; Lewis et al, 2017; Wang et al, 2012), although more work is necessary to explain how gestures might modulate speech-processing networks that contribute to semantic memory updating.

As noted above, speech-gesture incongruity effects in high WM load trials were qualitatively different from those in low load trials. Instead of reduced N400 effects, per our original hypothesis, speech in incongruent videos elicited more positive ERPs than those in congruent ones. In fact, this frontal positivity resembles the P3a, an ERP component thought to reflect attentional shifts toward unexpected stimuli, that exhibits a frontocentral distribution akin to that observed in the current study (Muller-Gass et al, 2007; Polich, 2007; Sussman et al, 2003). Further, P3a generation depends on task difficulty, as the same non-target stimuli that elicit a prominent P3a in difficult discrimination tasks do not do so in easier ones (Comerchero & Polich, 1999). That is, novel stimuli that are not task relevant can be processed relatively automatically when task demands are low, but, require a frontally mediated orienting response when task demands are high.

In the present study, the difficulty of the memory task apparently impacted participants' attention to the discourse videos presented during the maintenance period. The low load memory task presumably did not demand intense attentional focus, thereby facilitating more automatic processing of incongruent gestures. Under normal circumstances, gestures influence the interpretation of the accompanying speech, so that words occurring with semantically incongruent gestures elicit larger amplitude N400 than those with congruent gestures. However, because the high memory load trials were more difficult, the remaining WM resources were not sufficient to achieve the typically effortless integration of semantic information across the two modalities.

Despite some claims that the integration of the meaning of speech and gestures is fully automatic (McNeill, 1992; Kelly et al, 2010), a number of factors have been shown to interfere with speech-gesture integration (Kelly et al, 2007; Habets et al., 2011; Obermeier et al., 2011). For example, when speech and gestures unfold in a temporally asynchronous manner, semantic integration does not proceed automatically, but requires controlled processes (Obermeier et al., 2011). Indeed, some evidence suggests perceptual “unity” effects driven by the spatiotemporal coincidence or semantic congruency of stimuli across different modalities may also extend to speech and gesture pairings, such that semantic alignment promotes speech-gesture integration (Margiotoudi et al, 2014). Given that attentional resources are thought to be particularly important during WM operations that require people to reconcile conflicting information across modalities (Sepp et al, 2019), our memory load manipulation may have had a disproportionate effect on the incongruent gestures.

To evaluate this possibility, a post hoc analysis compared ERP activity as a function of WM load on words accompanied by congruent as opposed to incongruent gestures. This analysis confirmed that the WM load manipulation did not significantly affect ERPs to words in the congruent videos but had a robust impact on words accompanied by incongruent gestures (see Figure 5). Although the concurrent verbal WM task had little impact on the relatively automatic integration of speech with congruent gestures, it compromised the ability of participants to deal with incongruent gestures.

The disproportionate impact of verbal WM load on the processing of incongruent speech-gesture pairings is in keeping with neuroimaging studies of multimodal discourse processing. Such studies indicate that the semantic relatedness of speech and iconic gestures at least partially dictate the configuration of functional networks engaged in discourse processing (Willems et al, 2007; Green et al, 2009). For example, unrelated co-speech gestures disproportionately recruit bilateral inferior frontal, supplementary motor, and left inferior parietal regions—areas heavily implicated in the maintenance of and control over verbal information in WM (Baldo & Dronkers, 2006; see Nee et al, 2012 for review).

Research investigating the dynamics of multisensory integration indicates that when general processing demands are low, there is less need for top-down attentional resources to combine information across modalities into a coherent multimodal representation. However, when people are forced to divert relevant modality-specific resources toward a secondary task, the effectiveness of more automatic sensory integration mechanisms can become compromised (Talsma et al, 2010; Alsius, 2005). Findings of the present study thus point to commonalities in the neural mechanisms at play in the integration of semantic information in speech and gestures with those in more basic sensory integration.

4.2 Gesture and Working Memory

Previous work suggests the spontaneous production of meaningful gestures can facilitate lexical retrieval processes (Rauscher et al, 1996; Krauss, 1998; Pine et al, 2007), especially when WM resources are compromised (Goldin-Meadow et al, 2001; Gillespie et al, 2014; Wagner, et al., 2004). Such research suggests that the act of producing a meaningful gesture reduces the demands of speech production, presumably because the bodily motion primes the relevant information in semantic memory. Consistent with neuroimaging studies that indicate a common network of brain regions supporting the comprehension of gestures and speech (Dick et al, 2009; Xu et al, 2009; Özyürek, 2014), the current study underscores the role of meaningful co-speech gestures in the online access of word meaning (Wu & Coulson, 2010).

Although previous work in our lab has shown little evidence for a relationship between verbal WM resources and sensitivity to the congruity of speech and gestures, that work relied on behavioral responses to probes presented after the offset of multimodal discourse (Wu & Coulson, 2014). By contrast, the present study measured the brain response to words during the discourse itself and was thus better suited to detect a role for verbal WM in real-time processing. Importantly, these data suggest that while reducing the availability of verbal WM resources had little impact on speech accompanied by congruent gestures, it does modulate the brain response to spoken words when the meaning of accompanying gestures is less apparent.

Even with reduced verbal WM, however, the brain response differentiated congruent and incongruent gestures. Interestingly, the amplitude of the P3a response to incongruent gestures was not related to participants' scores on our measures of verbal WM capacity, but rather to individual differences in visuospatial WM. That is, participants with better scores on the Corsi block task exhibited larger P3a components in the high WM load incongruent condition (β=0.34µv; $p<0.05$). As the amplitude of P3a increases as a function of stimulus salience (Nittono, 2006), this finding is consistent with behavioral studies in our lab that indicate a relationship between visuospatial WM ability and sensitivity to co-speech iconic gestures (Wu & Coulson, 2014).

Similarly, Özer and Göksun (2019) have investigated how individual differences in verbal and visuospatial WM capacity relate to sensitivity to speech and gesture in multimodal discourse. They find that visuospatial WM ability relates to sensitivity to gestural information, while verbal WM relates to sensitivity to speech (Özer & Göksun, 2019). These differential effects reflect a modality-specific relationship between WM resources and the ability to interpret different channels of information in multimodal discourse. In the present study, the relationship between visuospatial skills and participants' sensitivity to incongruent gestures supports the hypothesis that these skills are important for the interpretation of gestures.

We conclude that under typical conditions, the neurocognitive architecture that supports controlled verbal WM operations is not recruited for the interpretation of meaningful co-speech iconic gestures. Instead, verbal WM resources are activated in situations where meaning is ambiguous or unclear. Thus, its role appears more relevant for resolving semantic uncertainty or temporarily buffering discordant streams of information than for mediating the extraction of visuospatial content from gestures. Future work could test this hypothesized relationship of verbal WM for speech-gesture integration by parametrically varying the relatedness of co-speech gestures.

Acknowledgments

Research was supported by NSF award #BCS-0843946 to S.C. and NIH T32 DC007361 to J.M.

References

Austin EE, & Sweller N. (2014). Presentation and production: The role of gesture in spatial communication. Journal of Experimental Child Psychology, 122, 92–103. [PubMed: 24549229]

Alsius A, Navarra J, Campbell R, & Soto-Faraco S. (2005). Audiovisual integration of speech falters under high attention demands. Current Biology, 15(9), 839–843. [PubMed: 15886102]

Amoruso L, Gelormini C, Aboitiz FA, González A, Manes F, Cardona J, & Ibanez A. (2013). N400 ERPs for actions: building meaning in context. Frontiers in human neuroscience, 7, 57. [PubMed: 23459873]

Baayen R, Davidson D, & Bates D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59(4), 390–412.

Baddeley AD (1986).Working memory. Oxford: Oxford University Press.

Baldo JV, & Dronkers NF (2006). The role of inferior parietal and inferior frontal cortex in working memory. Neuropsychology, 20(5), 529. [PubMed: 16938015]

Buchsbaum BR, Hickok G, & Humphries C. (2001). Role of left posterior superior temporal gyrus in phonological processing for speech perception and production. Cognitive Science, 25(5), 663–678.

Comerchero MD, & Polich J. (1999). P3a and P3b from typical auditory and visual stimuli. Clinical Neurophysiology, 110(1), 24–30 [PubMed: 10348317]

Conway AR, Kane MJ, Bunting MF, Hambrick DZ, Wilhelm O, & Engle RW (2005). Working memory span tasks: A methodological review and user's guide. Psychonomic bulletin & review, 12(5), 769–786. [PubMed: 16523997]

Daneman M, Carpenter PA Individual differences in working memory and reading Journal of Verbal Learning and Verbal Behavior, 19 (4) (1980), pp. 450–466

Dave S, VanHaerents S, & Voss J. (2020). Cerebellar theta and beta noninvasive stimulation rhythms differentially influence episodic memory versus language. bioRxiv.

Dick AS, Goldin-Meadow S, Hasson U, Skipper JI, & Small SL (2009). Co-speech gestures influence neural activity in brain regions associated with processing semantic information. Human Brain Mapping, 30(11), 3509–3526. [PubMed: 19384890]

Dick AS, Goldin-Meadow S, Solodkin A& Small SL. 2012 Gesture in the developing brain. Dev. Sci 15, 165–180. [PubMed: 22356173]

Fabbri-Destro M, Avanzini P, De Stefani E, Innocenti A, Campi C, & Gentilucci M. (2015). Interaction between words and symbolic gestures as revealed by N400. Brain topography, 28(4), 591–605. [PubMed: 25124860]

Gillespie M, James AN, Federmeier KD, & Watson DG (2014). Verbal working memory predicts co-speech gesture: Evidence from individual differences. Cognition, 132(2), 174–180. [PubMed: 24813571]

Goldin-Meadow S. (1998). The development of gesture and speech as an integrated system. New Directions for Child and Adolescent Development,1998(79), 29–42.

Goldin-Meadow S, & Alibali MW (2013). Gesture's role in speaking, learning, and creating language. Annual review of psychology, 64, 257–283.

Goldin-Meadow S, Nusbaum H, Kelly SD, & Wagner S. (2001). Explaining Math: Gesturing Lightens the Load. Psychological Science,12(6), 516–522. [PubMed: 11760141]

Gunter TC, Jackson JL, & Mulder G. (1995). Language, memory, and aging: an electrophysiological exploration of the N400 during reading of memory-demanding sentences. Psychophysiology, 32(3), 215–229. [PubMed: 7784530]

Habets B, Kita S, Shao Z, Özyurek A, & Hagoort P. (2011). The Role of Synchrony and Ambiguity in Speech–Gesture Integration during Comprehension. Journal of Cognitive Neuroscience, 23(8), 1845–1854. [PubMed: 20201632]

Hagoort P, & Berkum JV (2007). Beyond the sentence given. Philosophical Transactions of the Royal Society B: Biological Sciences, 362(1481), 801–811.

He Y, Luell S, Muralikrishnan R, Straube B, & Nagels A. (2020). Gesture's body orientation modulates the N400 during semantic integration of gesture and visual sentence. bioRxiv.

Holle H, & Gunter TC (2007). The Role of Iconic Gestures in Speech Disambiguation: ERP Evidence. Journal of Cognitive Neuroscience, 19(7)

Holler J, Kelly S, Hagoort P, & Ozyurek A. (2012). When gestures catch the eye: The influence of gaze direction on co-speech gesture comprehension in triadic communication. In the 34th Annual Meeting of the Cognitive Science Society (CogSci 2012) (pp. 467–472). Cognitive Society.

Iverson JM, & Goldin-Meadow S. (2005). Gesture Paves the Way for Language Development. Psychological Science, 16(5), 367–371. [PubMed: 15869695]

Kelly SD, Ward S, Creigh P, & Bartolotti J. (2007). An intentional stance modulates the integration of gesture and speech during comprehension. Brain and language, 101(3), 222–233. [PubMed: 16997367]

Kelly SD, Creigh P, & Bartolotti J. (2010). Integrating Speech and Iconic Gestures in a Stroop-like Task: Evidence for Automatic Processing. Journal of Cognitive Neuroscience, 22(4), 683–694. [PubMed: 19413483]

Krauss RM, & Hadar U. (1999). The role of speech-related arm/hand gestures in word retrieval. Gesture, Speech, and Sign, 93–116.

Kutas M, & Federmeier KD (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). Annual Review of Psychology, 62(1), 621–647.

Kuznetsova A, Brockhoff PB and Christensen RHB (2017). "lmerTest Package: Tests in Linear Mixed Effects Models." Journal of Statistical Software, 82(13), pp. 1–26.

Lau EF, Phillips C, & Poeppel D. (2008). A cortical network for semantics:(de) constructing the N400. Nature Reviews Neuroscience, 9(12), 920–933. [PubMed: 19020511]

Lewis AG, Schoffelen JM, Hoffmann C, Bastiaansen M, & Schriefers H. (2017). Discourse-level semantic coherence influences beta oscillatory dynamics and the N400 during sentence comprehension. Language, Cognition and Neuroscience, 32(5), 601–617.

Logan G. (1979). On the use of a concurrent memory load tomeasure attention and automaticity. Journal of Experimental Psychology: Human Perception and Performance, 5, 189–207.

Luck Steven & Vogel Edward. (2001). Multiple sources of interference in dual-task performance: the cases of the attentional blink and the psychological refractory period. Current directions in Psychological Science 10.1093/acprof:oso/9780198505150.003.0007.

Margiotoudi K, Kelly S, & Vatakis A. (2014). Audiovisual temporal integration of speech and gesture. Procedia-Social and Behavioral Sciences, 126, 154–155.

McNeill D (1992) Hand and mind: what gestures reveal about thought. University of Chicago Press, Chicago

McNeill D (2005) Gesture and thought. University of Chicago Press, Chicago

Ménoret M, Varnet L, Fargier R, Cheylus A, Curie A, Portes VD, Paulignan Y. (2014). Neural correlates of non-verbal social interactions: A dual-EEG study. Neuropsychologia, 55, 85–97. [PubMed: 24157538]

Nittono H. (2006). Voluntary stimulus production enhances deviance processing in the brain. International Journal of Psychophysiology, 59(1), 15–21. [PubMed: 16257077]

Muller-Gass A, & Schröger E. (2007). Perceptual and cognitive task difficulty has differential effects on auditory distraction. Brain research, 1136, 169–177. [PubMed: 17223092]

Nee DE, Brown JW, Askren MK, Berman MG, Demiralp E, Krawitz A, & Jonides J. (2013). A meta-analysis of executive components of working memory. Cerebral cortex, 23(2), 264–282. [PubMed: 22314046]

Oberauer K. (2019). Working Memory and Attention - A Conceptual Analysis and Review. Journal of cognition, 2(1), 36. 10.5334/joc.58 [PubMed: 31517246]

Obermeier C, Holle H, & Gunter TC (2011). What iconic gesture fragments reveal about gesture–speech integration: When synchrony is lost, memory can help. Journal of Cognitive Neuroscience, 23(7), 1648–1663. [PubMed: 20350188]

Özer D, & Göksun T. (2019). Visual-spatial and verbal abilities differentially affect processing of gestural vs. spoken expressions. Language, Cognition and Neuroscience, 1–19.

Özyürek A. (2014). Hearing and seeing meaning in speech and gesture: Insights from brain and behaviour. Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1651), 20130296

Pine KJ, Bird H, & Kirk E. (2007). The effects of prohibiting gestures on children's lexical retrieval ability. Developmental Science, 10(6), 747–754. [PubMed: 17973791]

Polich J. (2007). Updating P300: An integrative theory of P3a and P3b. Clinical Neurophysiology, 118(10), 2128–2148 [PubMed: 17573239]

Proverbio AM, Gabaro V, Orlandi A, & Zani A. (2015). Semantic brain areas are involved in gesture comprehension: An electrical neuroimaging study. Brain and Language, 147, 30–40 [PubMed: 26011745]

Rauscher FH, Krauss RM, & Chen Y. (1996). Gesture, speech, and lexical access: The role of lexical movements in speech production. Psychological science, 7(4), 226–231.

Richardson FM, Ramsden S, Ellis C, Burnett S, Megnin O, Catmur C, ... & Price CJ. (2011). Auditory short-term memory capacity correlates with gray matter density in the left posterior STS in cognitively normal and dyslexic adults. Journal of Cognitive Neuroscience, 23(12), 3746–3756. [PubMed: 21568634]

Sepp S, Howard SJ, Tindall-Ford S, Agostinho S, & Paas F. (2019). Cognitive load theory and human movement: Towards an integrated model of working memory. Educational Psychology Review, 1–25.

So W, Shum PL, & Wong MK (2015). Gesture is More Effective than Spatial Language in Encoding Spatial Information. Quarterly Journal of Experimental Psychology, 68(12), 2384–2401.

Sussman E, Winkler I, & Schröger E. (2003). Top-down control over involuntary attention switching in the auditory modality. Psychonomic bulletin & review, 10(3), 630–637. [PubMed: 14620357]

Talsma D, Senkowski D, Soto-Faraco S, & Woldorff MG (2010). The multifaceted interplay between attention and multisensory integration. Trends in cognitive sciences, 14(9), 400–410. [PubMed: 20675182]

Wagner SM, Nusbaum H, & Goldin-Meadow S. (2004). Probing the mental representation of gesture: Is handwaving spatial? Journal of Memory and Language, 50(4), 395–407

Wang L, Jensen O, Van den Brink D, Weder N, Schoffelen JM, Magyari L, ... & Bastiaansen M. (2012). Beta oscillations relate to the N400m during language comprehension. Human brain mapping, 33(12), 2898–2912. [PubMed: 22488914]

Willems RM, & Hagoort P. (2007). Neural evidence for the interplay between language, gesture, and action: A review. Brain and Language,101(3), 278–289. [PubMed: 17416411]

Willems RM, 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(7), 1235–1249. [PubMed: 18284352]

Wu YC, & Coulson S. (2011). Are depictive gestures like pictures? Commonalities and differences in semantic processing. Brain and Language, 119(3), 184–195 [PubMed: 21864890]

Wu YC, & Coulson S. (2014). Co-speech iconic gestures and visuo-spatial working memory. Acta Psychologica, 153, 39–50. [PubMed: 25282199]

Wu YC, & Coulson S. (2010). Gestures modulate speech processing early in utterances. NeuroReport, 21(7), 522–526. [PubMed: 20375745]

Wu YC, & Coulson S. (2007). How iconic gestures enhance communication: An ERP study. Brain and Language, 101(3), 234–245 [PubMed: 17222897]

Xu J, Gannon PJ, Emmorey K, Smith JF, & Braun AR (2009). Symbolic gestures and spoken language are processed by a common neural system. Proceedings of the National Academy of Sciences,106(49), 20664–20669

Yang J, Andric M, & Mathew MM (2015). The neural basis of hand gesture comprehension: A meta-analysis of functional magnetic resonance imaging studies. Neuroscience & Biobehavioral Reviews, 57, 88–104. [PubMed: 26271719]

Zhang Z, & Yuan K-H (2018). Practical Statistical Power Analysis Using Webpower and R. Granger, IN: ISDSA Press.

Highlights

- EEG recorded to speech with semantically congruent/incongruent gestures under high/low verbal load
- ERP to words with congruent gestures not impacted by VWM load suggesting automaticity
- Words with incongruent gestures elicit enhanced N400 with low load, P3a with high load
- Speech-gesture integration similar to basic sensory integration in use of top-down attention
- VWM recruited for speech-gesture integration when meaning of gestures is unclear

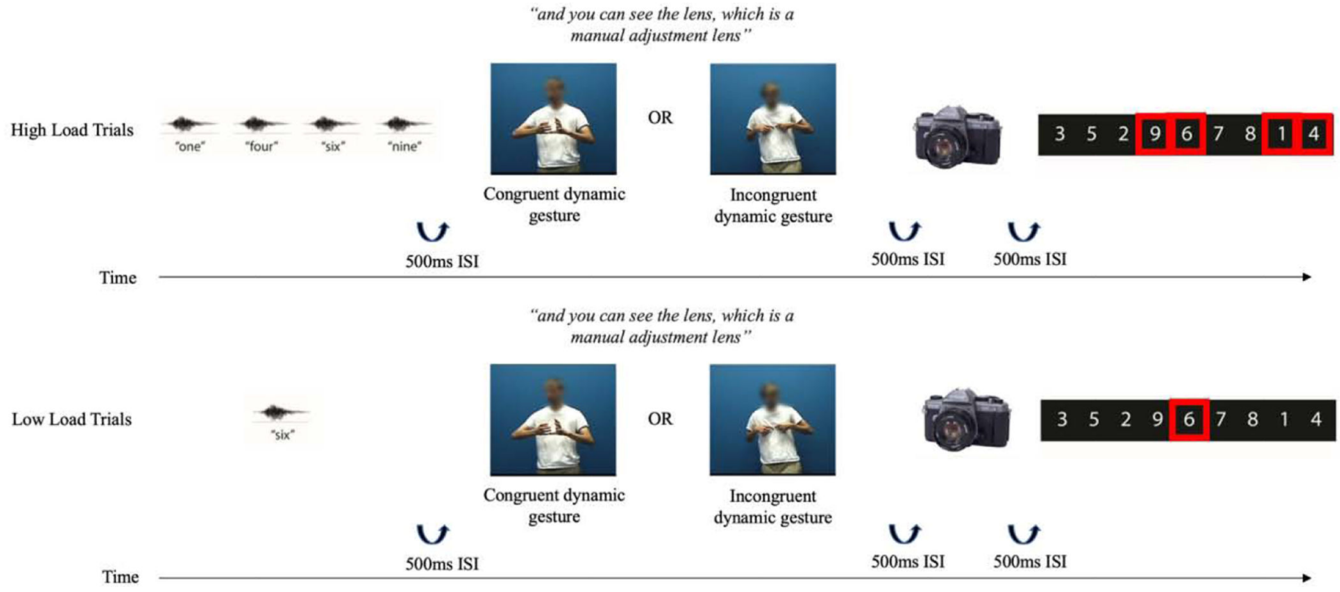
"and you can see the lens, which is a manual adjustment lens"
High Load Trials
"one"
"four"
"six"
"nine"
Congruent dynamic gesture
OR
Incongruent dynamic gesture
3 5 2 9 6 7 8 1 4
500ms ISI
500ms ISI
500ms ISI
Time
"and you can see the lens, which is a manual adjustment lens"
Low Load Trials
"six"
Congruent dynamic gesture
OR
Incongruent dynamic gesture
3 5 2 9 6 7 8 1 4
500ms ISI
500ms ISI
500ms ISI
Time

Figure 1:

Summary of experimental paradigm. High memory load trials involved 4 digits (SOA 1s) followed by a discourse prime and picture probe before free recall of digits was prompted. The low load condition involved 1 digit instead of 4.

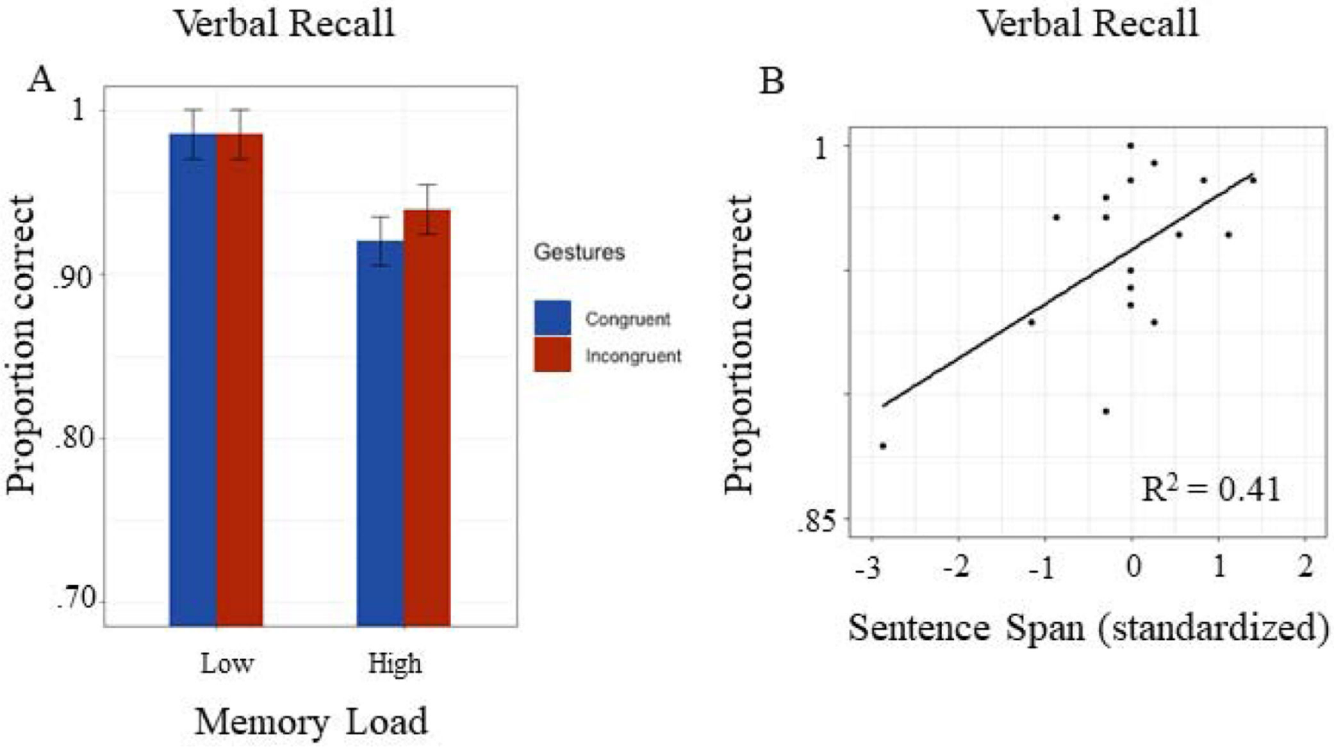

Figure 2.

A) Proportion of correct trials across Load and Video conditions (95% confidence intervals included). B) A simple linear regression between standardized total Sentence Span scores and performance on the secondary WM task (proportion of total trials correct).

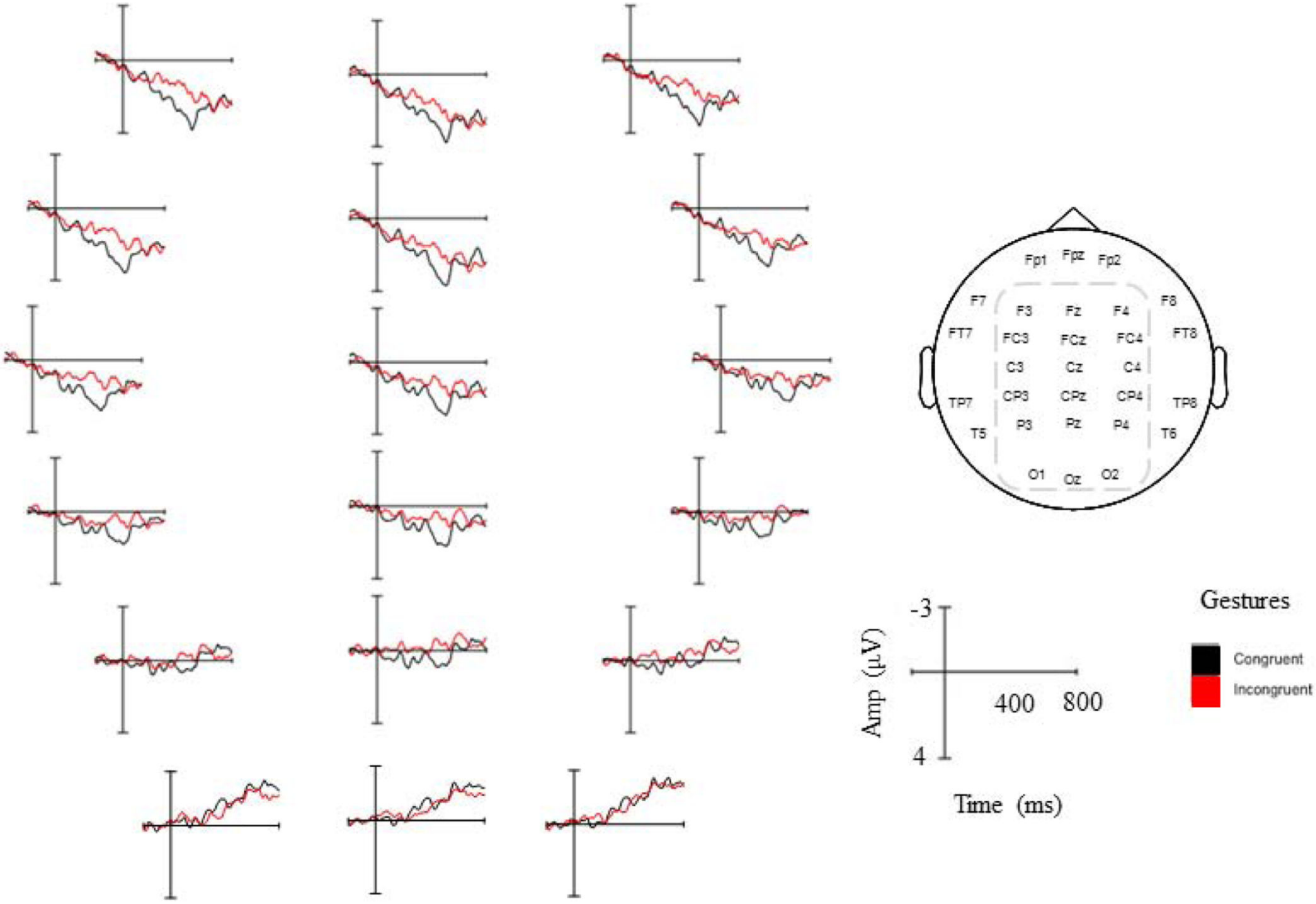

Figure 3.

Averaged ERPs time-locked to first words presented in discourse primes during low memory load trials. Word N400 amplitudes in incongruent videos were more negative on average during trials with mismatching speech and gestures.

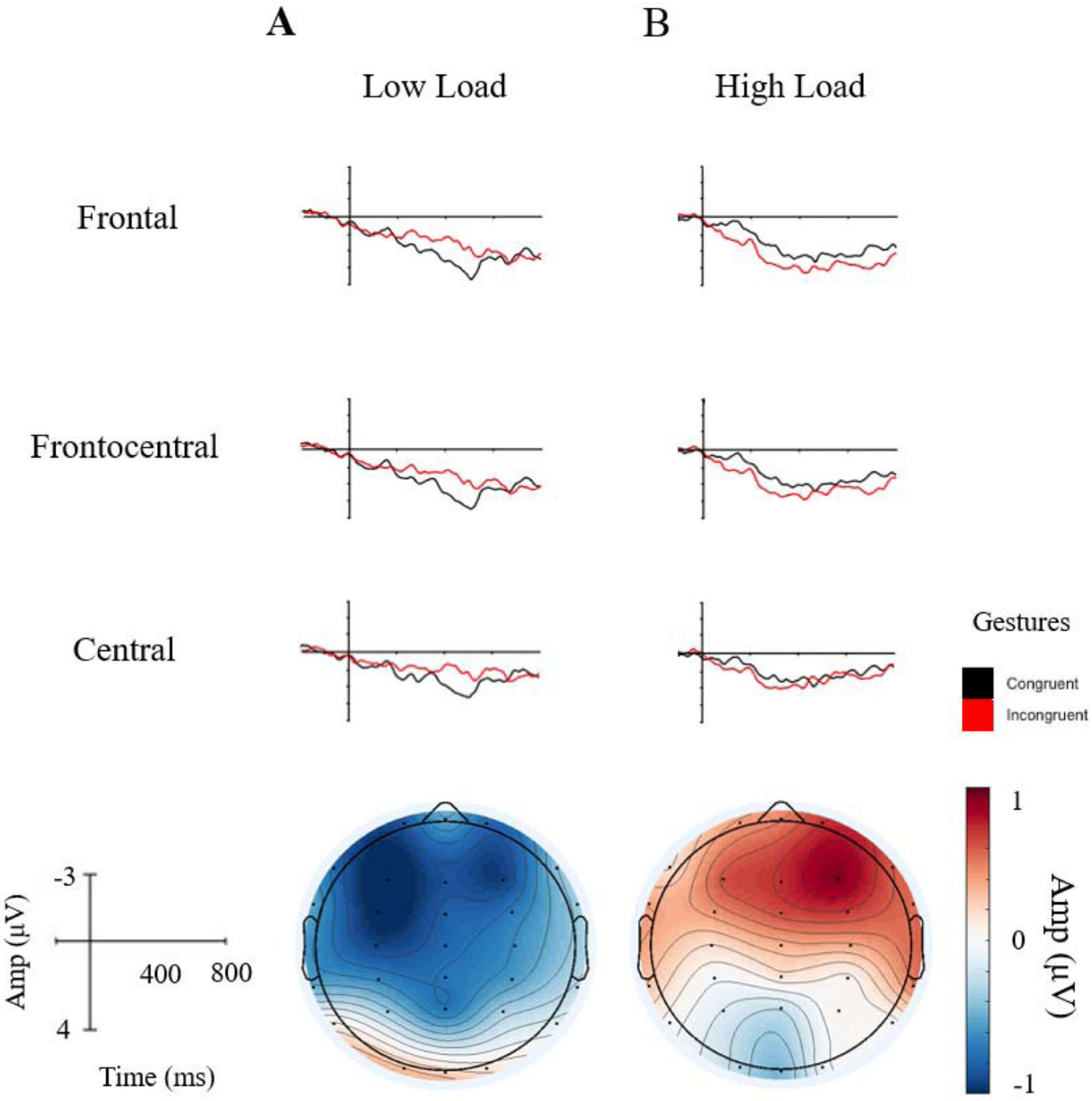

Figure 4.

A) Top: 3 composite ERPs calculated by averaging across 3 central electrodes in the low memory load condition (Frontal: F3,Fz,F4; Frontocentral: FC3, FCz, FC4; Central: C3, Cz, C4). Bottom: Topographical distribution of discourse congruity effect in low memory load trials plotted between 200 and 500ms post word onset. B) Top: 3 composite electrodes calculated in high memory load trials. Bottom: Distribution of discourse congruity effect in high load trials. Red waveforms correspond to initial content words encountered in incongruent videos, while black corresponds to words in the congruent videos.

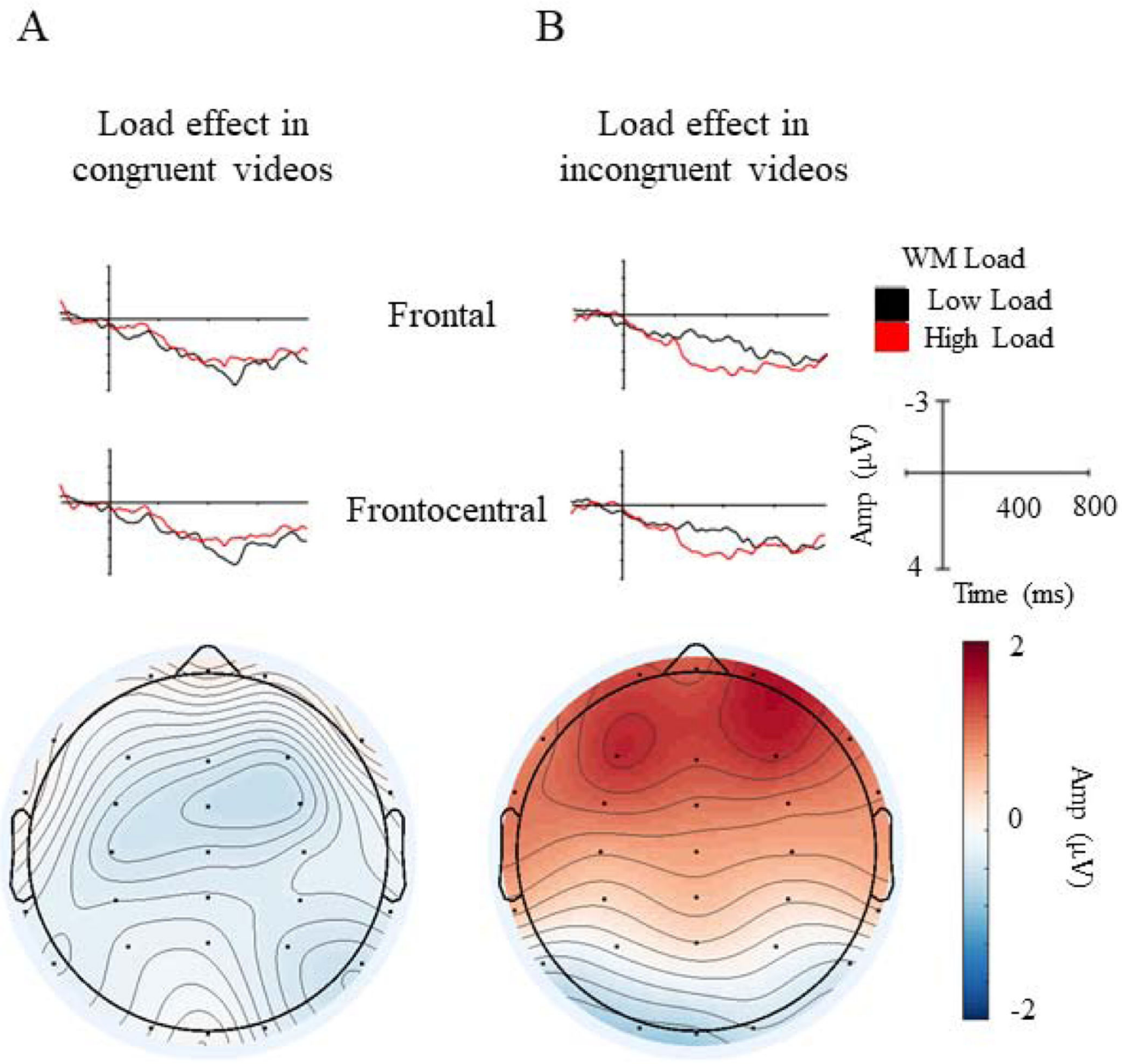

Figure 5.

A) Top: 2 composite ERPs calculated by averaging across 3 central electrodes in the low memory load condition (Fz: F3, Fz ,F4; FCz: FC3, FCz, FC4). Bottom: Topographical distribution of the memory load effect in across video conditions plotted between 200 and 500ms post word onset.