Acta Psychologica 153 (2014) 39–50

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy

Co-speech iconic gestures and visuo-spatial working memory

Ying Choon Wu [a,b,*], Seana Coulson [a,c]

[a] Center for Research in Language, UC San Diego 0526, 9500 Gilman Dr., La Jolla, CA 92093, USA
[b] Swartz Center for Computational Neuroscience, UC San Diego 0559, 9500 Gilman Dr., La Jolla, CA 92093, USA
[c] UC San Diego, Dept. of Cognitive Science 0515, 9500 Gilman Dr., La Jolla, CA 92093, USA

ARTICLE INFO

Article history:
Received 4 September 2013
Received in revised form 2 August 2014
Accepted 8 September 2014
Available online 1 October 2014

PsycINFO classification:
2343 Learning & Memory
2720 Linguistics & Language & Speech
2320 Sensory Perception

Keywords:
Dual task methodology
Gesture comprehension
Iconic gestures
Multi-modal discourse
Spatial cognition
Working memory

ABSTRACT

Three experiments tested the role of verbal versus visuo-spatial working memory in the comprehension of co-speech iconic gestures. In Experiment 1, participants viewed *congruent* discourse primes in which the speaker's gestures matched the information conveyed by his speech, and *incongruent* ones in which the semantic content of the speaker's gestures diverged from that in his speech. Discourse primes were followed by picture probes that participants judged as being either *related* or *unrelated* to the preceding clip. Performance on this picture probe classification task was faster and more accurate after congruent than incongruent discourse primes. The effect of discourse congruency on response times was linearly related to measures of visuo-spatial, but not verbal, working memory capacity, as participants with greater visuo-spatial WM capacity benefited more from congruent gestures. In Experiments 2 and 3, participants performed the same picture probe classification task under conditions of high and low loads on concurrent visuo-spatial (Experiment 2) and verbal (Experiment 3) memory tasks. Effects of discourse congruency and verbal WM load were additive, while effects of discourse congruency and visuo-spatial WM load were interactive. Results suggest that congruent co-speech gestures facilitate multi-modal language comprehension, and indicate an important role for visuo-spatial WM in these speech–gesture integration processes.

1. Introduction

Successful communication often requires multi-modal integration, whereby interlocutors combine information from the verbal channel with visual information about the speaker and the environment. For example, we have documented a speaker uttering the phrase, "manual adjustment lens," to describe a camera while making hand movements that resemble the act of focusing a telephoto lens. The speech and the gesture in this example provide complementary information — and by combining their meanings, it becomes evident that the speaker is describing the lens of a *camera*, and not some other optical device, such as a telescope or a pair of binoculars (Wu & Coulson, 2007). Although prior research indicates that listeners rapidly combine the meaning of speech and iconic gestures in examples such as this (Kelly, Kravitz, & Hopkins, 2004; Ozyurek, Willems, Kita, & Hagoort, 2007; Wu & Coulson, 2010), little is known about the cognitive resources mediating these integration processes.

Here, we focus on *depictive* or *iconic gestures* – that is, those which bear featural similarities to the concepts they represent – as prior research suggests iconic gestures impact semantic aspects of real-time discourse comprehension (Kelly et al., 2004; Ozyurek et al., 2007; Wu & Coulson, 2010). Given that iconic gestures depict visual properties such as shape and size, one obvious possibility is that visuo-spatial processes are important for listeners' success at relating information conveyed in the verbal modality to visual information conveyed in their accompanying gestures. The *visuo-spatial resources hypothesis* is a natural fit with gesture production models suggesting that people gesture in order to convey analogue information in mental images (McNeill, 1992), or to coordinate spatial aspects of a message with the propositional content in their speech (Kita, 2000). Indeed, the gesture production literature suggests that people are more likely to gesture when their speech has spatial or imagistic content (Hadar & Krauss, 1999; Hostetter & Hopkins, 2002; Lavergne & Kimura, 1987; Morsella & Krauss, 2004). Although the *comprehension* of gestures has received much less attention, the visuo-spatial resources hypothesis is consistent with research demonstrating similarities between patterns of brain response to iconic gestures and photographs of real world objects (Wu & Coulson, 2011), as well the finding that listeners use information available in speaker's iconic gestures to help formulate visually specific situation models (Wu & Coulson, 2007, 2010).

However, as their name suggests, co-speech gestures occur almost exclusively in the context of speech — and hence, their semantic analysis may depend heavily on verbal resources. The *verbal resources hypothesis* is in keeping with research suggesting that the meaning

* Corresponding author at: Center for Research in Language, UC San Diego, 9500 Gilman Drive 0526, La Jolla, CA 92093-0526, USA. Tel.: +1 858 869 7014; fax: +1 858 822 5097.
E-mail addresses: ywu@cogsci.ucsd.edu, yingchoon@gmail.com (Y.C. Wu).

http://dx.doi.org/10.1016/j.actpsy.2014.09.002
0001-6918/

of iconic gestures is highly ambiguous, and is determined largely by the meaning of the speech that accompanies them (Hadar & Pinchas-Zamir, 2004; Krauss, Dushay, Chen, & Rauscher, 1995). It is also consistent with neuroimaging research that indicates many of the brain areas mediating the interpretation of gesture, also mediate the interpretation of speech (Straube, Green, Weis, & Kircher, 2012; Willems, Ozyurek, & Hagoort, 2007). Finally, the two hypotheses are not mutually exclusive, as it is quite possible that speech–gesture integration recruits both verbal and visuo-spatial resources.

Given the function of co-speech gestures in real-time language comprehension, working memory (WM) is likely to play an important role in their interpretation. According to the now classic model advanced by Baddeley and Hitch (1974), WM is critical for online processing, serving to temporarily maintain and store perceptual information, and enabling the appropriate updating of representations in long term memory. Notably, WM is widely thought to be comprised of a central controller as well as at least two distinct, modality-specific subsystems dedicated to the maintenance of *visual* information via the visuo-spatial sketch pad, and auditory and *verbal* information via the phonological loop. If listeners tend to preferentially recruit visuo-spatial or verbal resources during speech–gesture integration, we would expect to observe a relationship between the impact of iconic gestures on discourse comprehension and the availability of either visuo-spatial or verbal WM resources (or both).

The present study explored this hypothesis using a two-fold approach. Experiment 1 adopted a correlational method, examining whether there was a relationship between individual differences in measures of either verbal or visuo-spatial WM capacity and individual differences in sensitivity to iconic gestures. In Experiments 2 and 3, we used a dual task paradigm to examine whether taxing different components of WM impact gesture comprehension, suggestive of a causal role for WM in speech–gesture integration. Accordingly, these studies assessed whether participants' ability to utilize the information in co-speech gestures was compromised by manipulating the load on either visuo-spatial (Experiment 2) or verbal (Experiment 3) WM. Finally, Experiment 4 was conducted to ensure that differences in the results of Experiments 2 and 3 did not stem from differences in the difficulty of the secondary verbal and visuo-spatial recall tasks used in those studies.

2 . Experiment 1

To explore the cognitive resources mediating speech–gesture integration, Experiment 1 examined the relationship between individual differences in WM capacity, as measured through verbal and visuo-spatial span tests, and sensitivity to speech–gesture congruency, as measured through a picture probe classification task. Healthy adults viewed short video clips of spontaneous discourse involving iconic gestures, and then classified subsequent photographs of objects and scenes (picture probes) as either related or unrelated to the discourse primes. Control primes were created by swapping the audio and video portions of the original video clips so that the gestures no longer exhibited a semantic relationship to the content of the utterance. Manipulating both the congruency of the discourse prime and the relatedness of the picture probe, we predicted that discourse congruency would have a much greater impact on related than unrelated picture probes.

Our reasoning stemmed from the following premises. First, if related picture probes were classified faster than unrelated ones, irrespective of the preceding discourse prime (a simple relatedness main effect), it could be argued that participants based their responses on the spoken aspects of each utterance, while ignoring the co-occurring gestures. Alternatively, if all probes were classified faster when primed by congruent versus incongruent speech and gestures (a simple main effect of congruency), the difference in classification times could be attributed to perceptual rather than semantic differences between the congruent and incongruent primes — that is, it could be argued that the observed effect was driven primarily by the different types of video editing performed in the construction of the discourse primes as opposed to the degree of semantic coherence between speech and gestures in the two types of trials. On the other hand, if participants were truly sensitive to speech-gesture congruency, we would expect to see much larger congruency effects in response to *related* picture probes, where congruent gestures might aid task performance, than for *unrelated* ones, where neither congruent nor incongruent gestures would be helpful.

Experiment 1 examined the relationship between the impact of speech–gesture congruency on picture probe comprehension and individual differences in verbal and visuo-spatial WM capacity. If visual analysis figured prominently in participants' abilities to conceptually integrate the speaker's speech and gestures, we would expect increasingly superior picture probe comprehension across individuals with increasing visuo-spatial WM capacity. Such an outcome would be likely if space plays a crucial role in the interpretation of iconic gestures – that is, if comprehenders draw mappings between spatially instantiated features of a gesture – such as hand shape, hand location, trajectory, rate of motion, and so forth – and imagistic representations stored in long term memory.

On the other hand, it is possible that verbal processes are more important than visuo-spatial ones for understanding multi-modal discourse. Gestures may, for example, directly activate language-based representations that impact comprehension. Alternatively, increased efficiency at verbal processing may free up attentional resources that allow for greater sensitivity to the semantic properties of gestures. In either case, the *verbal resources hypothesis* predicts a positive relationship between individuals' verbal WM capacity and the impact of iconic gestures on discourse comprehension.

Experiment 1 thus explored the verbal and visuo-spatial resources hypotheses by testing for a relationship between individual differences in verbal and visuo-spatial WM capacity and differences in sensitivity to speech–gesture congruency. Verbal WM capacity was assessed with the Sentence Span test (Daneman & Carpenter, 1980), in which participants hear progressively longer lists of unrelated sentences and recall the sentence final words when cued. Visuo-spatial WM capacity was assessed using the Corsi Block task (Berch, 1998; Milner, 1971), which involves remembering and reproducing progressively longer sequences of block locations. Through multiple regression analysis, the magnitude of the speech–gesture congruency effect on discourse comprehension was modeled using the magnitude of both verbal and visuo-spatial WM capacity as predictor variables. A positive linear relationship between verbal WM capacity and our measure of gesture sensitivity would support the verbal resources hypothesis. Likewise, a positive predictive relationship between visuo-spatial WM capacity and our measure of gesture sensitivity would support the visuo-spatial resources hypothesis.

3 . Methods

3.1. Participants

64 UCSD undergraduates (38 female) gave informed consent and received academic course credit for participation. All participants were fluent English speakers.

3.2. Span tasks

3.2.1. Corsi block task

The Corsi block-tapping task (Milner, 1971) is a widely used test of spatial skills and non-verbal WM. In the computerized variant implemented here, an asymmetric array of nine squares was presented on a monitor. On each trial, some or all of the squares would flash in sequence, though no square flashed more than once. Participants were instructed to reproduce each flash sequence immediately afterwards by clicking their mouse in the appropriate squares in the order that