Real-time Integration Of Gesture And Speech During Reference Resolution

Ellen Campana (ecampana@bcs.rochester.edu)
Department of Brain and Cognitive Sciences, University of Rochester
Rochester, NY 14627, USA

Laura Silverman (lauras@psych.rochester.edu)
Department of Clinical and Social Sciences in Psychology, University of Rochester
Rochester, NY 14627, USA

Michael K. Tanenhaus (mtan@bcs.rochester.edu)
Department of Brain and Cognitive Sciences, University of Rochester
Rochester, NY 14627, USA

Loisa Bennetto (bennetto@psych.rochester.edu)
Department of Clinical and Social Sciences in Psychology, University of Rochester
Rochester, NY 14627, USA

Stephanie Packard (sp002m@mail.rochester.edu)
Department of Brain and Cognitive Sciences, University of Rochester
Rochester, NY 14627, USA

Abstract

There is some disagreement among researchers about the role of gesture in comprehension; whether it is ignored, processed separately from speech, used only when speakers are having difficulty, or immediately integrated with the content of the co-occurring speech. The present experiment provides evidence in support of immediate integration. In our experiment participants watched videos of a woman describing simple shapes on a display in which the video was surrounded by four potential referents: the target, a speech competitor, a gesture competitor, and an unrelated foil. The task was to "click on the shape that the speaker was describing". In half of the videos the speaker used a natural combination of speech and gesture. In the other half, the speaker's hands remained in her lap. Reaction time and eye-movement data from this experiment provide a strong demonstration that as an utterance unfolds, listeners immediately integrate information from naturally co-occurring speech and gesture.

Background

Non-deictic manual gestures are ubiquitous in language production. People produce gesture even when their interlocuters cannot see them (Alibali, Heath and Myers, 2001) and can experience difficulty speaking when their hands are restricted (Graham and Heywood, 1975). There is some consensus that the production system allows different aspects of a single message to be expressed simultaneously over, and distributed between, linguistic utterances and gestures (Alibali, Kita and Young., 2000) However, researchers disagree about the role of gesture in comprehension; whether it is ignored (Krauss, Dushay, Chen and Rauscher., 1995), processed separately from language (Goldin-Meadow and Singer, 2003), used only when speakers are having difficulty (Rauscher, Krauss and Chen, 1996), or immediately integrated with the content of the co-occurring speech (McNeill, Cassell, and McCullough, 1994).

One factor that may have contributed to the lack of consensus is that gesture researchers have tended to focus on answering the question: is gesture communicative? (see Kendon, 1994 for a review). In doing so, they may have confounded two questions that could be considered independently: 1) do speakers intend to communicate using gesture in natural interaction? and 2) do comprehenders benefit from gesture when it naturally co-occurs with speech? For instance, some results suggesting a lack of intentionality have been taken as evidence that listeners' do not use gesture during comprehension (Krauss et al., 1994).

In the current study, we focus only on comprehenders. Our goal is to investigate how and when comprehenders make use of naturally co-occurring gesture and speech (leaving aside for the moment the question of whether speakers intend for gesture to be communicative). There is a growing body of evidence that listeners integrate many forms of extra-linguistic information when comprehending co-occurring speech, including embodiment constraints (Chambers, Tanenhaus, and Magnuson, 2004) gaze (Hanna and Brennan, in prep), and disfluency information (Ferreira, Lau and Bailey, in press; Arnold, Tanenhaus, Altmann, and Fagnano, 2004). Listeners take advantage of this information even when it is not intended by the speakers to be communicative. We would therefore expect listeners to take account of gestures as well. Much of this recent work has been done using the Visual World Paradigm (Tanenhaus et al. 1995; Cooper, 1974). Here, we apply this methodology to determine whether or not comprehenders benefit from gesture that co-occurs with speech

Figure 1: Volunteer modeling the ASL head-mounted eye-tracker that participants wore during the experiment. The eye-tracker recorded the position of the participant's gaze.

Method

We collected data from 10 participants, who were paid $7.50 for an hour of their time. Participants wore a lightweight head-mounted eye-tracker (figure 1) as they watched videos surrounded by four potential referents (figure 2).

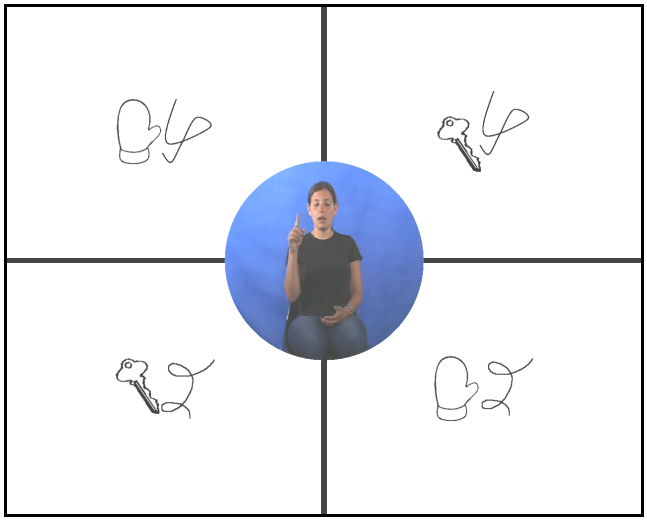


Figure 2: Example screen from the speech + gesture condition. The actor in the center of the screen describes the contents of the top left-hand quadrant, saying "a mitten and a curved line with one loop" as she gestures the shape of the loop.

The participants' task was to "click on what the speaker described." In half of the 14 trials the speaker in the video used a natural combination of speech and gesture (Campana et al, 2004). In the other half of the trials, the speaker's hands remained in her lap.

The stimuli were counterbalanced in two lists such that across participants each set of visual stimuli occurred with both speech + gesture videos and speech – only videos. The speech-only videos were directly modeled on the natural speech + gesture videos and contained identical verbal descriptions of the target objects. For example, for the set of potential referents shown in figure 2, the speaker in the video verbally described the target as "a mitten and a curved line with one loop." This verbal description was the same for both the speech + gesture and the speech-only conditions. However, in the speech + gesture condition the speaker also traced the shape of the loop in the air while in the speech-only condition her hands remained in her lap. The next two sections describe the experimental manipulation we did in more detail, focusing on: 1) the sets of potential referents that participants had to choose between for each trial and 2) the videos, specifically how we elicited natural gestures for this task.

Stimuli: Potential Referents

As outlined in the previous section, for each trial in our experiment there was a set of four potential referents. Each potential referent was comprised of two individual objects or features, one of which was chosen to be easy to describe verbally, and the other of which was chosen to be difficult to describe verbally, but easy to gesture. For each trial there were two of the former and two of the latter. The features were combined in such a way that each potential referent had one of each type of feature. Based on these features, the potential referents were classified as target, speech competitor, gesture competitor, or unrelated foil (Table 1).

Table 1: This table describes the four types of potential referents that appeared in each trial. "Objects" were the contents of a single quadrant on the participants' screen, but they were sometimes composed of several line drawings.

Object	Object Type
	**Target** Consistent with both speech and gesture for the entire description
	**Speech Competitor** Consistent with speech, but not gesture, up until a point near the end of the description
	**Gesture Competitor** Consistent with gesture, but not with speech
	**Foil** Inconsistent with both speech and gesture

The remainder of this section will go into more detail concerning what we meant by these descriptions, taking each one in turn.

The target item was the one that the speaker described – at the end of the description, this is the object that participants were expected to click on. The other types of items were

chosen based on their relationship to the speech and gesture as the speakers description unfolded.

The speech competitor was consistent with the speech up until a certain point in the utterance (we'll call this point the "point of disambiguation," or POD). In the example from figure 2, in which the description was "a mitten and a curved line with one loop," the speech competitor is consistent with the first part of the description, "a mitten and a curved line with", but it is inconsistent with the description as a whole given the final words "one loop." It should also be noted that the speech competitor in our study was always inconsistent with the gesture that the speaker in the video made in the speech + gesture condition.

The gesture competitor was inconsistent with the speech even very early on in description, but it was consistent with the gesture in the gesture + speech condition. In the example from figure 2, the speaker traces the shape of the loop in the speech + gesture condition. The gesture competitor is consistent with this gesture even though it is inconsistent with "a mitten" in the verbal description (it has a key instead).

Finally, the foil was unrelated to both the verbal description and the gesture in the gesture + speech condition. It was, however, systematically related to the other potential referents in the trial. It had one feature in common with the gesture competitor (the one that was not traced in the gesture + speech condition, e.g. the key) and one feature in common with the speech competitor (the one that was not included in the verbal description e.g. the curved line with two loops). The rationale for this was that we wanted to reduce the predictability of the target.

Stimuli: Videos

For each trial, the potential referents surrounded the visual stimuli, which were counterbalanced between two lists such that each list consisted of seven trials with speech + gesture videos and seven trials with speech-only videos. The main distinction between the two video conditions is the presence or absence of gesture. The verbal descriptions of the target objects are identical. In the speech + gesture condition, the speaker uses a natural combination of speech and gesture (Campana et al, 2004); whereas, in the speech-only condition, the speaker's hands remain in her lap.

For the speech + gesture condition, videos were captured in a single session. In the first portion (CC1), a volunteer model was run in our production experiment (Campana et al, 2004). This involved briefly showing her a set of four objects exactly like the visual stimuli described above, except that one was highlighted. Her task was to "get her partner to click on the square that is highlighted." Her partner had a similar display without the highlighting and was sitting at a table opposite her (off-camera) .After completing all of the trials, she was asked to do the entire set of trials a second time (CC), but to be sure to use her hands if she felt like it.

We selected videos from both sets as stimuli for the comprehension experiment. After reviewing all the videos, the spontaneous descriptions (CC1) were chosen over the second descriptions (CC) unless the speech was disfluent. Both of the two lists contained two videos from the first data collection (CC1) and five videos from the second data collection (CC).

In the speech-only condition, videos were captured in a second session. In order to replicate the speech + gesture videos without any hand movements, the volunteer model was provided with both written transcripts and sound clips of the speech + gesture condition prior to recording each trial. Videos for the speech-only condition were chosen on the basis of identical verbal description and similar facial expression, articulation, intonation, and timing.

Results

This experiment yielded both coarse-grained reaction-time data and fine-grained eye-movement data. Both sources of data are relevant for interpreting the results of the experiment, and both support the conclusion that during reference resolution, naturally co-occurring gesture and speech information are immediately integrated. In this section we will first describe the reaction time data and then go on to describe the eye-movement data.

Results: Reaction Time

We analyzed the reaction times with respect to the point of disambiguation (POD) in each individual video. As described in the previous section, the POD is the point in time at which the utterance (just the literal component) goes from being consistent with several potential referents to being consistent with only one. For the trial shown in figure 2 the POD would be the onset of the word "one" in the description "a mitten and a curved line with one loop." The POD is relevant theoretically to the reference comprehension literature and has been used in similar studies investigating the time course of reference resolution (Brown-Schmidt, Campana, and Tanenhaus, 2004; Chambers, Tanenhaus, Eberhard, Filip, & Carlson, 2002, Hanna, Tanenhaus & Trueswell 2003, Eberhard, Spivey-Knowlton, Sedivy & Tanenhaus, 1995; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995).

For our experiment, we chose to analyze reaction times with respect to the POD both because of it's theoretical relevance and because it allows us to interpret reaction times even in the face of variation in length between trials and between tokens in the two conditions (an inevitable consequence of using naturally-produced stimuli). We found that participants who saw the speech + gesture version of a given trial more often clicked on the correct target referent prior to the POD than participants who saw the speech – only version of the same trial ($T_1(9)=4.87$, $T_2(13)=3.21$, $p<.05$).

Results: Eye-movement Data

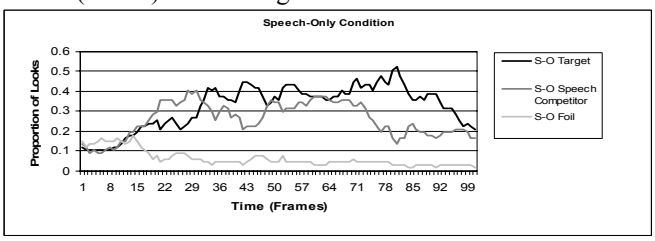
The eye-tracking methodology can provide very fine-grained information about language processing in context. In reference resolution experiments similar to ours, language-driven eye-movements have been observed as early as 250 ms after disambiguating information is encountered, little more than the time required to program and execute the required motor commands (Allopenna, Magnuson, & Tanenhaus, 1998).

In the context of our experiment participants' looks to the target, speech competitor, gesture competitor, and unrelated item can provide information about which entities they are considering as potential referents of the speaker's description.

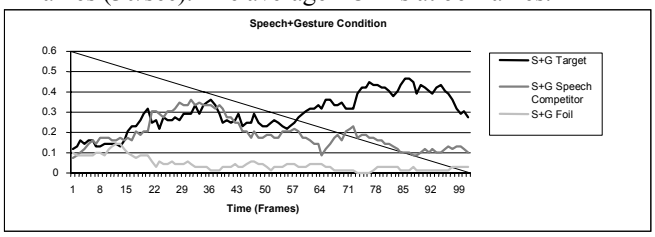
In our experiment we observed that participants who saw the speech + gesture version of a given trial were less likely to look at the speech competitor between the onset of the description and the POD than participants who saw the speech – only version of the same trial ($T_1(9)=2.27$, $T_2(13)=2.64$, $p<.05$).

Graphs of the eye-movements in the two conditions reveal longer-lasting competition from the speech competitor in the speech – only condition (graph 1) compared to the speech + gesture condition (graph 2). By competition we mean that even after looks to the target have surpassed looks to the speech competitor, looks to the speech competitor still remain above baseline (looks to the unrelated foil).

Graph 1: Proportion of looks over time to the target, speech competitor, and foil in the speech-only condition. Time is in frames (30/sec). The average POD is at 62 frames.



Graph 2: Proportion of looks over time to target, speech competitor, and foil in the speech + gesture condition. Time is in frames (30/sec). The average POD is at 66 frames.



Implications

At the beginning of this paper we described several hypotheses that have been advanced in the gesture literature concerning the relationship between gesture and speech during comprehension: 1) gesture could be ignored by comprehenders, 2) gesture could be processed separately and independently from speech, 3) gesture could be used by comprehenders only in situations where speakers seem to be having difficulty, and 4) gesture and speech could be immediately integrated during the process of comprehension. Our experiment provides evidence in support of the 4th hypothesis: immediate integration.

The videos we used as stimuli in our study were naturally-produced. Like most natural examples of this phenomena, the gestures were aligned with or slightly preceded speech (McClave, 1994). In addition, because gestures are holistic and spatial they either encode different features of the referent than speech, or they encode the same features as speech, but with a different timecourse (McNeill, 2000). In our experiment we took advantage of these properties by carefully constructing the original context in which our model produced the reference, and correspondingly the context in which our participants' understood them.

The visual stimuli consisted of potential referents that each had two features that were relevant to the task. In the speech + gesture condition, both of the target's features were described in speech and one was also gestured. Given the information contained in each channel, and the properties of gesture described above, if participants attended to both speech and gesture, and immediately integrated the two sources of information, they would be able to identify the target from among the set of potential referents much earlier than if they attended to just one channel information source. This is, in fact, what we found: participants were able to identify the target faster in the gesture + speech condition than in the speech - only condition. The eye-movement data is also consistent with the hypothesis that integration is immediate, and it provides little data in support of the competing hypotheses. Thus, our results demonstrate that as an utterance unfolds, listeners integrate information from naturally co-occurring speech and gesture.

Future Work

We are currently utilizing the methodology from this study to examine the time course of gesture and speech comprehension in individuals with high-functioning autism and Asperger Syndrome. Individuals with autism have significant impairments in social and communicative domains. Research has established that they experience difficulties comprehending the nonverbal behaviors of others, such as eye-gaze, pointing, and facial expressions (e.g., Klin, Jones, Schultz, Volkmar & Cohen, 2002; Goodhart & Baron-Cohen, 1993; Blair, Frith, Abell & Cipollotti, 2002). This is important since, the ability to decode nonverbal information is critical for appropriate behavior during social interactions. For example, Boyatzis & Sataprasad (1994) examined the relationship between children's peer popularity and their abilities to decode gestures and facial expressions. This study found that non-verbal abilities, and specifically gesture decoding skills were significantly related to peer popularity. Other evidence suggests that people entirely deprived of exposure to gestures (i.e., children who are congenitally blind) have social and communicative impairments that resemble those observed in sighted children with autism (e.g., Brown, 1997). Hobson (1993) has proposed an overlap in the developmental psychopathology of autism and congenital blindness, based on a shared inability to process the outward physical expressions of others' social-emotional experiences. Taken together, these studies suggest an important link between gesture comprehension abilities and social outcomes that may be critical to understanding the social and communicative difficulties observed in autism.

The goal of our future work involves, first establishing whether individuals with autism attend to and successfully decode meaningful gestures in the absence of speech. Next, it involves establishing how and when individuals with autism process gestures that occur in the presence of speech. Based on the findings of the present study, we know that typical adults immediately integrate information that is presented via gesture and speech. We expect that individuals with autism will show a different time-course and pattern of gesture and

speech processing. Specifically, empirical evidence suggests that individuals with autism have difficulties processing information from two modalities. For example, DeGeldger et al., (1991) found that individuals with autism experienced significant difficulty matching speech sounds with their corresponding mouth and lip movements. Similarly, Bryson (1972) found that individuals with autism performed significantly less well on a task where they matched items across verbal and visual modalities compared to a task where they matched objects within modalities. Research on attention also provides evidence that individuals with autism demonstrate an inability to shift attention between modalities (e.g., Courchesne et al., 1994).

Limitations of these studies are that none of them use spontaneous, naturally occurring social stimuli, and none of these studies look at the natural time course of verbal and non-verbal processing in autism. The stimuli used in our study (as described earlier) involve samples of spontaneous instances of gesture and speech collected during a gesture production task. In addition, the eye-tracking technology allows us to examine on-line gesture and speech processing and integration as it occurs.

This research has the potential to provide answers to questions central to understanding social and communicative functioning in autism. Furthermore it has clear implications for treatment. If gesture comprehension in autism is affected by difficulties with cross-modal processing, then interventions can focus on teaching children with autism to attend to gestures in the presence of speech.

Acknowledgements

This research was supported by NIH grant HD-27206 and (NIMH) F31 MH70119. We would like acknowledge several individuals for their hard work and contributions to this research: Rebecca Webb, Hsi-Mei (Betty) Huang, Kathryn McNamara, Kelley Knoch and Cheryl Carmichael.

References

Alibali, M. W., Heath, D. C., and Myers, H. J. (2001). Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen. *Journal of Memory and Language*, **44**, 1-20.

Alibali, M. W., Kita, S., and Young, A. J. (2000). Gesture and the process of speech production: We think, therefore we gesture. *Language and Cognitive Processes*, **15**, 593-613.

Allopenna, P.D., Magnuson, J.S., & Tanenhaus, M.K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping of models. Journal of Memory and Language, 38, 419-439.

Arnold, J.A., Tanenhaus, M.K. Altmann, R.J., & Fagnano, M. (2004). The old and, theee, uh, new: Disfluency and reference resolution. Psychological Science.

Blair, R.J.R., Frith, U., Smith, N., Abell, F. & Cipolotti, L. (2002). Fractionation of visual memory: Agency detection and its impairment in autism. Neuropsychologia, 40, 108-118.

Boyatzis, C. J., & Satyaprasad, C. (1994). Children's facial and gestural decoding and encoding: relations between skills and with popularity. *Journal of Nonverbal Behavior*, 18(1), 37-55.

Brown, R., Hobson, R.P., Lee, A. & Stevenson, J. (1997). Are there "autistic-like" features in congenitally blind children: *Journal of Child Psychology, Psychiatry, and Allied Disciplines, 38,* 693-703.

Brown-Schmidt, S., Campana, E. & Tanenhaus, M.K. (2005) Real-time reference resolution by nave participants during a task-based unscripted conversation. In J.C. Trueswell & M.K. Tanenhaus (eds.), World-situated language processing: Bridging the language as product and language as action traditions. Cambridge, MA: MIT Press.

Bryson, C. (1972). Short-term memory and cross-modal information processing in autistic children. *Journal of Learning Disabilities, 5,* 81-91.

Campana, E., Silverman, L., Tanenhaus, M. K., and Bennetto, L. (2004). Iconic Gesture Production in Controlled Referential Domains. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*. Chicago, IL, August 2004.

Chambers, C.G., Tanenhaus, M.K. & Magnuson, J.S. (2004). Action-based affordances and syntactic ambiguity resolution. Journal of Experimental Psychology: Learning, Memory & Cognition, 30, 687-696.

Chambers, C.G., Tanenhaus, M.K., Eberhard, K.M., Filip, H., & Carlson, G.N. (2002). Circumscribing referential domains during real-time language comprehension. *Journal of Memory and Language*, **47**, 30-49.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. Cognitive Psychology, 6, 84-107.

Courchesne, E., Townsend, J.P., Akshoomoff, N.A., Yeung-Courchesne, R., Press, G.A., Murakami, J.W., Lincoln, A.J., Jamies, H.E., Saitoh, O., Egaas, B., Haas, R.H., & Schreibman, L. (1994). A new finding: Impairment in shifting attention in autistic and cerebellar patients. In *Atypical cognitive deficits in developmental disorders: Implications for brain function*, (ed. S.H. Broman & J. Grafman). Hillsdale, NJ: Erlbaum.

de Gelder, B., Vroomen, J., & van der Heide, L. (1991). Face recognition and lipreading in autism. *European Journal of Cognitive Psychology, 3*, 69-86.

Eberhard, K.M., Spivey-Knowlton, M.J., Sedivy, J.C., & Tanenhaus, M.K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. Journal of Psycholinguistic Research, 24, 409-436.

Ferreira, F., Lau, E.F., & Bailey, K.G.D. (in press). Disfluencies, parsing, and tree-adjoining grammars. Cognitive Science.

Goldin-Meadow, S, and Singer, M. A. (2003). From children's hands to adults' ears: Gesture's role in the learning process. *Developmental Psychology*, **39**, 509-520.

Goodhart, F., & Baron-Cohen, S. (1993). How many ways can the point be made? Evidence from children with and without autism. *First Language*, 13, 225-233.

Graham, J. A., and Heywood, S. (1975). The effects of elimination of hand gestures and of codability on speech

performance. *European Journal of Social Psychology*, **2**, 189-195

Hanna, J.E., Tanenhaus, M.K. & Trueswell, J.C. (2003). The effects of common ground and perspective on domains of referential interpretation. Journal of Memory and Language., 49, 43-61

Hobson, R.P. (1993). *Autism and the development of mind.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Kendon, A. (1994). Do gestures communicate? A review. *Research on Language and Social Interactions*, 27, 175-200.

Klin, A., Jones, W., Schultz, R., Volkmar, F.R. & Cohen, D.J. (2002). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Arch. Gen. Psychiat., 59*, 809-816.

Krauss, R. M., Dushay, R. A., Chen, Y., and Rausher, F. (1995). The communicative value of conversational hand gestures. *Journal of Experimental Social Psychology*, **31**, 533-553.

McNeill, D. (Ed.) (2000). *Language and Gesture.* Cambridge: Cambridge University Press.

McNeill, D., Cassell, J., and McCullough, K.-E. (1994). Communicative effects of speech-mismatched gestures. *Research on Language and Social Interaction*, **27**, 223-237.

Rauscher, F. H., Krauss, R. M. and Chen Y. (1996). Gesture, speech, and lexical access: The role of lexical movements on speech production. *Psychological Science*, **7**, 226-231.

Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., & Sedivy, J.C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, **268**, 1632-1634.