Using Gesture to Resolve Referential Ambiguity During Online Language Processing: Differential Effects of Masking and Noise

Sharice Clough[1], Melissa C. Duff[1], Sarah Brown-Schmidt[2]

[1]Department of Hearing and Speech Sciences, Vanderbilt University Medical Center

[2] Department of Psychology and Human Development, Vanderbilt University

Language processing is incremental, multimodal, and situated in rich visual contexts. Gesture aids language comprehension[1,2], and its onset often proceeds its semantic affiliates in speech[3]. Listeners can use gesture to track referents during complex syntactic processing[4] and resolve ambiguous referents[5] such as homophones[6,7]. Yet, few studies have examined real-time multimodal language processing. We adapted the visual world paradigm[8] to examine how people integrate co-occurring speech and gesture to resolve temporary ambiguity in sentences across three experiments. Experiment 1 examines the use of gesture for predicting the target referent in speech, measured via eye gaze. Experiments 2 and 3 extend this work by examining whether the benefit of gesture interacts with the use of face coverings and noise, respectively.

We tracked participants' eye movements as they viewed a visual scene containing four pictures, randomly assigned to the four corners. A speaker appeared in the center (Fig 1A) and produced subject-verb-object sentences in English (e.g., "The girl will eat the very good sandwich."). The speaker produced either a meaningful gesture (e.g., *sandwich holding* gesture) or a meaningless grooming movement (e.g., arm scratch) on the verb of each sentence (e.g., "will eat"). In experiment 2, the speaker wore a mask on half of the trials. In experiment 3, the audio was noise-vocoded to reduce speech intelligibility on half of the trials. We examined fixations to regions of interest on the screen containing the target (e.g., sandwich), semantic competitor (e.g., apple), and two distractor items (e.g., piano and guitar) during the critical window between onset of the movement stroke and onset of the object in speech (e.g., "sandwich"). This analysis window was offset by 180ms due to the time needed to launch an eye-movement, minus a 20ms baseline to calculate the first-order autocorrelation AR(1)[9]. We used dynamic generalized linear mixed models[10] to predict fixations to the target item (1) or not the target item (0). All models included fixed effects for experimental conditions, AR(1), and time, with random intercepts for participant and item.

Experiment 1 used a within-subject design examining one factor containing two levels: Movement Type (meaningful gesture or meaningless grooming movement). 23 undergraduate students completed 1 session of 240 trials, lasting approximately 45 minutes. The meaningless grooming movement level was dummy coded as the reference. There was a significant effect of movement type ($\beta = 0.57$, $z = 8.48$, $p < .001$) (Fig 1B). Participants were 1.8 times more likely to look at the target item during the critical window (between onset of gesture stroke and mean onset of object in speech) when the speaker produced a gesture compared to when they produced a grooming movement. The significant effects of time ($\beta = 0.48$, $z = 20.96$, $p < .001$) and AR(1) ($\beta = 10.73$, $z = 60.81$, $p < .001$) reflect the tendency of participants to be more likely to fixate the target over time within a trial and the serial dependency of fixation location from one timepoint to another.

Experiment 2 used a within-subject design examining two crossed factors, each with two levels: Movement Type (meaningless grooming movement or meaningful gesture) and Face Covering (unmasked or masked). 35 undergraduate students completed 1 session of 320 trials, lasting approximately 60 minutes. The meaningless grooming movement level and the unmasked face level were dummy coded as the references. There was a significant effect of movement type ($\beta = 0.57$, $z = 17.90$, $p < .001$) (Fig 1C). Participants were 1.8 times more likely to look at the target item during the critical window when the speaker produced a gesture compared to when they produced a grooming movement. There was a significant effect of masking ($\beta = 0.15$, $z = 4.63$, $p < .001$). Participants were 1.2 times more likely to look at the target item during the critical window when the speaker was masked compared to unmasked. Movement type and Masking did not interact. Time and AR(1) were also significant predictors of fixation to target.

Experiment 3 used a within-subject design examining two crossed factors, each with two levels: Movement Type (meaningless grooming movement or meaningful gesture) and Noise Level (clear or 4-band noise-vocoded). 36 undergraduate students completed 1 session of 320 trials, lasting approximately 60 minutes. The meaningless grooming movement level and the clear speech level were dummy coded as the references. There was a significant effect of movement type ($\beta = 0.39$, $z = 10.31$, $p < .001$) (Fig 1D). Participants were 1.5 times more likely to look at the target item during the critical window when the speaker produced a gesture compared to when they produced a grooming movement. There was a significant effect of noise level ($\beta = -24$, $z = -5.91$, $p < .001$). Participants were 79% as likely to look at the target item during the critical window when the speech was noise-vocoded compared to clear. Movement type and Noise did not interact. Time and AR(1) were also significant predictors of fixation to target.

We demonstrated a significant and positive effect of gesture on predicting the target referent, replicated across three experiments using a novel adaptation of the visual world paradigm. We also demonstrated that participants were more likely to fixate on the target item when the speaker was wearing a mask but less likely when the speech was degraded with noise-vocoding, suggesting that availability in lip cues may also modulate visual attention. Although other studies have demonstrated that gesture has an increased benefit in noise[11,12], we found that the effect of gesture did not significantly differ between clear and degraded speech conditions. This raises questions about how listeners weigh and integrate information from different modalities across communication contexts that reflect the rich processing demands of daily life.

Index Terms: eye-tracking, language processing, gesture comprehension, speech-gesture integration, noise, masks

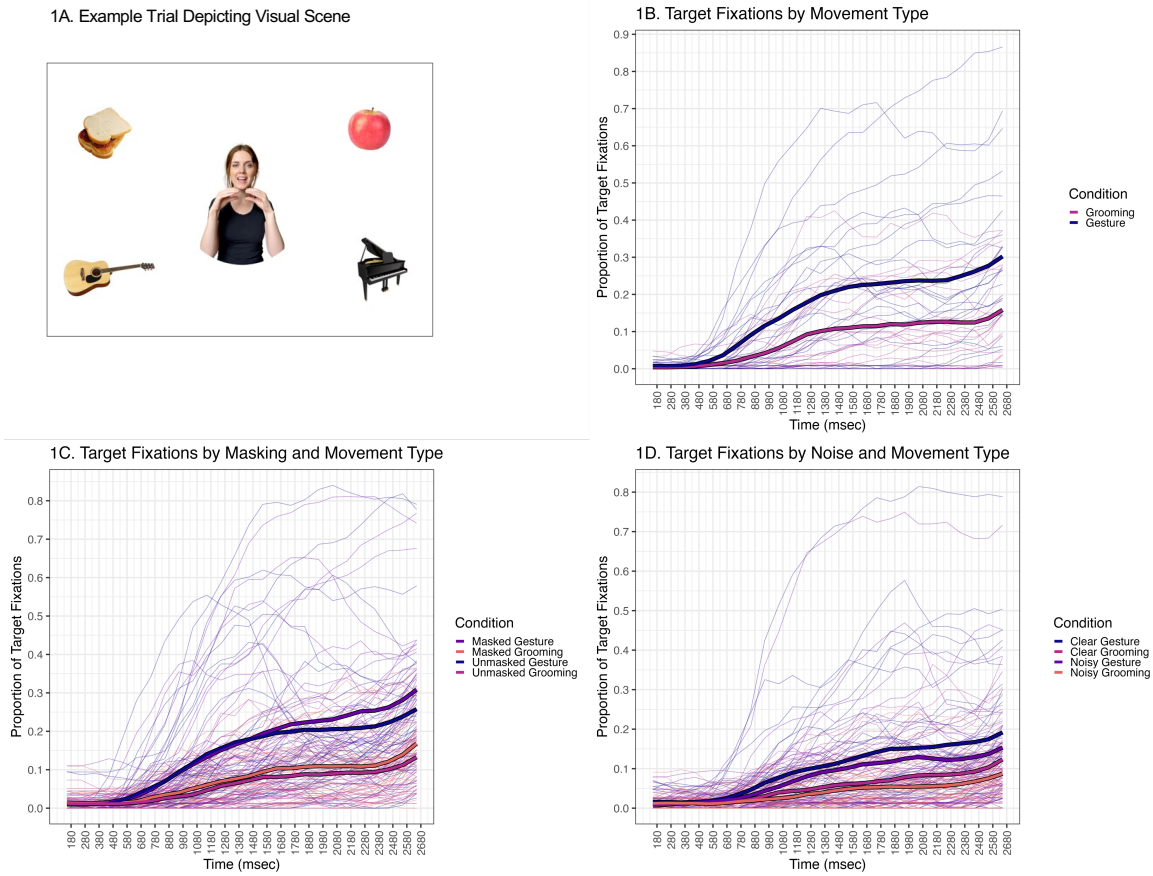

Figure 1A. *Example trial showing visual scene. On this trial, the speaker says, "The girl will eat the very good sandwich," while producing a sandwich-holding gesture on the phrase "will eat."*

Figure 1B. *Proportion of eye fixations to the target item (e.g., sandwich) during the critical window (between onset of the movement stroke and average onset of the object) across grooming movement and gesture trials. 0 msec represents onset of the gesture stroke. The analysis window is offset by 180 msec due to the time needed to launch an eye movement, minus a 20 msec baseline to calculate AR(1). Thick lines represent group means and thin lines represent individual participants.*

Figure 1C. *Proportion of eye fixations to the target item across grooming movement and gesture trials for masked and unmasked speakers.*

Figure 1D. *Proportion of eye fixations to the target item across grooming movement and gesture trials for noise-vocoded and clear speech trials.*

References

1. Hostetter AB. When do gestures communicate? A meta-analysis. *Psychol Bull*. 2011;137(2):297-315. doi:10.1037/a0022128
2. Dargue N, Sweller N, Jones MP. When our hands help us understand: A meta-analysis into the effects of gesture on comprehension. *Psychol Bull*. 2019;145(8):765-784. doi:10.1037/bul0000202
3. Fritz I, Kita S, Littlemore J, Krott A. Multimodal language processing: How preceding discourse constrains gesture interpretation and affects gesture integration when gestures do not synchronise with semantic affiliates ☆. *J Mem Lang*. 2021;117:104191. doi:10.1016/j.jml.2020.104191
4. Theakston AL, Coates A, Holler J. Handling agents and patients: Representational cospeech gestures help children comprehend complex syntactic constructions. *Dev Psychol*. 2014;50(7):1973-1984. doi:10.1037/a0036694
5. Goodrich Smith W, Hudson Kam CL. Children's use of gesture in ambiguous pronoun interpretation. *J Child Lang*. 2015;42(3):591-617. doi:10.1017/S0305000915000045
6. Holle H, Gunter TC. The role of iconic gestures in speech disambiguation: ERP evidence. *J Cogn Neurosci*. 2007;19(7):1175-1192. doi:10.1162/jocn.2007.19.7.1175
7. Obermeier C, Dolk T, Gunter TC. The benefit of gestures during communication: Evidence from hearing and hearing-impaired individuals. *Cortex*. 2012;48(7):857-870. doi:10.1016/j.cortex.2011.02.007
8. Tanenhaus MK, Spivey-Knowlton MJ, Eberhard KM, Sedivy JC. Integration of visual and linguistic information in spoken language comprehension. *Science (1979)*. 1995;268(5217):1632-1634. doi:10.1126/science.7777863
9. Boff KR, Kaufman Lloyd, Thomas JP. *Handbook of Perception and Human Performance*. (Boff KR, Kaufman Lloyd, Thomas JP (James P, eds.). Wiley; 1986.
10. Cho SJ, Brown-Schmidt S, Lee W yeol. Autoregressive Generalized Linear Mixed Effect Models with Crossed Random Effects: An Application to Intensive Binary Time Series Eye-Tracking Data. *Psychometrika*. 2018;83(3):751-771. doi:10.1007/s11336-018-9604-2
11. Drijvers L, Özyürek A. Visual context enhanced: The joint contribution of iconic gestures and visible speech to degraded speech comprehension. *Journal of Speech, Language, and Hearing Research*. 2017;60(1):212-222. doi:10.1044/2016_JSLHR-H-16-0101
12. Holle H, Obleser J, Rueschemeyer SA, Gunter TC. Integration of iconic gestures and speech in left superior temporal areas boosts speech comprehension under adverse listening conditions. *Neuroimage*. 2010;49(1):875-884. doi:10.1016/j.neuroimage.2009.08.058