Cognition 133 (2014) 692–697

Contents lists available at ScienceDirect

Cognition

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

Short Communication

Social eye gaze modulates processing of speech and co-speech gesture

Judith Holler [a,b,*], Louise Schubotz [a,f], Spencer Kelly [c], Peter Hagoort [a,e], Manuela Schuetze [a], Aslı Özyürek [d,a]

[a] Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
[b] University of Manchester, School of Psychological Sciences, Coupland Building 1, M13 9PL Manchester, UK
[c] Colgate University, Psychology Department, Center for Language and Brain, Oak Drive 13, Hamilton, NY 13346, USA
[d] Radboud University, Centre for Language Studies, Erasmusplein 1, 6525HT Nijmegen, The Netherlands
[e] Radboud University, Donders Institute for Brain, Cognition and Behaviour, Montessorilaan 3, 6525 HR Nijmegen, The Netherlands
[f] Max Planck Institute for Demographic Research, Konrad-Zuse-Straße 1, 18057 Rostock, Germany

ARTICLE INFO

Article history:
Received 29 March 2013
Revised 11 August 2014
Accepted 13 August 2014

Keywords:
Language processing
Co-speech iconic gesture
Eye gaze
Recipient status
Communicative intent
Multi-party communication

ABSTRACT

In human face-to-face communication, language comprehension is a multi-modal, situated activity. However, little is known about how we combine information from different modalities during comprehension, and how perceived communicative intentions, often signaled through visual signals, influence this process. We explored this question by simulating a multi-party communication context in which a speaker alternated her gaze between two recipients. Participants viewed speech-only or speech + gesture object-related messages when being addressed (direct gaze) or unaddressed (gaze averted to other participant). They were then asked to choose which of two object images matched the speaker's preceding message. Unaddressed recipients responded significantly more slowly than addressees for speech-only utterances. However, perceiving the same speech accompanied by gestures sped unaddressed recipients up to a level identical to that of addressees. That is, when unaddressed recipients' speech processing suffers, gestures can enhance the comprehension of a speaker's message. We discuss our findings with respect to two hypotheses attempting to account for how social eye gaze may modulate multi-modal language comprehension.

1. Introduction

Face-to-face communication is a multi-modal activity often involving multiple participants. Despite this, language comprehension has typically been investigated in uni-modal (i.e., just speech) and solitary (i.e., one listener) contexts. Here, we investigate language comprehension in the context of two other modalities omnipresent during face-to-face communication, co-speech gesture and eye gaze. Moreover, we explore the interplay of these modalities during comprehension in a dynamic context, where a speaker's eye gaze switches between two interlocutors, rendering them sometimes directly addressed, and sometimes relatively unaddressed, a typical characteristic of multi-party conversation.

Despite the uni-modal focus of traditional approaches to language comprehension, recent years have seen an increase in studies considering language as consisting of both speech and co-speech gestures. These studies have provided behavioural and neural evidence that co-speech gestures are processed semantically and integrated with speech during comprehension (e.g., Holle & Gunter, 2007;

* Corresponding author at: Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands. Tel.: +31 24 3521268.
E-mail address: judith.holler@mpi.nl (J. Holler).

http://dx.doi.org/10.1016/j.cognition.2014.08.008
0010-0277/

Kelly, Kravitz, & Hopkins, 2004; Kelly, Özyürek, & Maris, 2010; Willems, Özyürek, & Hagoort, 2007, 2009; Özyürek, Willems, Kita, & Hagoort, 2007; Wu & Coulson, 2005, 2007; Yap, So, Melvin Yap, Tan, & Teoh, 2011). Research has further shown that this integration process is not always automatic but sensitive to the perceived intentional coupling of gesture and speech—such as when observing a gesture performed by one person accompanying speech produced by another (Kelly, Creigh, & Bartolotti, 2010; Kelly, Ward, Creigh, & Bartolotti, 2007). A question that remains is whether the processing of multi-modal utterances is also modulated by social cues integral to the communicative situation, such as when a speaker's gaze conveys information about whom he/she is addressing.

Due to the saliency of the sclera in the human eye in contrast to other primate species, gaze is a powerful social cue in human interaction (e.g., Goodwin, 1981; Kendon, 1967; Rossano, 2012; Senju & Johnson, 2009). While some studies have investigated speech-gesture comprehension in the presence of gaze, they have done so without manipulating gaze direction as an independent cue (e.g., Kelly et al., 2004; Straube, Green, Jansen, Chatterjee, & Kircher, 2010; Wu & Coulson, 2007a).

One exception is a recent study (Holler, Kelly, Hagoort, & Özyürek, 2012) in which a speaker alternated her gaze between two recipients, rendering one of them addressed and the other unaddressed during each message she communicated. This study focused on how the manipulation of social gaze would influence the comprehension of uni-modal ("*she trained the horse*") and bi-modal ("*she trained the horse*" + whipping gesture) utterances. Following each utterance, participants saw a target word onscreen, matching either the preceding speech (speech-related targets [e.g., to train]) or the preceding gesture conveying complementary information (gesture-related targets [e.g., to whip]). Unaddressed recipients responded more slowly than addressees to gesture-related target lures following the bi-modal utterances. However, their response times for the uni-modal (speech-only) conditions did not differ from those of addressees (neither for the speech- nor the gesture-related targets). Participants in this study were required to focus their attention on the verbal modality, firstly, by making judgements about the *speech* they heard in the preceding video, and, secondly, by responding to *verbal* targets displayed onscreen. Explicitly focusing participants' attention on speech in this way might simply not be suitable for uncovering processing differences relating to the speech modality.

The present study uses a visually focused paradigm that avoided explicit attention to the preceding speech, or to words onscreen, to allow us to better observe potential differences in addressed and unaddressed recipients' processing of bi-modal messages (speech + gesture) *and* the processing of speech when speech is the only modality carrying semantic information. Like Holler et al. (2012), we simulated a triadic communication setting. Participants watched a speaker who was filmed in such a way that she appeared to be looking either at them or at another recipient, conveying speech-only or speech + gesture utterances referring to objects (e.g., "*he prefers the laptop*"). The gestures depicted a typical feature of the object mentioned (e.g., a typing gesture). Each message was followed by two object pictures and participants indicated which of these matched the preceding message as a whole. Thus, instead of probing participants' processing of *either* the speech-related *or* the gesture-related utterance components (Holler et al., 2012), the present study assesses comprehension of the message *overall* rather than probing its separate components.

We have two hypotheses regarding the influence of social eye gaze on multi-modal (speech + gesture) message comprehension. The *Parallel Attenuation Hypothesis* states that social eye gaze direction affects the processing of information in the speech and gesture modalities in a parallel fashion. Schober and Clark (1989) have shown that overhearers process speech less well than addressees in contexts without visual access to the speaker. In contexts that do provide such visual access, gaze direction is an important indicator of communicative intent. Semantic information provided by a speaker who averts her gaze to look at someone else may thus be perceived as intended for this other person. In face-to-face communication, unaddressed recipients may thus not only process speech less well than addressees, but their processing of co-speech gestures may be attenuated, too. If this hypothesis holds, we would expect overall message comprehension (i.e., including speech + gesture utterances) to be less efficient for unaddressed than for addressed recipients.

Alternatively, the *Cross-modal Enhancement Hypothesis* states that social eye gaze direction influences the processing of speech and gesture differently: When unaddressed recipients' speech processing suffers, gesture does not. This effect may be due to the fact that, when the speaker's eye gaze is averted, gesture is still available for/directed at the unaddressed recipient (since, in triadic communication, speakers tend to produce gestures in front of their body visually accessible for all participants, Özyürek, 2002). Access to gestural information might thus enhance unaddressed recipients' speech comprehension, resulting in addressed and unaddressed recipients comprehending the overall message equally well.

The present study aims to tease apart which of these hypotheses may best explain how, in the context of gaze-directional addressing (Lerner, 2003), recipients comprehend multi-modal language in a pragmatically richer context than has been investigated traditionally.

2. Method

2.1. Participants

32 right-handed, native German speakers (16 female) from Rostock University (tested at the Max Planck Institute for Demographic Research in Rostock) participated in the experiment (mean age 24.5 yrs) and were financially compensated (€10).

2.2. Design

We used a 2×2 within-participants factorial design, manipulating gaze direction of the speaker (direct gaze/

addressed recipient condition vs. averted gaze/unaddressed recipient condition) and modality of presentation (speech-only vs. speech + gesture).

2.3. Materials

All stimuli were presented (and responses recorded) in Presentation® software (www.neurobs.com) using a 15″ computer monitor and Sennheiser® closed-cup headphones.

2.3.1. Video clips

160 short sentences (canonical SVO structure) spoken by a female German actor were video-recorded. The sentences always referred to an object, e.g. *"he prefers the laptop"* (*"er bevorzugt den Laptop"*). Each sentence was recorded with or without gesture, and with gaze being direct or averted. The gestures always provided information about the object's shape, size or function (e.g., typing, Fig. 1) and accompanied the noun phrase of the sentence. The actor depicted these object features in a way that felt natural to her.

To avoid possible order effects during recording, for half of our stimuli, the actor first produced the stimuli for the direct gaze condition, followed by those for the averted gaze condition, and for the other half, the order was reversed. To ensure that our stimuli did not differ systematically between the two gaze conditions, we also showed each gesture video—in the absence of speech and with the head obscured—to a separate set of 12 participants and asked them to rate (1–7 Likert-scale) how well each gestural depiction matched the object it was meant to represent (verbal object label was displayed onscreen following the gesture-clip). A Mann–Whitney U-test showed that the gestures in the two gaze conditions depicted the objects equally well, $U = 13.00$ (sample size $n_1 = 6$, $n_2 = 6$), $p = .485$. We also subjected 20% of our speech-only stimuli (random selection) to an acoustic analysis in Praat (Boersma & Weenink, 2014) which confirmed that, despite the slight head turn the speaker performed (theoretically, the different positioning of the larynx could have influenced the loudness of speech), the average intensity of the speech did not differ in the two gaze conditions, $t(31) = 1.23$, $p = .229$.

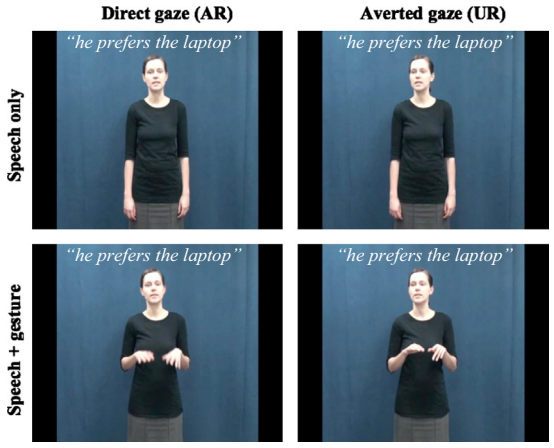

Fig. 1. The four different versions of video stimuli employed in the present study (the speech that accompanied each of the videos is marked by inverted commas). AR = addressed recipient, UR = unaddressed recipient.

Fig. 2. Example of a pair of object pictures.

To ensure that the gestures unambiguously referred to the *objects* mentioned, verbs were non-action verbs (since a typing gesture accompanying the sentence "he types on the laptop" could refer to both "typing" and "laptop") including "prefer" ("bevorzugen"), "like" ("mögen"), and "see" ("sehen").

Our manipulation of gaze direction and modality of presentation resulted in four versions of each item: 1. direct gaze (addressed recipient) speech-only, 2. direct gaze (addressed recipient) speech + gesture, 3. averted gaze (unaddressed recipient) speech-only, and 4. averted gaze (unaddressed recipient) speech + gesture (Fig. 1).

Each participant saw each of the 160 video clips (576 × 576 pix; 16.74 × 17.73 cm onscreen) in one of the four conditions, resulting in 160 experimental trials per participant (40 per condition), plus 24 filler trials (where, for variation, the gestures depicted actions rather than objects).

2.3.2. Object pictures

320 pictures, identical in size (400 × 400 pix; 12.39 × 12.39 cm), were edited in Adobe Photoshop® to show one object (in colour) on white background. 160 pictures showed objects mentioned in the stimulus sentences (e.g., a laptop), and 160 pictures were distractors (e.g., a towel) (Fig. 2)[1] (across participants, each picture-pair was shown with all four video-clips). Two native German speakers were informally presented with all pictures to ensure they were readily recognisable as the intended objects.

2.4. Procedure

Participants watched video-clips of the speaker who, they were told, had been asked to spontaneously create short messages based on line drawings and words displayed on a laptop screen (positioned out of shot, looked at by the speaker before each utterance). Participants were

[1] Our original design involved an additional manipulation concerning the relatedness of the two pictures in that, for half of them, the gesturally depicted information matched both the target and the distractor objects' affordances (e.g., the action depicted by the gesture in Fig. 1 would match piano [playing] and laptop [typing]), but for the other half of picture parings, it did not (e.g., laptop and towel). However, because we found no interaction effects related to this manipulation in either our error or our RT data (all Ps >.3) we collapsed across these two types of picture pairings and report the combined data only.

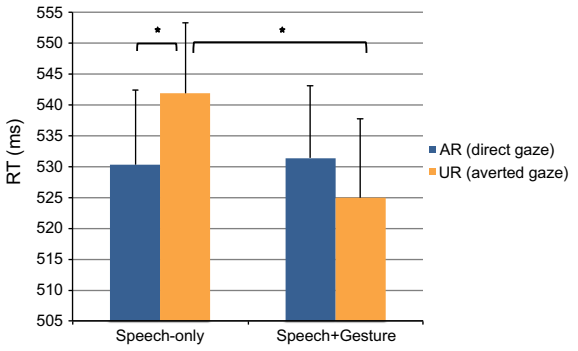

Fig. 3. Addressed recipients' (AR) and unaddressed recipients' (UR) mean reaction times (ms) in the speech-only and speech + gesture conditions (error bars represent SE).

also told that during the recordings, a second person had been sitting in the same room diagonally across from the speaker (all participants seemed to indeed imagine this second recipient as having been present, as indicated through post-experiment ratings of his/her personality characteristics). The speaker was supposedly instructed to sometimes address this (fictitious) participant (averted gaze condition), and sometimes the other (actual) participant via a video-camera positioned straight across from her (direct gaze condition). Following each video-clip (100 ms interval), participants saw two object pictures onscreen and were asked to indicate via button press which of the pictures matched the speaker's preceding message (in speech or gesture) (left button = left-hand picture, right button = right-hand picture; picture location on screen was counterbalanced across participants). Participants' reaction times (calculated as the difference between target picture onset and button press, in milliseconds) and response accuracy were recorded (followed by a fixation cross presented for 2–5 s before the next trial began). Before the experiment proper, participants completed six practice trials.

Participants were asked not to avert their gaze from the screen and were video-recorded (with their consent) during the entire experiment to allow for subsequent checks—these showed that everyone adhered to the instructions.

3. Results

Six trials from two participants were excluded from the analysis due to a technical error. An alpha-value of .05 was used throughout. All *p*-values reported are two-tailed.

3.1. Reaction times

For the analysis of reaction times, we excluded outliers (2.5 SD) and all incorrect responses (83 in total = 1.62% of trials). Fig. 3 shows the data for the 2 (gaze: direct vs. averted) × 2 (modality: speech-only vs. speech + gesture) repeated measures ANOVA. The main effect of gaze was not significant, $F(1,31) = .464$, $p = .501$, $\eta_p^2 = .015$, while the main effect of modality approached significance, $F(1,31) = 3.431$, $p = .074$, $\eta_p^2 = .100$. In addition, there was a significant interaction of gaze and modality, $F(1,31) = 5.947$, $p = .021$, $\eta_p^2 = .161$.[2]

In line with our hypotheses, we calculated two a priori contrasts (using paired-samples *t*-tests), comparing addressed and unaddressed recipients in their processing of uni-modal speech-only utterances and bi-modal speech + gesture utterances. These showed that unaddressed recipients were significantly slower than addressees at processing speech-only utterances, $t(31) = 2.547$, $p = .016$, $r = .085$, but that unaddressed and addressed recipients did *not* however differ in their processing of speech + gesture utterances, $t(31) = 1.112$, $p = .275$, $r = .047$.

Because the interaction was significant and inspection of the data encouraged additional comparisons not captured by our contrasts, we carried out two further standard post hoc tests (using a Bonferroni-adjusted alpha-level of .0125). These revealed that, while addressed recipients did not differ in their processing of speech-only and speech + gesture utterances ($t(31) = .200$, $p = .843$, $r = .008$), unaddressed recipients did—they responded more slowly to speech-only compared to speech + gesture utterances, $t(31) = 2.930$, $p = .006$, $r = .121$.

3.2. Errors

Our 2 (gaze: direct vs. averted) × 2 (modality: speech-only vs. speech + gesture) repeated measures ANOVA (based on the Adjusted Rank Transform Test for non-normally distributed data; Leys & Schumann, 2010) on error percentages revealed a significant modality effect ($F(31) = 7.257$, $p = .011$, $\eta_p^2 = .190$), with participants making fewer errors in the speech + gesture conditions (*Md* = 0.000, *Range* = 5.000) than in the speech-only conditions (*Md* = 3.000, *Range* = 7.500). No other effects were significant (gaze: $F(31) = 0.853$, $p = .363$, $\eta_p^2 = .027$; gaze × modality: $F(31) = 1.112$, $p = .300$, $\eta_p^2 = .035$), and neither were the a priori contrasts for the speech-only ($z = 1.088$, *N-ties* = 16, $p = .277$, $r = .192$) and speech + gesture conditions ($z = .054$, *N-ties* = 16, $p = .957$, $r = .010$).

4. Discussion

This study investigated the interplay of multiple communicative modalities during language comprehension by experimentally simulating a socially dynamic, multiparty communication setting. Specifically, we asked how addressed and unaddressed recipients, as signaled through the speaker's gaze, process speech with and without iconic gestures. The response time data revealed that unaddressed recipients were significantly slower than addressed recipients when processing speech alone. This finding is in line with past research showing that unaddressed recipients (in an overhearer role) process speech less well than addressed recipients in the absence of mutual visibility (Schober & Clark, 1989). Our results show

[2] A very similar pattern of results was obtained with an analysis by items: Main effect of gaze, $F(1,159) = .216$, $p = .643$, $\eta_p^2 = .001$; main effect of modality, $F(1,159) = 2.674$, $p = .104$, $\eta_p^2 = .017$; interaction of gaze and modality, $F(1,31) = 3.985$, $p = .048$, $\eta_p^2 = .024$.

that this pattern also holds for face-to-face contexts (at least for side-participants, Clark & Carlson, 1982).[3] Further, our findings reveal that the processing of linguistic information is not only influenced by concurrent *referential* speaker-gaze to objects in the immediate surrounding (e.g., Hanna & Brennan, 2007; Knöferle & Kreysa, 2012; Staudte & Crocker, 2011) but also by a speaker's *social* gaze. And, finally, as predicted, when recipients are not explicitly asked to make judgements about *verbal* targets in relation to *verbal* components of utterances (Holler et al., 2012), differences in how addressed and unaddressed recipients comprehend speech-only utterances do indeed emerge.

Crucially, our results show that unaddressed recipients processed bi-modal utterances significantly faster than uni-modal ones, while no such difference was found for addressees. Apparently, unaddressed recipients significantly benefitted from the gestures, allowing them to perform at the same level as addressees when matching the target pictures to preceding bi-modal utterances. These findings are in line with the *Cross-modal Enhancement Hypothesis*. When speaker-gaze is averted, processing of speech suffers but gesture does not, thus benefitting overall message comprehension.

The cognitive processes underlying the cross-modal enhancement effect in the present study may have resulted from a number of different mechanisms. One possibility is that gestures were semantically integrated with the verbal information, thus leading to a richer, more unified mental representation of the concept of 'laptop', for example. Alternatively, they may have lead to a stronger memory trace due to receiving related information from two different input streams (visual and verbal), with this information being associated but stored *separately*, *not* as a unified representation (much like a dually-coded representation à la Paivio (1986)).[4] Both of these interpretations are plausible given that iconic gestures can prime linguistic concepts during comprehension (Wu & Coulson, 2007b; Yap et al., 2011). One could also argue that the enhancement effect did not happen at the semantic level but that gestures, as mere visual movements, simply enhanced the attention to speech and thus the subsequent memory for the message (however, note that this would stand in contrast to some earlier findings showing that, while gestures that map semantically onto the information in the accompanying speech do influence speech memory, gestural movements not semantically related to speech, such as beats and incongruent gestures, do not (Feyereisen, 1998; Kelly, McDevitt, & Esch, 2009)). All these possibilities are compatible with our general finding that speech-gesture utterances are comprehended differently by addressed and unaddressed recipients; while unaddressed recipients process speech less well, gestures help their language comprehension, thus leading to a more enhanced representation of the event than when receiving information from speech alone, in line with the *Cross-modal Enhancement Hypothesis*.

The finding that co-speech gestures are beneficial when speech processing suffers nicely complements earlier research showing similar effects when the speech signal is not easily audible due to concurrent multi-speaker babble (Obermeier, Dolk, & Gunter, 2012). The present findings show that co-speech gestures are not only beneficial when the physical perception of speech is problematic, but also when pragmatic context leads to reduced processing of speech—despite the physical speech signal remaining intact. Whether gestures aid comprehension by recruiting the same cognitive mechanism in the context of being an unaddressed recipient and in the context of physically degraded speech is an interesting avenue for future research.

Our error data showed a main effect of modality—both addressed and unaddressed recipients made fewer errors after perceiving speech accompanied by gestures than speech alone. This corroborates past research showing that gestures can help language comprehension (e.g., Kelly, Barr, Church, & Lynch, 1999). It may seem surprising that we did not find a main effect of modality in our reaction time data for addressed recipients—thus contrasting with some earlier studies, such as Kelly et al. (2010, see experimental results referred to in footnote 2). However, note that in this same study by Kelly et al., no effect on error rates was found. This precedence demonstrates that differences between bimodal versus unimodal conditions may be detected using one measure but not another. It is also important to note that our study differs from earlier ones (e.g., Kelly et al., 2010) by using an off-line rather than an on-line measure of comprehension, and by including additional social information in our stimuli, such as the speaker's face. Finally, it is also possible that the lack of a main effect of modality in our addressed recipients' reaction time data may be due to a floor effect around 525/530 ms obscuring the difference between addressed recipients' responses in the bimodal and unimodal conditions (while this difference *does* appear for unaddressed recipients, however, due to their longer processing times in the speech-only condition).

In our study, gesture orientation was kept constant when eye gaze direction changed and this could have maintained attention to gesture in the unaddressed recipient condition. One reason for keeping gesture orientation constant was ecological validity. Previous research has

[3] Note that Schober and Clark's (1989) study showed that one crucial factor leading to the better comprehension of speech by addressed than unaddressed recipients (in their case, overhearers) was that addressed recipients had the chance to interact and thus to ground information with the speaker, whereas unaddressed recipients did not. Here, we found a difference in speech comprehension for addressed and unaddressed recipients despite neither being able to interact and ground with the speaker. Thus, future research employing an interactive paradigm might reveal even more pronounced effects.

[4] Theoretically, it could be argued that our participants may have shown improved performance in the bi-modal as compared to the uni-modal condition not because they processed speech accompanied by gestures better than speech alone, but because they processed gestures instead of speech. That is, participants might have had an easier time matching *visually* depicted information (e.g., a gesture for typing) to a *visual* image of a laptop than matching the word laptop to the visual image of a laptop. However, the following reason speaks against this assumption. As mentioned in Section 2.3.2, our original design contained both pairings of pictures that did and that did not share the gesturally depicted feature of the object mentioned. If participants had processed predominantly the gestures on their own rather than integrating them with the accompanying speech, they should have made more errors in those cases in which the gesture matched both rather than just one of the objects depicted (i.e., interference vs. facilitation trials). However, this was not the case.

shown that, in triadic communication, speakers perform their gestures in front of their body, visually accessible for all participants rather than moving them together with their eye gaze (Özyürek, 2002). Further, the change in gaze direction would have been confounded with perceiving the gestures from different visual angles, making it impossible to determine whether gaze direction itself modulates the comprehension of speech and gesture. Exploring the interplay of gesture orientation and gaze direction requires future research.

In conclusion, the present study has brought together three different modalities—speech, eye gaze and hand gesture—in a language comprehension paradigm, advancing our understanding of how perceived communicative intent, as signaled through a speaker's social gaze, influences the interplay of semantic modalities during comprehension in a face-to-face-like setting. The findings are striking since we have shown that the ostensive cue of eye gaze has the power to modulate how different recipients process speech and co-speech gestures. The next step is to investigate this interplay of modalities in even more situated and interactive settings. In the situated, triadic communication setting simulated here, the gestural modality can come to the aid of unaddressed recipients—when speech processing suffers, gestures help.

Acknowledgments

We would like to thank the Max Planck Institute for Demographic Research in Rostock for their provision of testing space and financial support for LS during part of this research project, the Radboud University Nijmegen for financial support in the form of participant payment, Natalie Sebanz, Günther Knoblich, Ivan Toni, Idil Kokal, and members of the Neurobiology of Language Department and the Gesture & Sign Language group at the Max Planck Institute for Psycholinguistics for helpful feedback during discussions of this study, Ronald Fischer and Nick Wood for assistance with programming and video editing, and the European Commission for funding this research (JH was supported through a Marie Curie Fellowship #255569, and through ERC Advanced Grant #269484INTERACT).

References

Boersma, P., & Weenink, D. (2014). Praat: doing phonetics by computer [Computer program]. Retrieved from http://www.praat.org/.

Clark, H. H., & Carlson, T. B. (1982). Hearers and speech acts. *Language, 58*, 332–373.

Feyereisen, P. (1998). Le rôle des gestes dans la mémorisation d'énoncés oraux. In S. Santi, I. Guaïtella, C. Cavé, & G. Konopczynski (Eds.), *Oralité et gestualité. Communication multimodale, interaction. Actes du colloque Orage 98* (pp. 355–360). Paris: L'Harmattan.

Goodwin, C. (1981). *Conversational organization: Interaction between speakers and hearers*. New York: Academic Press.

Hanna, J. E., & Brennan, S. E. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory & Language, 57*, 596–615.

Holle, H., & Gunter, T. C. (2007). The role of iconic gestures in speech disambiguation: ERP evidence. *Journal of Cognitive Neuroscience, 19*, 1175–1192.

Holler, J., Kelly, S., Hagoort, P., & Özyürek, A. (2012). When gestures catch the eye: The influence of gaze direction on co-speech gesture comprehension in triadic communication. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Meeting of the Cognitive Science Society* (pp. 467–472). Austin, TX: Cognitive Society.

Kelly, S. D., Barr, D., Church, R. B., & Lynch, K. (1999). Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of Memory & Language, 40*, 577–592.

Kelly, S. D., Creigh, P., & Bartolotti, J. (2010). Integrating speech and iconic gestures in a Stroop-like task: Evidence for automatic processing. *Journal of Cognitive Neuroscience, 22*, 683–694.

Kelly, S. D., Kravitz, C., & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain & Language, 89*, 253–260.

Kelly, S. D., McDevitt, T., & Esch, M. (2009). Brief training with co-speech gesture lends a hand to word learning in a foreign language. *Language & Cognitive Processes, 24*, 313–334.

Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science, 21*, 260–267.

Kelly, S. D., Ward, S., Creigh, P., & Bartolotti, J. (2007). An intentional stance modulates the integration of gesture and speech during comprehension. *Brain & Language, 101*, 222–233.

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica, 26*, 22–63.

Knöferle, P., & Kreysa, H. (2012). Can speaker gaze modulate syntactic structuring and thematic role assignment during spoken sentence comprehension? *Frontiers in Cognitive Science, 3*, 538.

Lerner, G. H. (2003). Selecting next speaker: The context-sensitive operation of a context-free organization. *Language in Society, 32*, 177–201.

Leys, C., & Schumann, S. (2010). A nonparametric method to analyze interactions: The adjusted rank transform test. *Journal of Experimental Social Psychology, 46*, 684–688.

Obermeier, C., Dolk, T., & Gunter, T. C. (2012). The benefit of gestures during communication: Evidence from hearing and hearing-impaired individuals. *Cortex, 48*, 857–870.

Özyürek, A. (2002). Do speakers design their cospeech gestures for their addressees? The effects of addressee location on representational gestures. *Journal of Memory & Language, 46*, 688–704.

Özyürek, A., Willems, R. M., Kita, S., & Hagoort, P. (2007). On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. *Journal of Cognitive Neuroscience, 19*, 605–616.

Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford, UK: Oxford University Press.

Rossano, F. (2012). Gaze behavior in face-to-face interaction. Published PhD thesis, Radboud University, Nijmegen.

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology, 21*, 211–232.

Senju, A., & Johnson, M. H. (2009). The eye contact effect: Mechanisms and development. *Trends in Cognitive Sciences, 13*, 127–134.

Staudte, M., & Crocker, M. (2011). Investigating joint attention mechanisms through spoken human–robot interaction. *Cognition, 120*, 268–291.

Straube, B., Green, A., Jansen, A., Chatterjee, A., & Kircher, T. (2010). Social cues, mentalizing and the neural processing of speech accompanied by gestures. *Neuropsychologia, 48*, 382–393.

Willems, R. M., Özyürek, A., & Hagoort, P. (2007). When language meets action: The neural integration of speech and gesture. *Cerebral Cortex, 17*, 2322–2333.

Willems, R. M., Özyürek, A., & Hagoort, P. (2009). Differential roles for left inferior frontal and superior temporal cortex in multimodal integration of action and language. *NeuroImage, 47*, 1992–2004.

Wu, Y. C., & Coulson, S. (2005). Meaningful gestures: Electrophysiological indices of iconic gesture comprehension. *Psychophysiology, 42*, 654–667.

Wu, Y. C., & Coulson, S. (2007a). How iconic gestures enhance communication: An ERP study. *Brain & Language, 101*, 234–245.

Wu, Y. C., & Coulson, S. (2007b). Iconic gestures prime related concepts: An ERP study. *Psychonomic Bulletin & Review, 14*, 57–63.

Yap, D. F., So, W. C., Melvin Yap, J. M., Tan, Y. Q., & Teoh, R. L. S. (2011). Iconic gestures prime words. *Cognitive Science, 35*, 171–183.