Here's not looking at you, kid!
Unaddressed recipients benefit from co-speech gestures when speech processing suffers

Judith Holler (judith.holler@mpi.nl)[1,2]
Louise Schubotz (louise.schubotz@mpi.nl)[1]
Spencer Kelly (skelly@colgate.edu)[3]
Peter Hagoort (peter.hagoort@mpi.nl)[1,5]
Manuela Schütze (manuela.schuetze@mpi.nl) [1]
Aslı Özyürek (asli.ozyurek@mpi.nl)[1,4]

1 Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525XD Nijmegen, The Netherlands
2 University of Manchester, School of Psychological Sciences, Coupland Building 1, M13 9PL Manchester, UK
3 Colgate University, Psychology Department, Center for Language and Brain, Oak Drive 13, Hamilton, NY 13346, USA
4 Radboud University, Centre for Language Studies, Erasmusplein 1, 6525HT Nijmegen, The Netherlands
5 Radboud University, Donders Institute for Brain, Cognition and Behaviour, Montessorilaan 3, 6525 HR Nijmegen, The Netherlands

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 these different modalities, and how perceived communicative intentions, often signaled through visual signals, such as eye gaze, may influence this processing. We address this question by simulating a triadic communication context in which a speaker alternated her gaze between two different recipients. Participants thus viewed speech-only or speech+gesture object-related utterances when being addressed (direct gaze) or unaddressed (averted gaze). Two object images followed each message and participants' task was to choose the object that matched the message. Unaddressed recipients responded significantly slower than addressees for speech-only utterances. However, perceiving the same speech accompanied by gestures sped them up to a level identical to that of addressees. That is, when speech processing suffers due to not being addressed, gesture processing remains intact and enhances the comprehension of a speaker's message.

Keywords: language processing; co-speech iconic gesture; eye gaze; recipient status; communicative intent; multi-party communication.

Introduction

Human face-to-face communication is a multi-modal activity and often involves multiple participants. Despite this, language comprehension has typically been investigated in uni-modal (i.e., just speech) and solitary (i.e., one passive listener) contexts. The present study investigates language comprehension in the context of two other modalities omnipresent during face-to-face communication, co-speech gesture and eye gaze. Moreover, it explores the interplay of these modalities during comprehension in a situated, dynamic social context involving multiple interlocutors in different roles.

There is, by now, a plethora of empirical evidence demonstrating that speech and co-speech gestures are semantically integrated during comprehension (e.g., Holle & Gunter, 2007; Holle, Gunter, Rüschemeyer, Hennenlotter, & Iacoboni, 2008; Kelly, Özyürek, & Maris, 2010; Özyürek, Willems, Kita, & Hagoort, 2007; Willems, Özyürek, & Hagoort, 2007, 2009). However, only two recent studies have begun to explore to what extent this integration is automatic, and to what extent it is controlled and influenced by the pragmatics of communication, such as the perceived intentional coupling of gesture and speech (e.g., when observing a gesture performed by one person accompanying speech produced by another) (Kelly, Ward, Creigh, & Bartolotti, 2007; Kelly, Creigh, & Bartolotti, 2010). The findings suggest that the semantic integration of gesture and speech is indeed sensitive to the intentional coupling of the speech and gesture modalities.

A question that remains is whether this holds when we situate speech and gesture comprehension in a context that is much closer to natural communication, such as in a face-to-face context, where speech and gesture are accompanied by additional nonverbal social cues, such as eye gaze. Due to the saliency of the sclera and the contrast it forms with the iris in the human eye, gaze direction is not only omnipresent but also an extremely powerful social cue in human face-to-face interaction (Senju & Johnson, 2009). While some studies have investigated speech and gesture comprehension in the presence of eye gaze, they have typically done so without manipulating eye gaze direction as an independent cue (e.g., Green, Straube, Weis, Jansen, Willmes, Konrad, & Kircher, 2009; Kelly, Kravitz, &

Hopkins, 2004; Skipper, Goldin-Meadow, Nusbaum, & Small, 2009; Straube, Green, Jansen, Chatterjee, & Kircher, 2010; Wu & Coulson, 2005, 2007).

One exception is a recent study by Holler, Kelly, Hagoort, and Özyürek (2012). Their study involved one speaker alternating her gaze between two recipients, thus rendering one of them addressed and the other unaddressed during each message she communicated. Despite this study involving multi-modal messages consisting of speech and gesture, the study was designed to primarily yield insights into the influence of eye gaze direction on the processing of the gestural component of bi-modal utterances. Thus, while showing that addressed and unaddressed recipients process gestures differently, the findings revealed no effect of eye gaze on the processing of speech. However, as the authors state themselves, this does not necessarily mean that addressed and unaddressed recipients do not differ in how they process speech; one reason being that the paradigm applied in their study required participants to focus attention on the verbal modality to make judgements about the *speech* they heard. This explicit attentional focus might have masked effects of eye gaze on speech processing that may be revealed in other contexts.

There are some studies that provide us with good reasons to assume that this is indeed the case. For example, Schober and Clark (1989) showed that overhearers process speech less well than addressees in a referential communication task. While this study did not involve a manipulation of eye gaze direction (nor a face-to-face context), it demonstrates that recipient status can have a significant impact on how we process language. This evidence is complemented by more recent studies that did investigate speech processing in the context of gaze. For example, Staudte and Crocker (2012) showed that a robot's eye gaze towards objects in the interlocutors' environment influenced participants' reference resolution, while Knöferle and Kreysa (2012) demonstrated that a person's eye gaze towards objects influences how participants process speech with respect to thematic role assignment and syntax.

Based on this earlier research, we predict that *social* eye gaze, indicating communicative intent and recipient status in conversation, also influences the processing of speech. Thus, the present study investigates how, in a multi-party setting, different types of recipients (as signaled through a speaker's eye gaze direction) process speech, and speech accompanied by gestures. To do so, we developed a visually focused paradigm that avoids explicit attention to speech to allow us to better observe potential differences in addressed and unaddressed recipients' processing of both uni-modal speech-only and bi-modal speech-gesture utterances.

Like Holler et al. (2012), we implement our task in a situated, triadic communicative setting. However, in our task, participants watched a speaker conveying speech-only or speech + gesture utterances referring to objects (e.g., 'he prefers the laptop'). The gestures accompanying these utterances in the bi-modal condition were always iconic in nature and depicted a typical feature of the object (such as its function, e.g., a typing gesture). These messages were followed by two object images, one of them having been mentioned in the utterance. The task was simple – speakers were asked to indicate as quickly as possible which of the two images was related to the speaker's preceding message. This paradigm allows us to test, firstly, how different types of recipients process speech when it is the only modality carrying semantic information, and, secondly, how they process semantic messages that are communicated bi-modally, via speech and co-speech gesture.

More specifically, we are also interested in seeing whether the findings from our study are in line with the *Competing Modalities Hypothesis* proposed by Holler et al. (2012). This hypothesis states that unaddressed recipients focus more on gesture than do addressed recipients, since they are processing information from fewer (visual) modalities overall (i.e., no eye gaze, since the speaker's eyes are averted to the other participant). They can therefore devote more cognitive resources to the gestures, and, as a consequence, they process the gesturally depicted meaning *more* than addressees. In contrast to Holler et al. (2012), whose paradigm was designed to tap primarily into co-speech gesture processing, we here test this hypothesis in a paradigm that allows us to measure the processing of both gesture *and* speech. That is, if, in the present study, we do observe an effect of recipient status on the processing of speech in a way that is in line with past research (e.g., Schober & Clark, 1989) - meaning unaddressed recipients process speech less well - then the enhanced processing of co-speech gestures may benefit unaddressed recipients' comprehension of the speaker's message and compensate for some (or even all) of the speech processing disadvantage.

As an alternative, Holler et al. (2012) proposed the *Fuzzy Representation Hypothesis*. This hypothesis predicts that unaddressed recipients perceive gestures as being less intended for them than for the gazed at recipient. They therefore process gestures *less* clearly than addressees, and, as a consequence, end up with a fragmented, or fuzzy, representation of the gesturally depicted meaning. If the Fuzzy Representation Hypothesis is true, then we should see no benefitting effect of gestures on the processing of speech. Rather, unaddressed recipients might be slowed down even more when trying to process bi-modal utterances, since not only the speech poses difficulties for them, but also the gestures.

The present study aims to tease apart which of these two hypotheses may best explain how addressed and unaddressed recipients (as indicated by the speaker's eye gaze direction) comprehend multi-modal language in a pragmatically much richer communication context than has been traditionally investigated, that is, in a context that bears somewhat more resemblance to the kind of *joint activity* that human communication is (Clark, 1996).

Method

Participants

32 right-handed, native German speakers (16 female) participated in the experiment (mean age 24.5yrs).

Design

We used a 2x2 within-participants factorial design, manipulating the gaze direction of the speaker (direct gaze/addressed recipient condition vs. averted gaze/unaddressed recipient condition) as well as the modality of presentation (speech-only vs. speech+gesture).

Materials and Apparatus

Video clips 160 short sentences of a canonical SVO structure were constructed. Sentences always referred to an object combined with a non-action verb (see below for more detail), e.g. *'he prefers the laptop'* ('er bevorzugt den Laptop'). The iconic gestures accompanying the sentences always referred to the object that was mentioned in speech and provided information about its shape, function, or size (see Fig.1, for a gesture depicting the act of typing).

In order to guarantee that the gestures unambiguously referred to the *objects* mentioned, verbs were carefully selected to be as neutral as possible and were never action verbs. Hence, rather than more commonplace constructions like 'he types on the laptop' where the typing gesture could refer to both 'typing' and 'laptop', verbs like 'prefer' ('bevorzugen'), 'like' ('mögen'), or 'see' ('sehen') were used in the sentences. Our manipulation of both gaze direction and modality of presentation required each sentence to be recorded in four versions: 1. direct gaze (addressed) speech-only, 2. direct gaze (addressed) speech+gesture, 3. averted gaze (unaddressed) speech-only, and 4. averted gaze (unaddressed) speech+gesture (Fig. 1).

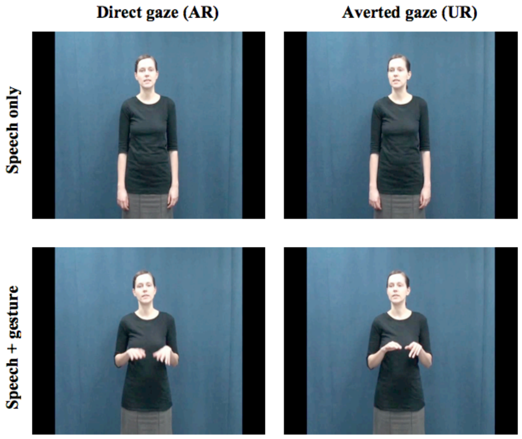

Figure 1: Four different versions of the 'laptop' stimulus. AR = addressed recipient, UR = unaddressed recipient.

Object pictures We created a total of 320 object pictures. 160 of these were pictures of the objects mentioned in the 160 stimulus sentences (e.g., a picture of a laptop), and an additional 160 pictures were selected to serve as unrelated pictures, such that the 'laptop' would be presented alongside a 'towel', for example (Fig. 2). Object pictures were searched via Google Images and further edited in Adobe Photoshop to have all objects presented in the same quality and size on a white background.

Prior to testing, all 320 pictures were judged by two raters (female native German speakers who did not participate in the main experiment) for their ease of identification.

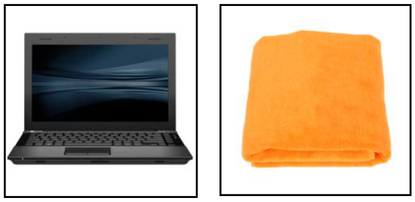

Figure 2: Example of a pair of object pictures.

Each participant saw each of the 160 video clips in one of the four conditions exactly once, resulting in 160 experimental trials per participant (40 trials per condition), plus 24 filler trials, yielding 184 trials overall. To avoid confounding effects of the order in which the pictures were presented on the screen, this order was counterbalanced.

Videos and object pictures were presented on a 15" laptop screen using Presentation software (http://www.neurobs.com). The audio signal of the videos was presented via high quality Sennheiser headphones.

Procedure

Participants were tested individually. At the beginning of each testing session, participants were familiarised with the experimental set-up and the course of the experiment, and were seated in front of the experiment laptop where they received their instructions.

Participants were told that they would see a number of pre-recorded video clips of a speaker (in fact a confederate) who, they were told, spontaneously formed short sentences based on line drawings and single words displayed on a screen not displayed in the video shot. They were also told that during the recordings, a second person was present in the room, sitting diagonally across from the speaker. The speaker was supposedly instructed to sometimes address this other (fictitious) participant when producing her utterances (averted gaze condition), and to sometimes address the (actual) participant via a video camera positioned straight across from her (direct gaze condition). Participants were instructed that following each video clip, they would see two pictures of objects on the screen, and that it was their task to indicate via button press which of the two pictures best matched the speaker's message (left button for the left-hand picture, right button for the right-hand picture). They were asked to react as quickly and as accurately as possible. Reaction times of participants' left/right responses were recorded via a button box, as were response accuracies.

In order to ensure that participants were actually watching the video clips and not basing their decision on the spoken part of the message only, they were explicitly asked to look at the screen during the entire course of the experiment. This was further enforced by the presence of a surveillance camera (our checks showed that no participant had looked away), which all participants agreed to be video-recorded with during the experiment.

Before the beginning of the experiment proper, participants completed a total of six practice trials. As in the actual experiment, each trial consisted of a video clip, followed immediately by the two object pictures, which stayed onscreen until the participants pressed a button. After their response, participants saw a fixation cross for a random time interval between 2 and 5 seconds before the next trial started.

Results

A total of six trials from two participants were excluded from the analysis beforehand because of a technical error. An alpha value of .05 was used throughout our statistical analyses. All p-values reported are two-tailed.

For the analysis of the reaction times[1], we excluded all incorrect responses, 83 in total (= 1.62% of all trials). Also excluded from the analysis were responses more than 2.5 SD above or below each subject's mean reaction time (this resulted in 118 responses being excluded: 40 in the speech-only condition, direct gaze, 31 in the speech-only condition, averted gaze, 23 in the speech+gesture condition, frontal gaze, and 24 in the speech+ gesture condition, averted gaze).

Figure 3 shows the reaction time data for the 2 (gaze direction: direct vs. averted) x 2 (modality of presentation: speech-only vs. speech+gesture) repeated measures ANOVA. The results yielded a significant interaction, F(1,31) = 5.947, p = .021. The main effect of modality was not significant, F(1,31) = 3.431, p = .074, and neither was the main effect of gaze, F(1,31) = .464, p = .501.

In line with our hypotheses, we calculated two a priori contrasts (using paired-samples t-tests), comparing addressed and unaddressed recipients' processing of uni-modal speech-only utterances, as well as their processing of the bi-modal speech+gesture utterances. The first comparison showed that unaddressed recipients (M = 542ms) were significantly slower than addressees (M = 530ms) at processing speech-only utterances, t(1,31) = 2.547, p = .016. The second comparison, however, showed that unaddressed (M = 525ms) and addressed (M = 531ms) recipients did *not* differ in their processing of speech+gesture utterances, t(1,31) = 1.112, p = .275.

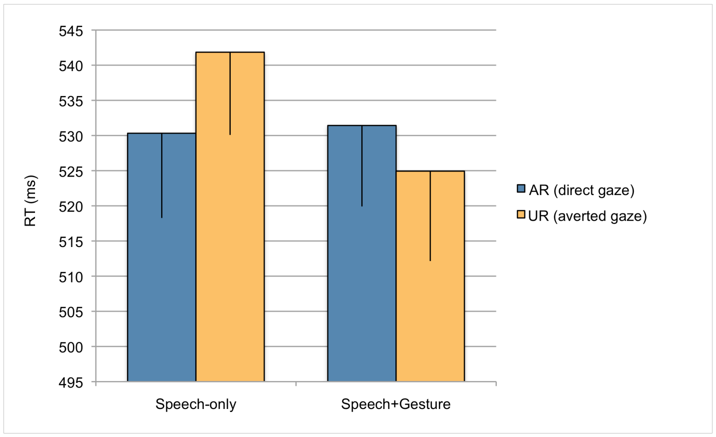

Figure 3: Addressed recipients' (AR) and unaddressed recipients' (UR) reaction times (ms) in the speech-only and speech+gesture conditions (error bars represent SE).

Discussion

This study investigated multi-modal language processing in a situated, socially dynamic communication setting involving multiple parties. The specific question we tried to answer is how different types of recipients, as signaled through a speaker's eye gaze, process speech and speech accompanied by iconic gestures, in a triadic communication scenario. The findings revealed a significant interaction between modality and recipient status. More precisely, they show, first and foremost, that the processing of speech-only utterances is indeed affected by recipient status in our task, since unaddressed recipients were significantly slower in this condition than were addressed recipients. Crucially, addressed and unaddressed recipients did not differ in their processing of speech+gesture utterances. That is, unaddressed recipients significantly benefitted from the information depicted in the gestural modality, allowing them to perform at the same level as addressees when perceiving bi-modal rather than uni-modal utterances.

The findings are thus very much in line with the Competing Modalities Hypothesis (Holler et al., 2012). Unaddressed recipients appear to focus their cognitive resources on the processing of co-speech iconic gestures. At the same time, the findings allow us to further refine this hypothesis; because we found that unaddressed recipients do not process speech more quickly than addressed recipients, the competition effect seems to apply to the visual modalities (gesture and gaze) only. In other words, due to not having to process eye gaze, unaddressed recipients can focus more on gesture and, as a consequence

[1] The analysis of participants' error rates yielded a significant modality effect, with both types of recipients being more accurate in the bi-modal than in the uni-modal condition. No other effects were significant.

process this information more. Their increased processing capacity due to the absence of direct gaze does not, however, affect their processing of speech-only utterances.

The reason as to why, in contrast to Holler et al. (2012), we found a numerical but no reliable difference between addressed and unaddressed recipients' processing of bi-modal utterances (i.e., unaddressed recipients were slightly faster in the bi-modal condition than addressed recipients were, but not significantly so) is likely to be due to our change in paradigm. As argued in the Introduction, the explicit attentional focus on the verbal modality in Holler et al.'s (2012) study might have masked differences in the processing of speech – an assumption that we were able to corroborate here. In the present study, we purposefully shifted participants' attention towards the visual modality (by asking them to identify *pictures*) in order to uncover potentially previously masked differences in speech processing, while being aware that this shift in paradigm might, in turn, reduce differences in the processing of visual (i.e., gestural) information between addressed and unaddressed recipients. The present study thus complements that by Holler et al. (2012) nicely. Together, they offer us a more comprehensive insight into how different recipients process uni-modal and bi-modal utterances in the presence of eye gaze.

What remains to be investigated are the exact cognitive mechanisms underlying our Competing Modalities account. Currently, we are unable to determine whether the iconic co-speech gestures benefit unaddressed recipients' processing of speech because they are semantically integrated with the verbal information - thus leading to a richer, unified mental representation of the concept of 'laptop', for example – or whether they 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 and *not* as a unified representation (much like a dually-coded representation à la Paivio (1986)). Future studies, preferably involving on-line measures suitable for dipping directly into semantic integration processes, are needed to answer this question.

In conclusion, the present study has brought together three different modalities in a language processing paradigm, and it advances our understanding of how perceived communicative intent, as signaled through a speaker's eye gaze, influences the interplay of these modalities during comprehension in a situated, face-to-face-like (rather than solitary) 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 semantic information carried by two concurrent modalities, speech and co-speech gestures. Moreover, we have shown that in situated face-to-face settings involving multiple recipients, the gestural modality can benefit unaddressed recipients – when speech processing suffers, gestures help.

Acknowledgments

We would like to thank the University of Rostock for their provision of testing space, the Radboud University and MaxNet initiative for financial support, the Neurobiology of Language Department and the Gesture & Sign Research Group at the Max Planck Institute for Psycholinguistics for helpful feedback during discussions of this study, and the European Commission for funding this research (JH was supported through a Marie Curie Fellowship #255569, and through ERC Advanced Grant #269484INTERACT).

References

Clark, H. (1996). Using language. Cambridge: Cambridge University Press.

Green, A., Straube, B., Weis, S., Jansen, A., Willmes K., Konrad, K., & Kircher, T. (2009). Neural integration of iconic and unrelated coverbal gestures: a functional MRI study. *Human Brain Mapping, 30,* 3309–3324.

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

Holle, H., Gunter, T. C., Rüschemeyer, S.A., Hennenlotter, A., & Iacoboni, M. (2008). Neural correlates of the processing of co-speech gestures. *NeuroImage, 39,* 2010-2024.

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., 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., 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 and Language, 101, 222-233.

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.

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.

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.

Skipper, J. I., Goldin-Meadow, S., Nusbaum, H. C., & Small, S. L. (2009). Gestures orchestrate brain networks for language understanding. *Current Biology, 19,* 661-667.

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. (2007). How iconic gestures enhance communication: An ERP study. Brain and Language, 101, 234-245.