

# Understanding discourse in face-to-face settings: The impact of multimodal cues and listening conditions

Anna Krason<sup>1\*</sup>, Rosemary Varley<sup>2</sup>, Gabriella Vigliocco<sup>1</sup>

<sup>1</sup> Experimental Psychology, University College London, UK

<sup>2</sup> Language and Cognition, University College London, UK

\*Corresponding author

Department of Experimental Psychology, University College London,

26 Bedford Way, WC1H 0AP, London, UK

[anna.krason.15@ucl.ac.uk](mailto:anna.krason.15@ucl.ac.uk)

## **ABSTRACT**

In face-to-face contexts, discourse is accompanied by various cues, like gestures and mouth movements. Here, we asked whether the presence of gestures and mouth movements benefits discourse comprehension under clear and challenging listening conditions and, if so, whether this multimodal benefit depends on the communicative environment in which interlocutors are situated. In two online experiments, participants watched videoclips of a speaker telling stories and answered yes-no questions about the content of each story. The speaker in the videos was spontaneously gesturing (or kept her hands still) and was wearing a surgical mask (or had her lips visible). The experiments differed in the communicative environment. In Experiment 1, the speaker narrated stories in silence whereas the listener (participants) heard them in clear or degraded speech conditions (analogous to watching the news on TV in a quiet or noisy café). In Experiment 2, the speaker narrated the stories once in silence and once while listening to background noise, and the listener heard them in clear or degraded speech condition, respectively (analogous to listening to a friend in a quiet or noisy café). Across the experiments, we found that co-speech gestures facilitated discourse comprehension regardless of the listening conditions or the presence of a mask. In contrast, mouth movements were primarily helpful in challenging listening conditions. These findings indicate that both cues matter to listeners but to a different extent. Moreover, we found that the multimodal benefit to comprehension was similar regardless of the communicative environment. Thus, this study demonstrates the importance of both co-speech gestures and mouth movements to discourse comprehension, offering insights into the dynamic interplay between these cues under different communicative environments.

## **INTRODUCTION**

Discourse refers to a linguistic unit larger than a sentence and that often describes a connected set of events. It is generally agreed that listeners process discourse by building situation models of the described events using general knowledge, communicative context, and prior experiences (Zwaan, Langston, et al., 1995; Zwaan, Magliano, et al., 1995; Zwaan & Radvansky, 1998). Studies investigating discourse comprehension have primarily centered around the linguistic information conveyed by speech or text. However, in face-to-face settings, spoken discourse is accompanied by an abundance of communicative cues, such as hand gestures, mouth and facial movements, body posture, eye gaze, and prosody, that support the creation of situation models and in turn impact discourse comprehension (Clark, 1996; Goodwin, 1981; Kendon, 1980, 2004; McNeill, 1992; Schegloff, 1984). Moreover, the importance of visual cues became particularly prominent during the COVID-19 pandemic because wearing health masks muffles speech and hides important mouth and facial movements, while at the same time it might drive listeners to rely upon other available cues, like gestures. It is still unclear to what extent these, often referred to as “non-linguistic” cues, improve comprehension as most studies have investigated isolated words accompanied by either gestures or mouth movements, ignoring the possibility that these cues may interact with each other and with the auditory signal. The aim of the current study is to fill this gap by looking at how spontaneously produced co-speech gestures and wearing a face mask impact discourse comprehension under clear and challenging listening conditions.

### **Co-Speech Gestures in Comprehension**

Co-speech gestures (or gesticulations; Kendon, 1988) are spontaneous and idiosyncratic hand movements that are time-locked to speech. They are ubiquitous in face-to-face discourse and studies have shown that they serve a communicative function

(Hostetter, 2011; Kendon, 2004). Representational gestures, which refer to physical or abstract features and properties of objects, actions, elements of events in an imagistic way (e.g., rapidly moving index and middle fingers to represent a walking action or pointing behind oneself to refer to past times), facilitate discourse comprehension by activating the semantic system (Kendon, 1988; Kita et al., 1997; McNeill, 1992). These gestures are particularly beneficial when the information they convey is spatial or motoric (e.g., describing directions), which has been explained in terms of a construction of a spatially organized situation model (Cutica & Bucciarelli, 2013). It has been suggested that representational gestures are integrated with speech during online comprehension, as they are automatically processed even when the semantic information they convey mismatches the speech (Green et al., 2009; Kelly et al., 2010; McNeill et al., 1994). In line with Kintsch's model of comprehension (Kintsch, 1998; van Dijk & Kintsch, 1983; Kintsch & van Dijk, 1978), representational gestures could be integrated both at the semantic and situation model levels thus supporting interpretation of the discourse.

While many studies have investigated the representational gestures speakers produce while narrating (e.g., Alibali et al., 2000; Brown & Gullberg, 2008; Kita & Özyürek, 2003; McNeill & Levy, 1982), only a few have looked at the role of these gestures in discourse comprehension. In three studies, Dargue and Sweller (2018, 2019, 2020) presented participants with videos of a speaker narrating short stories and asked them to recall as much information as possible. On half of the occasions, the speaker also produced iconic (representational) gestures that were either "typical" (most frequently occurring gestures obtained from a norming retell task) or "atypical" (a less common form of gestures created by the experimenters). A general conclusion from these studies was that typical, but not atypical gestures, improve free recall performance, which was further explained in terms of greater semantic relatedness between the information provided by typical gestures and speech

(Dargue & Sweller, 2018, 2019, 2020). Similar findings were reported by McKern et al. (2021), who used the same set of materials but additionally manipulated the clarity of the audio using background noise to increase task difficulty. Again, only typical gestures were found to facilitate narrative comprehension and this effect was particularly large in people with poorer non-verbal memory, as measured with a delayed recall task. The authors interpreted this finding in terms of typical gestures being most beneficial during semantic processing because they leave a long-term memory trace thanks to their iconicity. There was no interaction with speech clarity, suggesting that commonly occurring iconic gestures benefit speech comprehension regardless of task difficulty. However, the lack of effect of atypical iconic gestures can be related to the fact that they were generated artificially and therefore they might have introduced a mismatch with the auditory information, which led to poorer performance relative to when typical gestures accompanied speech.

Although some have argued that discourse comprehension is primarily influenced by representational gestures because they support the creation of situation models by bringing events to interlocutors' eyes (Feyereisen, 2006; Murgiano et al., 2021), other types of gestures may also play a role. These include beat gestures, which are rhythmic movements tightly coupled with speech prosody that mark salient phrases or words (Krahmer & Swerts, 2007), and have been found to elicit enhanced activation in the auditory cortex (Hubbard et al., 2009); pragmatic gestures, such as moving an open, flat hand towards the listener, that serve an interactive function (e.g., turn-taking; Bavelas et al., 1992, 1995); and deictic gestures, which include abstract pointing gestures indicating the location of a physically absent referent (e.g., pointing to a particular location in space to refer to the same protagonist in a story; Kendon, 1988, 2004; McNeill, 1992), and have been found to support discourse cohesion. For instance, Gunter et al. (2015) used EEG to measure the impact of abstract pointing gestures that indicated the location of the protagonists described in a story that was

either congruent (i.e., consistent with previously established location) or incongruent (i.e., inconsistent with previously established location). Larger N400 and P600 effects, which are event-related potentials often associated with language processing difficulty (Kutas & Federmeier, 2011), were found for the incongruent gestures, suggesting that listeners automatically process these gestures during comprehension, similarly to the iconic ones (e.g., Kelly et al., 2010). Furthermore, deictic gestures can reactivate situational models even when gestures are not in view anymore (Sekine & Kita, 2017) or when sentence meaning is ambiguous (Sekine & Kita, 2015; Smith & Kam, 2012).

To summarize, previous research has shown that representational and deictic gestures support discourse comprehension thanks to the semantic relatedness between gesture and speech information. Less is known about the extent to which observing other types of gestures, such as beat or pragmatic gestures, benefits discourse comprehension and whether any such benefit is influenced by gesture frequency or kinematics (e.g., amplitude). Moreover, in all the studies described above the gestures were scripted rather than naturally occurring which may introduce differences in processing. Our study contributes to this line of investigation by looking at non-scripted spontaneous gestures (including representational, deictic, pragmatic, and beat gestures).

### **Visual Speech Cues and the Effect of Masks**

Face and mouth movements are also central cues for speech comprehension. Mouth movements, in particular, primarily tap onto the phonological level of processing and have been shown to aid speech recognition by constraining lexical competition (Lachs & Pisoni, 2004; Peelle & Sommers, 2015; Tye-Murray et al., 2007). Most of what we know about the role of mouth movements in speech comprehension comes, however, from studies investigating isolated words (or embedded in sentences) while less is known about their impact on spoken discourse comprehension. Using a shadowing task, Reisberg et al. (1987)

demonstrated that the presence of mouth movements speeds up the processing of both clear and distorted spoken passages. A facilitatory effect of mouth movements was also reported for accuracy performance. Arnold and Hill (2001) presented short passages in auditory-only and audiovisual modalities followed by comprehension questions. The passages were either presented in listeners' native versus non-native language, accompanied by different accents, and semantically/syntactically complex versus simple sentences. The results showed that comprehension of narratives was always better when mouth movements were visible, suggesting that visual speech information is automatically processed during comprehension of passages.

The interest in the role of mouth movements has grown in the last couple of years because the COVID-19 pandemic has forced people to wear face masks covering mouth and cheeks and making speech muffled. Studies investigating the effect of face masks on speech intelligibility have demonstrated that both cognitive and listening effort increase (Giovanelli et al., 2021), which is particularly detrimental for people with hearing impairment (Saunders et al., 2021) and cochlear implants (Homans & Vroegop, 2021). In addition, emotional processing is hindered (Mheidly et al., 2020), which further leads to decreased trust, particularly in clinical populations (Malik et al., 2021). Interestingly, people are still quite accurate at processing emotions even when the mouth is covered with a mask provided the rest of the body is visible (Ross & George, 2022). This finding suggests that interlocutors flexibly weight the incoming information and if one cue is unavailable (or less informative), they make use of other cues (Krasen et al., 2021; Skipper et al., 2009; Zhang et al., 2021).

Masks can be of different sizes, shapes, and materials, however this does not appear to affect speech intelligibility as shown in work by Brown et al. (2021). The authors tested a group of younger and older adults on a comprehension task with sentences embedded in noise. Participants' task was to watch videoclips of a speaker producing sentences and type

down what they had heard. The speaker in the videos was asked to wear different health masks (surgical, cloth with filter, cloth without filter, transparent, or no mask). The results show that wearing a mask (regardless of its type) hinders sentence identification in noise in both younger and older adults to a similar extent. This finding is in line with studies demonstrating that mouth movements are particularly beneficial in noisy listening conditions (Grant & Seitz, 2000; Ma et al., 2009; Ross et al., 2007; Schwartz et al., 2004; Sumby & Pollack, 1954).

Finally, Pycha, Cohn, and Zellou (2022) investigated auditory sentence comprehension under various background noise conditions using a final word identification task. They manipulated factors such as whether the speaker was wearing a face mask (versus not), speaking style (clear or casual, i.e., when a speaker was asked to produce speech in a natural way), and speaker information (image of a person with or without a mask). Across two experiments, the authors showed that speech intelligibility was higher when the speaker was wearing a mask, which was interpreted in terms of the Lombard effect (Junqua, 1993), such that the speaker adjusted their voice to compensate for the presence of the mask, making the speech more intelligible. Comprehension also improved when the speech was clear rather than casual and when listeners knew that the speaker was wearing a mask. These results demonstrate that speech comprehension is a dynamic social act involving speaker-listener interactions, which depend on the information that interlocutors have about each other (see also Trujillo et al., 2021).

In brief, mouth and facial movements facilitate speech processing, but their effect becomes more prominent when the auditory encoding is challenging. For example, observing an interlocutor's mouth movements in noisy listening conditions can help disambiguate what is being said. Moreover, while wearing a health mask covering important facial cues became a standard during the COVID-19 pandemic, it is not clear whether and if so, masks affect



discourse comprehension. Finally, as with co-speech gestures, little is known whether the benefit of mouth movements depends on their interactions with other available cues.

### **Interactions of Hand and Mouth Movements**

Traditionally, different cues such as gestures and mouth movements have been studied in isolation, thus neglecting the possibility that cues may interact during speech comprehension and production. Only recently, the co-occurrence of multiple cues alongside speech has been investigated, with a handful of studies looking at connected speech (Trujillo et al., 2021; Zhang et al., 2021). For example, in a series of EEG experiments, Zhang et al. (2021) presented participants videos of a speaker uttering short passages while spontaneously producing gestures (including representational and beat gestures). The speaker's mouth movements were always visible as it is most often the case in naturalistic settings. The authors quantified linguistic information (word predictability), as well as the information contained in speech (prosody), and visual cues (gestures and mouth movements) and measured oscillatory changes in the N400 event-related potential. The results show that multimodal cues always modulate the comprehension of spoken passages, but they do so in an interactive and dynamic way. For example, mouth movements become particularly useful when gestures are also present, suggesting a multimodal enhancement to comprehension.

Trujillo et al. (2021) asked whether the Lombard effect is modulated by multimodal (including gesture and mouth) cues. The authors investigated face-to-face dyadic interactions, during which participants wore headphones and listened to a multi-talker babble. One of the participants had to communicate (in any manner they chose) action-related verbs to their experimental partner whose task was to correctly identify these verbs. The results demonstrate that speakers adjust all multimodal cues when faced with their listeners in adverse listening conditions, suggesting that the Lombard effect is a multimodal, rather than auditory-only, phenomenon and that these modulations are helpful for both interlocutors. In

addition, gestures interact with speech such that louder voice was reported particularly in the absence of gestures, whereas more exaggerated gestures were observed when the voice was neutral. Thus, gestures can be used as an additional channel to overcome noise. Modulations of mouth movements were only found in some participants, suggesting individual differences.

Other studies have further demonstrated that the communicative environment the interlocutors are situated in, i.e., whether a speaker/listener is in a noisy or quiet environment, matters to language comprehension. For example, Garnier, Menárd, and Alexandre (2018; see also Garnier, Henrich & Dubois, 2010) investigated speech intelligibility in clear and degraded conditions across three communicative scenarios: interactive audiovisual, in which the speaker interacts with the listener face-to-face, auditory-only, and reading aloud. The results demonstrated that visual facial cues (just as acoustics) are modulated in the presence of background noise, but crucially this effect was contingent upon the visibility of the listener to the speaker. Overall, these findings are in line with Pycha et al. (2022) and Trujillo et al. (2021) and suggest context-dependent modulations where audiovisual adjustments to noise are at least partially made for the listener.

Two important conclusions can be drawn from the above review. First, visual cues modulate linguistic information and they do so in a manner that depends on the informativeness of other cues. Second, speakers automatically adjust the way they produce visual cues in adverse listening conditions (multimodal Lombard effect), and these adjustments are suggested to be at least in part listener oriented. It is still unclear, however, whether these cues impact discourse comprehension differently depending on listening and speaking environments. We contrast two face-to-face communicative environments in this study: watching TV news in a café where the news presenter always produces cues in a quiet environment but the listener experiences them in silence (clear speech condition) or in noise

(degraded speech condition), versus listening to a friend in a café where the friend and the listener are situated in the same communicative environment (either quiet or noisy).

### **The Present Study**

In this study, we asked (i) whether and how visual cues, such as co-speech gestures and mouth movements, benefit discourse comprehension under clear and challenging listening conditions, and (ii) whether producing these visual cues in different communicative environments (i.e., in noise or in silence) impacts comprehension differently. In two separate experiments, we presented participants with videos of an actress telling stories about everyday events or retelling an episode from a Tom & Jerry or Sylvester & Tweety Pie cartoon. Participants were asked to answer eight yes-no comprehension questions about the content of each video-story. We manipulated the following: (i) the actress in the videos was either spontaneously gesturing or kept her hands still alongside her body while telling the stories (gesture present versus absent condition); (ii) the face of the actress was fully visible or her lips and cheeks were covered with a surgical mask (mouth present versus absent condition); (iii) the audio in the videos was clear (unedited version) or background noise was manually added (clear versus degraded speech condition). The decision to use background noise (rather than any other noise) and a health mask (rather than e.g., blurring the area of the lips) was driven by their semi-naturalistic character as people often need to process discourse in a noisy environment and wearing a mask has become a standard since the COVID-19 pandemic. The manipulations remained constant across the experiments, but the communicative environment in which the speaker narrated the stories differed between experiments. In Experiment 1, the speaker produced the story in silence and the listener watched the videos in either silence or noise. In Experiment 2, both the speaker and the listener were in a quiet or a noisy environment. For the degraded (noisy) condition in Experiment 2, the actress was wearing earphones and listening to cafeteria-type noise.

On the basis of prior studies investigating multimodal speech comprehension, we predicted that the presence of visual cues would facilitate comprehension of discourse (Zhang et al., 2021), and this effect would be most robust in adverse listening conditions (Drijvers & Özyürek, 2017; Krason et al., 2021). Moreover, we predicted that gestures would benefit comprehenders to a larger extent than mouth movements, given that they impact linguistic processing differently, i.e., gestures primarily support semantic processing while mouth movements support phonological processing (Hirata & Kelly, 2010). Additionally, if speakers automatically make multimodal adjustments while narrating in noise (as in the case of when the actress wore earphones and listened to background noise), and if these adjustments are, at least partially, intended for the listener (Trujillo et al., 2021; Garnier et al., 2018; Pycha et al., 2022), it is likely that we will observe a larger advantage of gestures and mouth movements on discourse comprehension in Experiment 2 than in Experiment 1.

The experiments were carried out under University College London Ethical Approval (0143/003) and were preregistered using <https://aspredicted.org/>. The preregistrations<sup>1</sup>, data, example materials, and the R code are publicly available and can be found on the Open Science Framework (OSF): <https://osf.io/6zxuw/>.

## **EXPERIMENT 1**

Experiment 1 assesses the impact of co-speech gestures, visible mouth movements, and listening conditions on discourse comprehension. In this experiment, the speaker always produced speech and multimodal cues in a quiet environment, while the listener (participants) processed the incoming information in either a quiet or noisy environment (e.g., similarly to watching TV news in a quiet or noisy café).

---

<sup>1</sup> Note the difference in labelling the experiments here (Experiment 1 versus Experiment 2) and in the preregistrations (Experiment 1a versus Experiment 1b).

## Methods

### *Participants*

We recruited 100 native speakers of American English via Prolific (<http://www.prolific.co/>). They were all right-handed monolinguals and reported no language, hearing, vision, or neurological impairments. The number of participants was determined using three separate methods. Simulations analysis with *simr* package (Green & MacLeod, 2016) based on coefficients from our previous study investigating interactions between gestures and mouth movements (Krasen et al., 2021) suggested that 80 participants would be enough to detect a three-way interaction with almost 95% power at an alpha level at 0.05. Using two other sample size calculations methods, we found that 100 participants would be needed to detect an effect size of 0.3 (GPower analysis; Faul et al, 2009) and have a good-to-excellent level of replicability and good level of precision (Trafimow, 2018; see the preregistration for more details).

Prior to the analyses, outliers were identified and removed. One individual performed below 3SD from the group mean and one other had average reaction time above 3SD from the group mean. The data from the remaining 98 participants (*Mean age* = 30.33, *SD* = 6.31, 49 females, 48 males, 1 preferred not to say) were used for the analyses.

### *Materials*

The stimuli consisted of 16 short stories, each accompanied by eight yes-no comprehension questions about the content of the stories. Half of the stories and the corresponding questions were taken from the Discourse Comprehension Test (DCT; Brookshire & Nicholas, 1993), which is a narrative comprehension assessment developed for adults with brain damage that has been standardized on neurologically healthy groups. The

DCT contains ten different stories but only eight were selected for the purpose of this study. The other half of the stories were written by the experimenters and were based on four episodes of Tom & Jerry and four episodes of Sylvester & Tweety Pie cartoons. Cartoon-type stories have been previously used to investigate the use of gestures during storytelling and elicitation tasks (e.g., Brown & Chen, 2013; Brown & Gullberg, 2008; Gullberg & Kita, 2009; Kita & Özyürek, 2003; McNeill & Levy, 1982). Cartoon stories are also more concrete and involve more actions, thus we expected to see a larger proportion of representational gestures relative to other types of gestures in these stories. The DCT stories are more abstract and should involve fewer such gestures. The reason for including different types of stories was to ensure variability and minimize the effects of fatigue, boredom, and/or story familiarity. The DCT and cartoon stories were matched as closely as possible by the number of words, number of unfamiliar words (i.e., words that fall outside of the classification of the 10,000 most frequently used words; Carroll, Davis, & Richman, 1971), number of sentences, mean sentence length (average number of words per sentence), number of subordinate clauses, and listening difficulty (i.e., the average number of syllables that are more than one per word for each sentence, averaged across sentences per story; the Easy Listening Formula from Fang, 1966). The questions for the cartoon stories were also modeled on those of the DCT, focusing on information salience (questions about main ideas that were central to the stories versus details that were not crucial to understanding the message of the stories) and directness (questions about the information directly stated in the stories versus implied information that could be inferred). Thus, there were four types of questions: main idea-stated (MI-S), main idea-implied (MI-I), detail-stated (DT-S), and detail-implied (DT-I), and each type appeared twice per story: once with the correct “yes” response and once with the correct “no” response. See Table 1 for the comparison between DCT and cartoon stories, with example stories and questions. Full materials are available on the project’s OSF page.

**Table 1**

*Comparison between the selected DCT (Brookshire & Nicholas, 1993) and cartoon stories*










*(based on episodes from Tom & Jerry and Sylvester & Tweety Pie).*

Story Type	DCT (n=8)	Cartoon (n=8)
<b>Example Stories</b>	“Neil Williams was short of money. The new term was about to begin and he didn’t have enough money to pay his tuition. So, one day, he walked to his parents’ home and borrowed their car. Then he drove to the bank to get a student loan. The loan officer at the bank was a tough old woman who always said she had never made a bad loan. (...)”	“One afternoon last summer, Sylvester had a clever plan to catch Tweety Pie. He hid in a room in a building of ‘Bird Watchers’ Society’, which was just across the street from Tweety Pie’s apartment. Tweety Pie was happily swinging in his cage on the windowsill. Sylvester looked out from the window and used binoculars to get a better view on Tweety Pie’s apartment. In less than two minutes, he spotted Tweety Pie’s cage. (...)”
<b>Example Questions</b>	1. Was Neil a high school student? (No) [MI-I] 2. Did Neil’s parents live nearby? (Yes) [DT-I] 3. Did Neil go to the bank to get a loan? (Yes) [MI-S] 4. Did Neil need the money to start a new business? (No) [MI-S] 5. Did Neil own a car? (Yes) [DT-S] 6. Did Neil go to the bank in the morning? (No) [DT-I] 7. Did Neil tell the woman that he had a cheese sandwich for lunch? (No) [DT-S] 8. Did Neil get the loan? (Yes) [MI-I]	1. Was Sylvester hiding in the same building where Tweety Pie? (No) [DT-S] 2. Was Tweety Pie outside the building? (No) [DT-I] 3. Did Sylvester spot Tweety Pie’s cage through binoculars? (Yes) [MI-S] 4. Was Tweety Pie scared of Sylvester? (Yes) [DT-I] 5. Did Sylvester use the main door to enter the Tweety Pie’s building? (Yes) [DT-S] 6. Did Sylvester follow all the signs of Tweety Pie’s building? (No) [MI-I] 7. Was Sylvester escorted out of the building? (No) [MI-S] 8. Did Sylvester have a headache at the end? (Yes) [MI-I]
<b>Example Attention Check Questions</b>	1. Was the actress wearing a black top?	1. Was the actress standing?
<b>Average number of words (All stories = 199.8)</b>	203.3	196.3
<b>Average number of unfamiliar words (All stories = 2.7)</b>	2.4	3
<b>Average number of sentences (All stories = 13.7)</b>	13.9	13.5
<b>Sentence length (All stories = 14.6)</b>	14.7	14.6
<b>Average number of subordinate clauses (All stories = 9.9)</b>	9.3	10.5
<b>Listening difficulty (All stories = 4.6)</b>	4.6	4.7

The stories were uttered by an actress who was a right-handed, native speaker of American English and they were video-recorded. The recordings were made with a professional camera (Zoom Q4) and took place at the actress' house due to COVID-19 restrictions at the time of the recording. The actress was asked to tell the stories under the following conditions (i) with her hands still on her lap and her face clearly visible, (ii) with her hands still on her lap, and wearing a surgical mask, (iii) with spontaneous gestures, and her face clearly visible, (iv) with spontaneous gestures, and wearing a surgical mask. Each story was recorded in four conditions at once, rotating across conditions. As a result, there were 64 (16 stories \* 4 conditions) different video-clips (~1.20-minute long). Next, all 64 videos were duplicated to create a degraded speech condition. That is, we extracted the audio from each videoclip and added cafeteria background noise of 0dB signal-to-noise ratio (SNR) using a personalized MATLAB script (MATLAB and Statistics Toolbox Release, 2021a), courtesy of Prof. Stuart Rosen. The degraded audio files were then combined with the corresponding videos resulting in 64 new video-files with background noise. The goal of manipulating speech clarity in this way was to represent a communicative environment, in which the cues were produced by the speaker in silence, but were heard/seen in noise (e.g., watching TV news in a noisy café). Altogether, there were eight experimental conditions (2 Gesture Presence x 2 Mouth Presence x 2 Speech Clarity) and each participant was exposed to each condition twice, completing 16 stories within the experiment.

**Figure 1** *Experimental conditions and the communicative environment from Experiment 1.*



EXPERIMENT 1				
Gesture Present		Gesture Absent		Communicative Environment
Mouth Present	Mouth Absent	Mouth Present	Mouth Absent	
				 1. Speaker in a <u>quiet</u> environment 2. Listener in a <u>quiet</u> environment
				

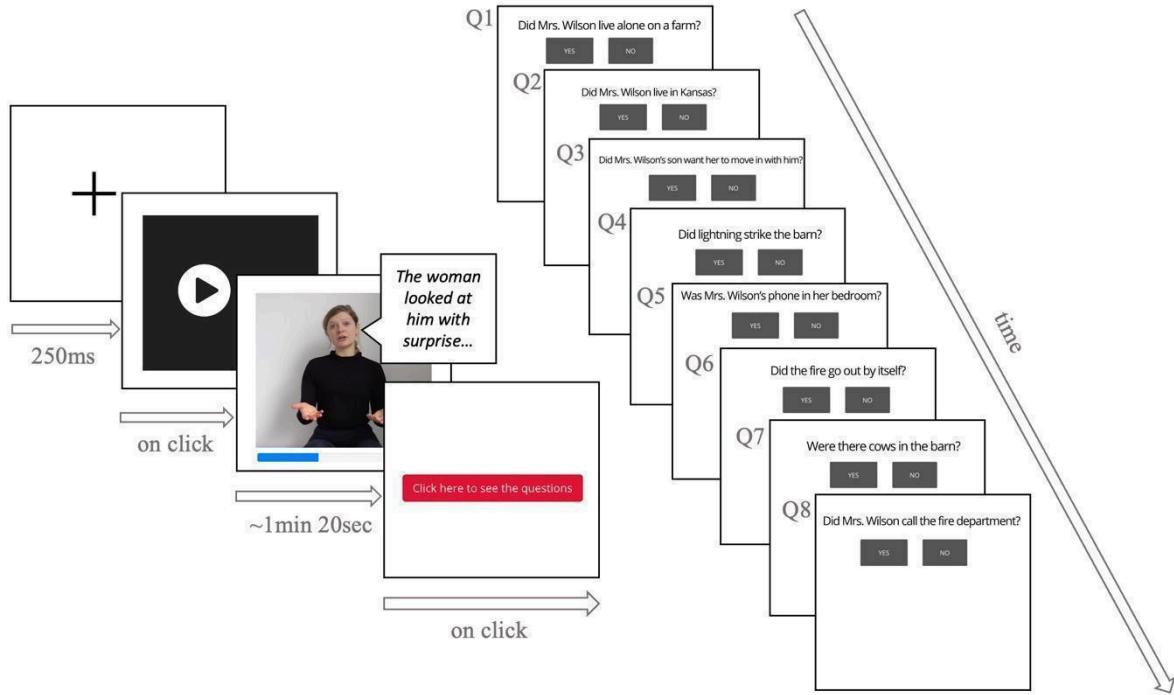
*Note:* Listener = Participant in the study; Speaker = Actress narrating stories. The yellow headphones icon indicates listening to cafeteria noise.

## Procedure

The experiment was created in Gorilla (<https://gorilla.sc/>) and lasted approximately 40 minutes. The task was to watch 16 videos and answer eight yes-no questions per video about the stories' content. The order of the videos and the conditions they were presented in were randomized across participants to ensure there was no order or learning effect. At the beginning of the experiment, participants were exposed to a practice story (in a clear speech condition with gestures and visible mouth movements) followed by eight yes-no questions appearing one after another. After each question, participants received feedback in the form of a green tick if they answered correctly or a red cross if they answered incorrectly. Feedback was provided to ensure that participants understood the task and became familiar with the different question types. At the end of the practice, participants were presented with written information reminding them that sometimes the sound would be noisy, and the speaker would wear a face mask, followed by a fragment of a video in degraded speech condition, with absent gestures and mouth movements covered with a mask as an example. The practice story was taken from the DCT practice samples and differed from the stories presented in the main experiment.

The main trials started with a fixation cross (250ms) followed by a video. Participants initiated the videos by clicking the “play” button. After each video, a series of eight questions appeared on the screen one by one with a fixation cross in between them (250ms). Moving to the next question required participants to choose an answer by clicking “yes” or “no” buttons. When participants answered all eight questions, a screen with “next” button appeared and participants could either take a short, self-timed break or could proceed to the next video. There was no feedback provided during the main task. See Figure 2 for an example of an experimental trial. Participants were asked to complete the experiment within an hour and there was a progress bar indicating how far in the experiment they were. Additionally, we introduced four attention checks that appeared on random occasions within the experiment. The attention checks were yes-no questions about the videos, not related to the content of the stories (see Table 1 for examples). The mean accuracy score on those trials was 3.65/4 (SD = 0.60) and no participant was removed based on them.

**Figure 2** *Example of an experimental trial with eight yes-no comprehension questions. The video depicts a speaker with visible mouth movements and producing spontaneous co-speech gestures.*



## Data Analysis

We carried out logistic mixed-effects regression on the accuracy scores in R (RStudio Team, 2015) using *afex* (Singmann et al., 2020). We fitted the model with all predictors of interest, including Gesture Presence (present versus absent), Mouth Presence (present versus absent), Speech Clarity (clear versus background noise), and up to the three-way interaction, following Winter (2019). Control variables of Story Type (cartoon versus DCT) and Question Type (MI-S, MI-I, DT-S, DT-I) were included in the model. All variables were categorical and were sum coded to test for main effects. *emmeans* package (Lenth, 2020) was used to obtain estimated marginal means. We also included by-Participant and by-Story intercepts to deal with non-independence of the data and account for participants' and stories' variability. We tried fitting the maximum random structure following Barr et al. (2013) and tested the models' fit using Principal Component Analysis (PCA) with the *RePsychLing* package (Baayen et al., 2015). We simplified the random structure based on the effect that had the least variance. To ensure convergence, *bobyqa* optimizer was used. The best fitting model

contained the random slope of Speech Clarity by-Participant. The size and the direction of the effects were assessed based on the coefficients following Jaeger (2008). There was no multicollinearity (all variance inflation factors  $<1.6$  based on *car* package; Fox & Weisberg, 2019). Prior to the analyses, outliers were identified and consequently removed. This included 12 trials that were answered quicker than 200ms and three questions due to overall accuracy below 3SD from the mean. For data cleaning and plotting, we used *tidyverse* (Wickham et al., 2019), *ggplot2* (Wickham, 2016), *sjPlot* (Lüdtke, 2022) and *beeswarm* (Eklund & Trimble, 2021) packages.

### ***Results and Discussion of Experiment 1***

There was a significant main effect of Gesture Presence ( $b = 0.07$ ,  $SE = 0.02$ ,  $z = 2.99$ ,  $p < 0.01$ ), indicating that participants performed overall better when the gestures were present compared to when they were absent. This predictor did not interact with any other predictors. There was also a significant main effect of Mouth Presence ( $b = 0.13$ ,  $SE = 0.02$ ,  $z = 5.81$ ,  $p < 0.001$ ) and Speech Clarity ( $b = 0.21$ ,  $SE = 0.02$ ,  $z = 8.59$ ,  $p < 0.001$ ), as well as the interaction between these two predictors ( $b = -0.08$ ,  $SE = 0.02$ ,  $z = -3.48$ ,  $p < 0.001$ ). Follow-up Bonferroni pairwise comparison showed that comprehension was particularly hindered when the speech was degraded, and mouth movements were covered by the mask, but equally good (as in the clear condition) when the mouth movements were visible (all  $p$ 's  $< 0.001$ ). There was no mask effect in the clear condition ( $p > 0.05$ ). The control variable of Question Type was also significant ( $p$ 's  $< 0.001$ ), with questions referring to main ideas being answered more accurately than questions about details, and implied main ideas being answered more accurately than stated main ideas. No other effects were significant. By-Participant intercept explained 32% of additional variance, whereas by-Story intercept explained 3.5% of additional variance. Slope of Speech Clarity was  $< 1\%$ , suggesting a similar relationship between accuracy performance and Speech Clarity variable for all

participants. Descriptive statistics are shown in Table 2, with full results presented in Table 3 (top panel) and plotted in Figure 3 (left panel).

The results from Experiment 1 show that naturally occurring co-speech gestures and speech-linked mouth movements influence how listeners comprehend spoken discourse. Specifically, we found that gestures facilitate comprehension across various listening conditions, while mouth movements are beneficial only when listeners process discourse in a noisy environment. This finding aligns with the idea that gestures and mouth movements tap into different processing levels. Gestures often operate at a higher level of semantics and pragmatics due to their iconicity, which supports the construction of situation models (Feyereisen, 2006; Murgiano et al., 2021). In contrast, mouth movements operate at a lower, sensorimotor level of phonology and phonetics. As such, observing mouth movements becomes particularly useful in noisy environments, as they help disambiguate speech sounds, thereby facilitating comprehension. The findings also potentially indicate a hierarchical organization of visual cues during discourse comprehension, with gestures showing greater influence than mouth movements (Hirata & Kelly, 2010). However, it is important to note that in the current experiment, the speaker was consistently in a quiet environment, and it remains unclear whether similar effects of gestures and mouth movements would be observed if the speaker were to produce these cues in a noisy environment (Trujillo et al., 2021).

Finally, the results concerning the question type provide some novel insight into how comprehenders process discourse. We found that listeners performed better when answering questions about implied main ideas compared to those explicitly stated, in line with a predominantly inferential and heuristic view of discourse processing (Kintsch, 1998; Kintsch & van Dijk, 1978; McNamara & Magliano, 2009). Indeed, in real-life conversations, speakers frequently leave “informational gaps” (Clark & Haviland, 1977) that listeners must fill by

constructing explicit or implicit inferences based on general knowledge and previous experiences to derive meaning (Graesser et al., 1994).

## EXPERIMENT 2

Experiment 2 assesses the impact of co-speech gestures, visible mouth movements, and listening conditions on discourse comprehension, when both the speaker and the listener are in the same communicative environment, i.e., either in a quiet or noisy environment. Investigating the role of multimodal cues in such a communicative environment is important, as both noisy and quiet conditions are very commonly occurring ones in face-to-face communication.

### Methods

#### *Participants*



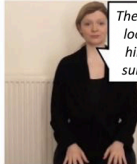

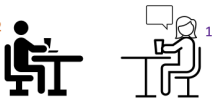





We conducted power calculations using *mixedpower* package (Kumle & Draschkow, 2021). The analysis consisted of building a mixed-effect regression model based on the coefficients from Experiment 1 and testing power for each effect size using different sample sizes. The results indicated that recruiting 100 participants will be enough to detect the interaction between mouth movements and speech clarity with more than 90% power (at alpha level of 0.05, with 1000 stimulations). As such, we recruited 100 native speakers of American English via Prolific (<http://www.prolific.co/>). As in Experiment 1, they were all right-handed monolinguals, with no language, hearing, vision, or neurological deficits. The data from four individuals were removed: One individual performed below 3SD from the group mean and three other individuals had average reaction time above 3SD from the group mean. The data from the remaining 96 participants (*Mean age* = 31.18, *SD* = 6.02, 47 females, 49 males) were used for the analyses.

## ***Materials***

The materials for Experiment 2 included the same 16 stories (eight DCT and eight cartoons) with eight yes-no comprehension questions per story. The stories were uttered anew by the same actress and were video-recorded with a Zoom Q4 camera at her house one year later. All narrated stories were presented in a randomized order to ensure that the actress maintained spontaneity and naturalness in her performance. This approach also aimed to prevent the actress from memorizing the stories and/or modifying her voice/movements in a predetermined manner. The actress narrated the stories with and without gestures, as well as with and without face mask, following the conditions in Experiment 1 and resulting in a total of 64 recorded stories. Additionally, the actress reproduced all 64 stories under the same gesture and mask conditions, but this time she also wore white, wireless earphones and listened to the 0dB SNR cafeteria background noise. Implementing this procedure was hypothesized to drive the speaker to modulate her voice and multimodal cues she produced in the noisy listening environment (eliciting a multimodal Lombard effect, e.g., Trujillo et al., 2021). Note that the background noise in the earphones is the same noise that was later used to degrade the audio for the degraded speech clarity condition (see more below). The volume level in the earphones was adjusted so that it was medium-to-very loud but did not create any discomfort to the actress. Given that participants were not informed about the actress experiencing noise through her earphones and considering the size of the earphones, we believe that this visual manipulation had no significant impact on participants' performance. Finally, the audio from the 64 videos where the actress was wearing earphones was extracted and cafeteria background noise of 0dB SNR was added, mimicking the steps introduced for Experiment 1. This was done to maintain consistency in the degraded speech condition to which participants were exposed across experiments. Participants watched all 16 stories and

were exposed to each condition twice. Figure 2 depicts experimental manipulations in Experiment 2.

**Figure 2** *Experimental conditions and the communicative environment from Experiment 2.*

EXPERIMENT 2					
Gesture Present		Gesture Absent		Communicative Environment	
Mouth Present	Mouth Absent	Mouth Present	Mouth Absent		
				 <p>1. Speaker in a <u>quiet</u> environment 2. Listener in a <u>quiet</u> environment</p>	
				 <p>1. Speaker in a <u>noisy</u> environment 2. Listener in a <u>noisy</u> environment</p>	

*Note:* Listener = Participant in the study; Speaker = Actress narrating stories. The yellow headphones icon indicates listening to cafeteria noise.

### ***Procedure***

Experiment 2 was built in Gorilla (<https://gorilla.sc/>) and followed the exact procedure as in Experiment 1 (see “Experiment 1, Methods, Procedure”).

### ***Data Analysis***

Following Experiment 1, we analyzed the data using a logistic mixed-effect regression to assess the effect of Gestures Presence, Mouth Presence, Speech Clarity, and their interactions (see “Experiment 1, Data Analysis and Results” for more details about the models). There was no multicollinearity (variance inflation factors <1.6) and outliers were removed (27 trials that were answered quicker than 200ms and 2 questions due to overall accuracy below 3SD from the mean).



## ***Results and Discussion of Experiment 2***

There was a significant main effect of Gesture Presence ( $b = 0.05$ ,  $SE = 0.02$ ,  $z = 2.14$ ,  $p=0.03$ ), such that participants performed better when the gestures were present compared to when they were absent. The effect of gesture is consistent with the one found in Experiment 1, however, with a smaller effect size. There was also a significant main effect of Mouth Presence ( $b = 0.12$ ,  $SE = 0.02$ ,  $z = 5.15$ ,  $p<0.001$ ) and Speech Clarity ( $b = 0.25$ ,  $SE = 0.03$ ,  $z = 8.65$ ,  $p<0.001$ ), as well as a marginal interaction between the two ( $b = -0.04$ ,  $SE = 0.02$ ,  $z = -1.71$ ,  $p=0.09$ ). Considering the main effects observed for Mouth Presence and Speech Clarity, which indicate that observing mouth movements is beneficial for listeners and that degraded speech is more challenging than clear speech, the marginal interaction further suggests that mouth movements are especially advantageous when speech is degraded as opposed to when it is clear ( $p<0.001$ ; Bonferroni pairwise comparison). This finding aligns with the significant interaction identified in Experiment 1. Finally, like in Experiment 1, the control variable of Question Type was also significant ( $p's < 0.001$ )<sup>2</sup>. Questions referring to main ideas were answered more accurately than questions about details, and implied main ideas were answered more accurately than stated main ideas. Participant intercept explained almost 30% of additional variance, and Story intercept explained almost 4%. Speech Clarity slope effect did not substantially differ between participants although it explained more variance than in Experiment 1 (2.7%). No other effects were significant. Descriptive statistics are shown in Table 2, with full results presented in Table 3 (bottom panel) and plotted in Figure 3 (right panel).

**Table 2**

---

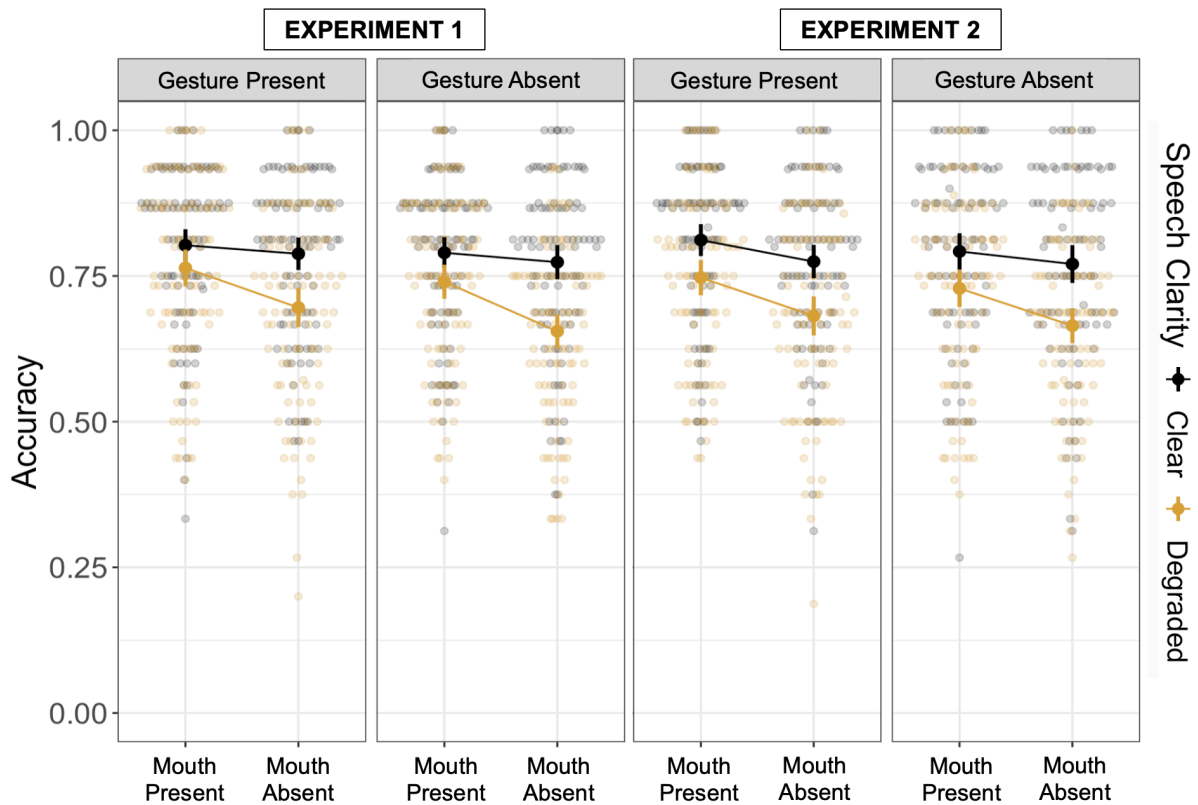
<sup>2</sup> We also tested models with Question Type in interactions with Gesture Presence, Mouth Presence, and Speech Clarity to further investigate the effect of Question Type on accuracy performance. Only the interaction between Mouth Presence1:QuestionType3 in Experiment 1 was significant ( $p=0.02$ ), such that participants performed better on DT-S (compared to the average effect of all question types) when mouth movements were present than absent. Full results from these models can be found on OSF.

Descriptive statistics (proportion correct and standard deviations in brackets) from Experiments 1 and 2.

Clear Speech				Degraded Speech			
Gesture Present		Gesture Absent		Gesture Present		Gesture Absent	
Mouth Present	Mouth Absent	Mouth Present	Mouth Absent	Mouth Present	Mouth Absent	Mouth Present	Mouth Absent
EXPERIMENT 1 (n=98)							
0.80 (0.40)	0.79 (0.41)	0.79 (0.41)	0.77 (0.42)	0.76 (0.43)	0.70 (0.46)	0.74 (0.44)	0.66 (0.48)
EXPERIMENT 2 (n=96)							
0.81 (0.40)	0.78 (0.42)	0.79 (0.41)	0.77 (0.42)	0.74 (0.44)	0.69 (0.47)	0.73 (0.45)	0.67 (0.47)

**Figure 3**

Plotted accuracy results from Experiments 1 (left) and 2 (right).



**Table 3**

Results of the mixed-effects logistic regression models for Experiment 1 (top) and Experiment 2 (bottom).

EXPERIMENT 1					
Random effects:	Variance	SD			
Participant(Intercept)	0.32	0.57			
SpeechClarity1	0.01	0.09			
Story(Intercept)	0.04	0.19			
Fixed effects:	b	SE	z	p	
(Intercept)	1.23	0.08	15.77	< 2e-16	***
GesturePresence1	0.07	0.02	2.99	0.00	**
MouthPresence1	0.13	0.02	5.81	0.00	***
SpeechClarity1	0.21	0.02	8.59	< 2e-16	***
GesturePresence1:MouthPresence1	-0.01	0.02	-0.44	0.66	
GesturePresence1:SpeechClarity1	-0.02	0.02	-0.80	0.42	
MouthPresence1:SpeechClarity1	-0.08	0.02	-3.48	0.00	***
GesturePresence1:MouthPresence1:SpeechClarity1	0.01	0.02	0.27	0.79	
Control variables:	b	SE	z	p	
StoryType1	0.02	0.05	0.45	0.65	
QuestionType1	0.44	0.04	10.71	< 2e-16	***
QuestionType2	-0.29	0.04	-7.90	0.00	***
QuestionType3	-0.30	0.04	-8.41	< 2e-16	***
EXPERIMENT 2					
Random effects:	Variance	SD			
Participant(Intercept)	0.30	0.54			
SpeechClarity1	0.03	0.16			
Story(Intercept)	0.04	0.20			
Fixed effects:	b	SE	z	p	
(Intercept)	1.19	0.08	15.17	< 2e-16	***
GesturePresence1	0.05	0.02	2.14	0.03	*
MouthPresence1	0.12	0.02	5.15	0.00	***
SpeechClarity1	0.25	0.03	8.65	< 2e-16	***
GesturePresence1:MouthPresence1	0.01	0.02	0.26	0.80	
GesturePresence1:SpeechClarity1	-0.01	0.02	-0.45	0.65	
MouthPresence1:SpeechClarity1	-0.04	0.02	-1.71	0.09	
GesturePresence1:MouthPresence1:SpeechClarity1	0.01	0.02	0.24	0.81	
Control variables:	b	SE	z	p	
StoryType1	0.02	0.05	0.44	0.66	
QuestionType1	0.42	0.04	10.22	< 2e-16	***
QuestionType2	-0.25	0.04	-6.47	0.00	***
QuestionType3	-0.29	0.04	-7.88	0.00	***

Note: Sum contrasts: GesturePresence1 = 'Present', MouthPresence1 = 'Present', SpeechClarity1 = 'Clear', StoryType1 = 'DCT', QuestionType1 = 'MI-I', QuestionType2 = 'DT-I', QuestionType3 = 'DT-S'. Significance levels: \*0.05; \*\*0.01; \*\*\*p<0.001

Experiment 2 successfully replicated the results of Experiment 1 for the clear condition and demonstrated similar effects for the degraded condition with new participants and re-recorded video-materials. We found that observing gestures is beneficial across noisy

and quiet listening conditions, and observing mouth movements is (marginally) more important in the noisy condition. Overall, these findings support the hypothesis that visual cues are weighted differently during discourse comprehension. As a reminder, the main difference between the experiments was that the speaker was situated in the same communicative environment as the listener in Experiment 2 but was always in a quiet environment in Experiment 1. This manipulation did not seem to affect the way visual cues impact discourse comprehension, as demonstrated by similar effect sizes (beta coefficients in Experiment 1 for Gesture Presence = 0.07, Mouth Presence x Speech Clarity = -0.08; beta coefficients in Experiment 2 for Gesture Presence = 0.05, Mouth Presence x Speech Clarity = -0.04)<sup>3</sup>. Finally, Experiment 2 also replicated the finding that comprehenders make inferences while processing discourse, as questions about implied main ideas were processed more accurately than stated ones.

## **FOLLOW-UP ANALYSES**

Across the experiments, we observed a robust, albeit small, effect of gestures on discourse comprehension. Since gestures were unscripted in our study, the speaker had the freedom to produce different types of gestures to different extents and it is not clear whether specific gestures (e.g., representational) may drive this effect. Thus, we conducted follow-up analyses to investigate the impact of gesture type, frequency, and amplitude on discourse comprehension.

---

<sup>3</sup> We also tested a logistic mixed-effect model with aggregated data from both experiments and an additional predictor of Experiment (Experiment 1 versus 2). The model showed a non-significant effect of Experiment ( $b = 0.02$ ,  $SE = 0.04$ ,  $z = 0.40$ ,  $p = 0.69$ ), providing little evidence that observing visual cues produced by a speaker-in-silence (Experiment 1) versus in-noise (Experiment 2) impacted comprehension differently. Full results are presented on the OSF.

## **Gesture Type and Frequency Analysis**

The analysis consisted of annotating all gestures in the videos (Gesture Present condition only) into one of three types: (a) representational and deictic gestures (including iconic and metaphoric gestures, abstract and concrete points, as well as emblems, i.e., conventionalized gestures such as thumbs up to represent an agreement) to describe their meaningful nature, (b) beat gestures, and (c) pragmatic gestures, following the lab coding manual (available upon request). Comparable analyses were conducted by Zhang et al. (2021), where they investigated the influence of spontaneously produced iconic, deictic, and beat gestures on discourse comprehension in neurotypical adults. We then tested a logistic mixed-effect regression model with gesture types (Representational/Deictic Gestures, Beat Gestures, and Pragmatic Gestures; all centered on the mean) as predictors, Story Type as a control variable, accuracy as the dependent variable, and random intercepts of Participant and Story. Only trials when the gestures were present were tested.

AK annotated all the gestures in Experiment 1 using ELAN (version 6.3, 2022). Coding reliability (10% of the total videos) between AK and a student assistant showed high Intraclass Correlation (Cronbach's  $\alpha > 0.9$ ). The actress produced on average 59 (SD=6.14) and 60 (SD=5.24) gestures per video when mouth movements were present versus absent, respectively. Most of the gestures were representational/deictic (43%), followed by pragmatic (32%), and beat gestures (25%). Numerically, there were more representational/deictic gestures produced during cartoon than DCT stories, and more beats than representational/deictic gestures during DCT than cartoon stories (see Figure 4). A greater frequency of representational/deictic gestures in cartoons compared to DCT was expected, given the inherently action-oriented nature of cartoons. Conversely, DCT stories, being more abstract, were expected to involve fewer instances of these gestures. The

statistical analysis showed no significant effects on accuracy performance (see Table 4), and thus, gesture coding was not carried out for Experiment 2.

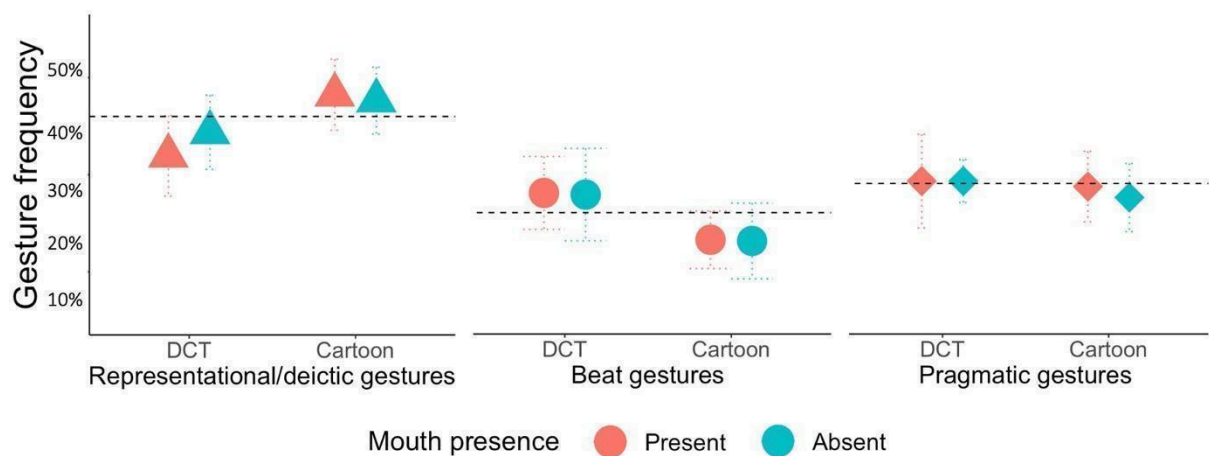
**Table 4**

*Results of the Gesture Type and Frequency Analysis.*

	<i>b</i>	<i>SE</i>	<i>z</i>	<i>p</i>
(Intercept)	1.27	0.07	17.44	<2e-16
MeaningfulGestures_z	0.01	0.05	0.22	0.83
BeatGestures_z	-0.02	0.05	-0.38	0.70
PragmaticGestures_z	0.06	0.05	1.28	0.20
StoryType1	0.00	0.06	-0.07	0.94

**Figure 4**

*Results of the Gesture Type and Frequency Analysis. The x-axis represents gesture type (representational/deictic gestures, beat gestures, pragmatic gestures) averaged by story type (DCT, cartoon) and y-axis represents gesture frequency (percentage). Color refers to the mouth presence condition (red=present, turquoise=absent). Error bars are standard deviations from the mean.*



## Gesture Amplitude Analysis

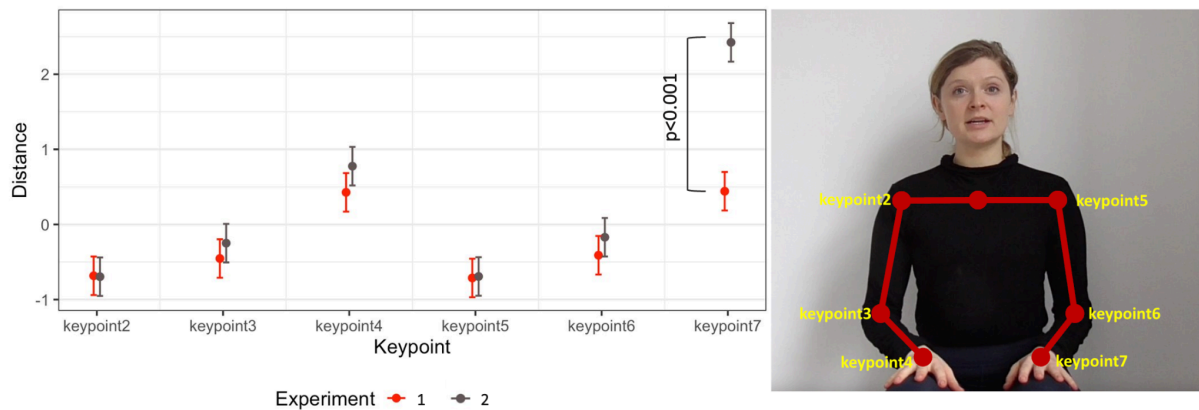
The goal of the Gesture-Amplitude analysis was to provide initial assessment of whether the size of gestures produced in Experiment 1 differed from those produced in Experiment 2. Similar analyses were run in, e.g., Trujillo et al. (2021) to assess the Lombard effect. We randomly selected one of the stories from our stimuli and analyzed the speaker's kinematics produced in silence (from a video from Experiment 1) and in noise (from a video from Experiment 2). The analysis consisted of an automatic 2D keypoint detection of movements of arms and hands using OpenPose (Cao et al., 2019). That is, we generated x and y coordinates for six of the body parts: right and left shoulders, elbows, and wrists. We then calculated the distance between the coordinates for each sentence (14 sentences per story) and for each frame. Larger distance indicates bigger movements. Finally, for each sentence, we took the average distance value for each keypoint, scaled it, and used it as our dependent variable in a linear mixed-effect regression analysis (with the *lme4* package; Bates et al., 2015). The predictors included: Keypoint (with six levels, corresponding to the x and y coordinates for the six body parts), Experiment (Experiment 1 versus 2), and the interaction between the two. We also entered Sentence as a random intercept. The categorical variables were dummy coded.

The results of the Gesture-Amplitude analysis showed a significant effect of Keypoints 4 (right wrist;  $b = 1.11$ ,  $SE = 0.17$ ,  $t = 6.38$ ,  $p < 0.001$ ) and 7 (left wrist;  $b = 1.13$ ,  $SE = 0.17$ ,  $t = 6.46$ ,  $p < 0.001$ ) with larger distance relative to other keypoints. Crucially, there was also a significant interaction between Keypoint and Experiment ( $b = 1.99$ ,  $SE = 0.25$ ,  $t = 8.09$ ,  $p < 0.001$ ) with larger distance for Keypoint 7 (left wrist) in Experiment 2 than Experiment 1. Overall, these results showed that the speaker produced bigger movements (using her left hand) when in noise, suggesting that she made gestural adjustments to deal with challenging listening conditions. More broadly, these post-hoc analyses provided initial

evidence for the multimodal Lombard effect (Trujillo et al., 2021). Figure 5 depicts the findings.

## Figure 5

*Results of the Gesture-Amplitude Analysis. The left panel: The x-axis represents different keypoints used in the analysis and the y-axis is the scaled distance. Colors represent experiment type. Error bars are standard errors of the mean. The right panel: schematic representation of keypoints annotation.*



These follow-up analyses yield two key findings. First, the speaker produced a variety of gestures while narrating stories, but the type and frequency of these gestures did not significantly impact comprehension scores. Second, the speaker used her non-dominant hand more often when producing the stories in noise. This finding is in line with the multimodal Lombard effect (Trujillo et al., 2021).

## GENERAL DISCUSSION

The present study is the first to investigate the benefit of multimodal cues, including gestures and mouth movements, for auditory discourse comprehension under clear and challenging (with background noise) listening conditions. We presented videos of a speaker



who, in half of the instances, spontaneously gestured and wore a health mask that covered communicative facial cues while narrating stories. This design aimed to resemble everyday communicative settings. We investigated the impact of interactions between visual cues and listening conditions on listeners' comprehension. Additionally, we looked at whether any potential multimodal benefit to comprehension depends on the communicative environment.

Across two experiments, we found that comprehending discourse is easier when spontaneous co-speech gestures accompany speech. This effect was small but robust and did not depend on gesture frequency/type, speech clarity, or mouth presence. We also showed that mouth movements support discourse comprehension. This effect, however, was most prominent in challenging listening conditions. Finally, we demonstrated that the multimodal benefit to comprehension was similar regardless of the communicative environment the speaker was situated in.

### **Multimodal Cues Matter to Listeners (but to a Different Extent)**

Multimodal cues, such as gestures and mouth movements, are part-and-parcel of face-to-face communication. Our findings support this view by demonstrating that these cues matter to listeners: They improve discourse comprehension, but they do so to a different extent.

Spontaneously produced co-speech gestures benefit discourse comprehension regardless of the presence of other cues or the clarity of the speech signal. This finding was observed in Experiment 1 and was then demonstrated in Experiment 2. It is also in line with previous studies on discourse comprehension showing beneficial (albeit small to medium) effect of gestures (for a meta-analysis see Dargue et al., 2019). Such gestural benefit to comprehension has primarily been attributed to typical iconic (Dargue & Sweller, 2018, 2020; McKern et al., 2021) and deictic (Gunter et al., 2015; Sekine & Kita, 2015, 2017; Smith & Kam, 2012) gestures, which were also the most frequently produced types of

gestures in this study. These gestures are known for their role in supporting creation of situation models (e.g., Clark, 1996; McNeill, 1992; Cutica & Bucciarelli, 2013), by engaging prior perceptual-motor experiences to simulate embodied events (e.g., Glenberg et al., 2013; Glenberg & Gallese, 2012; Zwaan, 2014), and the more informative they are, the more benefit to comprehension there is (Dargue & Sweller, 2018; Krason et al., 2021; Zhang et al., 2021).

In this study, other types of co-speech gestures were also present, and their role in supporting linguistic processing has been noted, albeit differently than representational or deictic gestures. For example, pragmatic gestures—the second most common type of gestures here—may have facilitated discourse processing by maintaining an engaging and interactive character of the narratives (Bavelas et al., 1992; 1995) or by marking specific structures or functions of discourse (e.g., indicating implied information or intensifying an argument; Kendon, 2004). The benefit of beat gestures—the least frequent gestures (although note that there were more beat gestures in the DCT stories than in the cartoon stories)—to comprehension is still debatable. Some have argued that they impact auditory comprehension to a lesser extent than representational gestures (Feyeresen, 2006; Zhang et al., 2021; Macoun & Sweller, 2016) and others have suggested that beat gestures might only be useful in processing the word to which they are time-locked to rather than any other information around (Igualada et al., 2017). Here, we have demonstrated that unscripted, spontaneously produced co-speech gestures collectively contribute to discourse comprehension benefit.

While we did not find evidence that the frequency, type, and amplitude of gestures differentially impacted discourse comprehension, future studies could focus on better quantifying the informativeness of various gestures in relation to discourse comprehension. Researchers could also investigate the influence of gestures produced by different individuals to better understand how gesture idiosyncrasy impacts comprehension.

Although some studies have shown that co-speech (iconic) gestures are particularly helpful in challenging listening conditions (Drijvers & Özyürek, 2017; Holle et al., 2010; Krason et al., 2021; Obermeier et al., 2012), our results demonstrate that listeners always take gestures into account during face-to-face discourse even when listening conditions are clear. We propose that the different results can be explained in terms of task difficulty. That is, performance on word/gesture comprehension tasks (e.g., free recall, Drijvers & Özyürek, 2017; picture verification, Krason et al., 2021; or assessing congruency between gesture and homonymous words, Holle et al., 2010) is most often at ceiling when the speech is clear so the effect of gestures could only be observed when the task is more challenging (i.e., when the speech was degraded). This is not the case in our discourse comprehension task; thus, we observe the effect of gestures in both clear and degraded conditions.

Additional evidence comes from recent studies by McKern et al. (2022) and Wilms et al. (2022), who showed that iconic gestures benefit narrative recall and sentence recognition, respectively, regardless of the noise. These findings, along with ours, jointly support the speech-gesture integration account according to which co-speech gestures are automatically and obligatorily processed alongside speech (e.g., Kelly et al., 2010; McNeill et al., 1994). Zhang et al. (2021) further showed that representational and beat gestures (as well as prosody and mouth movements) modulate the incoming linguistic information in naturalistic settings, and our study extends their finding by demonstrating that these modulations lead to quantifiable improvements in discourse comprehension.

In contrast to co-speech gestures, mouth movements are only used when speech is challenging. This effect was consistent across experiments and is well-established in the literature (e.g., Sumby & Pollack, 1954; Ross et al., 2007; Ma et al. 2009). Studies have suggested that seeing a speaker's mouth movements supports phonological processing and constrains lexical competition, which is crucial in adverse listening conditions when the

auditory channel provides only sub-optimal acoustic information for phonological processing during comprehension (for a review see Peelle & Sommers, 2015). Our study has further demonstrated that the audiovisual speech enhancement in challenging listening conditions holds even with naturalistic manipulation of mouth presence, i.e., when a speaker wears a surgical mask. In contrast, a similar close coupling between listening conditions and gestures is less probable as gesture primarily supports semantic and pragmatic processing levels.

A number of recent studies has investigated the effect of masks on speech recognition. Toscano and Toscano (2021) examined the impact of surgical mask, two different cloth masks, and a N95 respirator on auditory-only sentence recognition under different noise levels. They found that wearing a mask had little impact on the accuracy performance under low level noise, but it mattered, and depended on a mask type, under high level noise. Interestingly, the surgical mask was the only type that did not show an effect under any noise conditions. Building on these findings, Brown et al. (2021) further demonstrated that wearing a mask (regardless of its type) makes sentence comprehension in noise more challenging than when mouth movements are visible. Here, we showed a similar mask effect for discourse comprehension. Given that Toscano and Toscano (2021) found that the use of surgical masks has no impact on auditory processing of speech in noise, the effect of mask found in our study can be attributed to the fact that the mask obscured information from visual speech, rather than muffled spoken signal.

The finding that mouth movements are primarily helpful when speech is degraded but gestures benefit comprehension across different speech clarity conditions suggests that these visual cues are flexibly weighted by listeners (Krasen et al., 2021; Zhang et al., 2021) because they support linguistic processes differently (Hirata & Kelly, 2010). That is, whereas mouth movements support temporal and phonological encoding of the incoming speech, co-speech gestures support processing of high-level information (e.g., semantic encoding)

and listeners will use a cue or a combination of cues that is the most informative in a given context (Krason et al., 2021; Zhang et al., 2021). Some researchers have also suggested a “double enhancement” effect, i.e., greater benefit to spoken word comprehension when both gesture and mouth cues are present, particularly in challenging listening conditions (Drijvers & Özyürek, 2017). Here we showed that when co-speech gestures convey sufficient information to correctly interpret spoken passages, mouth movements become less important.

Finally, we also contrasted two possible communicative environments, in which cues can be produced and perceived. One scenario assumes that a speaker produces cues in a quiet environment, but listeners perceive them in noise (e.g., watching TV news in a noisy café; Experiment 1) and the other scenario depicts a situation where a speaker produces cues in noise and listeners also perceive them in noise (e.g., having a conversation in a noisy café; Experiment 2). Such manipulation was introduced to investigate whether the multimodal benefit depends on the listening conditions in which the cues are produced. Although we found that the speaker produced bigger movements with her left (non-dominant) hand in Experiment 2 than in Experiment 1, we did not find evidence that this gestural adjustment had a significant effect on comprehension (see Footnote 3 and effect sizes). Larger left-hand movements potentially suggest that the speaker was trying to compensate for the noise disruption (which is in line with the multimodal Lombard effect; Trujillo et al., 2021). However, since the speaker did not engage in direct interaction with the listener (participants), it remains unclear to what extent this absence of interaction may account for the observed lack of additional benefit to comprehension. For instance, in studies such as Trujillo et al. (2021) and Garnier et al. (2018), listeners actively participated in a conversation with the speaker, potentially enhancing multimodal modulations and their dynamic weighting. It is also possible that the relatively small size of the videos may have prevented participants from noticing the difference in gesture amplitude, resulting in null findings.

## **Advancing Naturalistic Language Comprehension Models**

The work presented in this paper is a building block in the development of naturalistic language comprehension models (Hasson et al., 2018; Nastase et al., 2021; Skipper, 2015; Vigliocco et al., 2024). It advances our understanding of discourse processing by extending findings on the benefits of visual cues, including various types of gestures, such as representational, deictic, beat, and pragmatic gestures, as well as mouth movements, to (semi-) naturalistic comprehension. We challenged the assumption that gestures are extraneous rather than central to language comprehension by demonstrating that listeners consistently use the information from gestures during discourse comprehension, even in clear listening conditions. Our findings also align with the notion that co-speech gestures often modulate the semantic or pragmatic level of processing, while mouth movements tap onto phonetics/phonology, providing insights into how and when gestures and mouth movements support discourse processing. Specifically, listeners use gestures across various listening conditions to support the creation of situation models and use mouth movements to disambiguate speech in challenging listening conditions.

Studying interactions between communicative cues during discourse processing poses several challenges. There are large individual differences in how much people benefit from gestures and mouth movements and it is unclear if similar multimodal benefits to discourse would be found for other populations. For example, individuals with poorer memory abilities extract less communicative information from gestures (McKern et al., 2021; Schubotz et al., 2020), and second language learners tend to benefit less than native speakers (Drijvers & Özyürek, 2020; Zhang, Ding, et al., 2021). Similar findings were observed for people with post-stroke aphasia, who, despite benefitting from visible mouth movements, showed a smaller effect than age-matched individuals without aphasia (Krasn et al., 2023). Moreover, co-speech gestures and mouth movements are not the only communicative cues during

face-to-face interactions. Studies have shown that, for instance, word predictability and prosody similarly modulate brain activity during naturalistic discourse comprehension (Zhang et al., 2021). Researchers have also demonstrated a strong link between other cues, such as beat gestures and prosody (Esteve-Gibert & Prieto, 2013), pragmatic gestures and negation movements (e.g., headshake; Kendon, 2004), eye gaze and co-speech gestures (Gullberg & Holmqvist, 1999), as well as between nodding and blinking (Hömke et al., 2018). All these cues are integral to face-to-face communication, and situating language in a physical and communicative environment is crucial for a thorough understanding of language processing *in situ* (Holler, 2022; Holler & Levinson, 2019; Levinson & Holler, 2014; Reggin et al., 2023; Murgiano et al., 2021). The current study supports this view.

## ACKNOWLEDGMENTS

We thank Prof. Stuart Rosen for his expertise and support with the auditory signal degradation. We would also like to thank Asude Eracikbas for her help with reliability coding. The work presented in this article was supported by the European Research Council Advanced Grant (ECOLANG, 743035) and the Royal Society Wolfson Research Merit Award (WRM\R3\170016) awarded to Prof. Gabriella Vigliocco.

## AUTHOR NOTE

Authors declare no conflict of interest.

The experiments were preregistered using <https://aspredicted.org/>. The preregistrations, data, example materials, and the R code are publicly available on the Open Science Framework (OSF): <https://osf.io/6zxuw/>.

## REFERENCES

- Alibali, M. W., Kita, S., & Young, A. J. (2000). Gesture and the process of speech production: We think, therefore we gesture. *Language and Cognitive Processes*, 15(6), 593–613. <https://doi.org/10.1080/016909600750040571>
- Arnold, P., & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology (London, England: 1953)*, 92 Part 2, 339–355.
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Bates, D. M. (2015.). *lme4: Mixed-effects modeling with R*. 145.
- Bavelas, J. B., Chovil, N., Coates, L., & Roe, L. (1995). Gestures specialized for dialogue. *Personality and Social Psychology Bulletin*, 21(4), 394–405. <https://doi.org/10.1177/0146167295214010>
- Bavelas, J. B., Chovil, N., Lawrie, D. A., & Wade, A. (1992). Interactive gestures. *Discourse Processes*, 15(4), 469–489. <https://doi.org/10.1080/01638539209544823>
- Brown, A., & Chen, J. (2013). Construal of Manner in speech and gesture in Mandarin, English, and Japanese. *Cognitive Linguistics*, 24(4), 605–631. <https://doi.org/10.1515/cog-2013-0021>
- Brown, A., & Gullberg, M. (2008). Bidirectional crosslinguistic influence in L1-L2 encoding of Manner in speech and gesture: A study of Japanese speakers of English. *Studies in Second Language Acquisition*, 30, 225–251. <https://doi.org/10.1017/S0272263108080327>
- Brown, V. A., Van Engen, K. J., & Peelle, J. E. (2021). Face mask type affects audiovisual speech intelligibility and subjective listening effort in young and older adults.



*Cognitive Research: Principles and Implications*, 6(1), 49.

<https://doi.org/10.1186/s41235-021-00314-0>

Cao, Z., Martinez, G., Simon, T., Wei, S.-E., & Sheikh, Y. (2019). OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP, 1–1.

<https://doi.org/10.1109/TPAMI.2019.2929257>

Carroll J. B., Davies, P., & Richman, B. (1971). Word Frequency Book. New York: American Heritage Publishing Co., Inc.

Clark, H. H. (1996). *Using Language*. Cambridge University Press.

<https://doi.org/10.1017/CBO9780511620539>

Cutica, I., & Bucciarelli, M. (2013). Cognitive change in learning from text: Gesturing enhances the construction of the text mental model. *Journal of Cognitive Psychology*, 25(2), 201–209. <https://doi.org/10.1080/20445911.2012.743987>

Dargue, N., & Sweller, N. (2018). Not all gestures are created equal: The effects of typical and atypical iconic gestures on narrative comprehension. *Journal of Nonverbal Behavior*, 42(3), 327–345. <https://doi.org/10.1007/s10919-018-0278-3>

Dargue, N., & Sweller, N. (2020). Two hands and a tale: When gestures benefit adult narrative comprehension. *Learning and Instruction*, 68, 101331.

<https://doi.org/10.1016/j.learninstruc.2020.101331>

Dargue, N., Sweller, N., & Jones, M. P. (2019). When our hands help us understand: A meta-analysis into the effects of gesture on comprehension. *Psychological Bulletin*, 145(8), 765–784. <https://doi.org/10.1037/bul0000202>

Drijvers, L., & Özyürek, A. (2017). Visual Context Enhanced: The Joint Contribution of Iconic Gestures and Visible Speech to Degraded Speech Comprehension. *Journal of*

- Speech, Language, and Hearing Research*, 60(1), 212–222.  
[https://doi.org/10.1044/2016\\_JSLHR-H-16-0101](https://doi.org/10.1044/2016_JSLHR-H-16-0101)
- Fang, I. E. (1966). The “Easy listening formula”. *Journal of Broadcasting*, 11(1), 63–68.  
<https://doi.org/10.1080/08838156609363529>
- Feyereisen, P. (2006). Further investigation on the mnemonic effect of gestures: Their meaning matters. *European Journal of Cognitive Psychology*, 18(2), 185–205.  
<https://doi.org/10.1080/09541440540000158>
- Garnier, M., Henrich, N., & Dubois, D. (2010). Influence of Sound Immersion and Communicative Interaction on the Lombard Effect. *Journal of Speech, Language, and Hearing Research*, 53(3), 588–608. [https://doi.org/10.1044/1092-4388\(2009/08-0138\)](https://doi.org/10.1044/1092-4388(2009/08-0138))
- Garnier, M., Ménard, L., & Alexandre, B. (2018). Hyper-articulation in Lombard speech: An active communicative strategy to enhance visible speech cues? *The Journal of the Acoustical Society of America*, 144(2), 1059–1074. <https://doi.org/10.1121/1.5051321>
- Giovanelli, E., Valzolgher, C., Gessa, E., Todeschini, M., & Pavani, F. (2021). Unmasking the Difficulty of Listening to Talkers With Masks: Lessons from the COVID-19 pandemic. *I-Perception*, 12(2), 2041669521998393.  
<https://doi.org/10.1177/2041669521998393>
- Glenberg, A. M., & Gallese, V. (2012). Action-based language: A theory of language acquisition, comprehension, and production. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, 48(7), 905–922.  
<https://doi.org/10.1016/j.cortex.2011.04.010>
- Glenberg, A. M., Witt, J. K., & Metcalfe, J. (2013). From the Revolution to Embodiment: 25 Years of Cognitive Psychology. *Perspectives on Psychological Science*, 8(5), 573–585. <https://doi.org/10.1177/1745691613498098>

- Goodwin, C. (1981). *Conversational Organization: Interaction Between Speakers and Hearers*.
- Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, 108(3 Pt 1), 1197–1208. <https://doi.org/10.1121/1.1288668>
- 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(10), 3309–3324. <https://doi.org/10.1002/hbm.20753>
- Gullberg, M., & Kita, S. (2009). Attention to Speech-Accompanying Gestures: Eye Movements and Information Uptake. *Journal of Nonverbal Behavior*, 33(4), 251–277. <https://doi.org/10.1007/s10919-009-0073-2>
- Gunter, T. C., Weinbrenner, J. E. D., & Holle, H. (2015). Inconsistent use of gesture space during abstract pointing impairs language comprehension. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.00080>
- Hasson, U., Egidi, G., Marelli, M., & Willems, R. M. (2018). Grounding the neurobiology of language in first principles: The necessity of non-language-centric explanations for language comprehension. *Cognition*, 180, 135–157. <https://doi.org/10.1016/j.cognition.2018.06.018>
- Hirata, Y., & Kelly, S. D. (2010). Effects of Lips and Hands on Auditory Learning of Second-Language Speech Sounds. *Journal of Speech, Language, and Hearing Research*, 53(2), 298–310. [https://doi.org/10.1044/1092-4388\(2009/08-0243\)](https://doi.org/10.1044/1092-4388(2009/08-0243))
- Holle, H., Obleser, J., Rueschemeyer, S.-A., & Gunter, T. C. (2010). Integration of iconic gestures and speech in left superior temporal areas boosts speech comprehension under adverse listening conditions. *NeuroImage*, 49(1), 875–884. <https://doi.org/10.1016/j.neuroimage.2009.08.058>

- Homans, N. C., & Vroegop, J. L. (2021). Impact of face masks in public spaces during COVID-19 pandemic on daily life communication of cochlear implant users. *Laryngoscope Investigative Otolaryngology*, 6(3), 531–539.  
<https://doi.org/10.1002/lio2.578>
- Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological Bulletin*, 137(2), 297–315. <https://doi.org/10.1037/a0022128>
- Hubbard, A. L., Wilson, S. M., Callan, D. E., & Dapretto, M. (2009). Giving speech a hand: Gesture modulates activity in auditory cortex during speech perception. *Human Brain Mapping*, 30(3), 1028–1037. <https://doi.org/10.1002/hbm.20565>
- Igualada, A., Esteve-Gibert, N., & Prieto, P. (2017). Beat gestures improve word recall in 3- to 5-year-old children. *Journal of Experimental Child Psychology*, 156, 99–112.  
<https://doi.org/10.1016/j.jecp.2016.11.017>
- Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4), 434–446.  
<https://doi.org/10.1016/j.jml.2007.11.007>
- Junqua, J.-C. (1993). The Lombard reflex and its role on human listeners and automatic speech recognizers. *Journal of the Acoustical Society of America*, 93(1), 510–524.  
<https://doi.org/10.1121/1.405631>
- 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(2), 260–267. <https://doi.org/10.1177/0956797609357327>
- Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance in M. *The Relationship of Verbal and Nonverbal Communication*, 25.
- Kendon, A. (1988). How gestures can become like words. In *Cross-cultural perspectives in nonverbal communication*. (pp. 131–141). Hogrefe & Huber Publishers.

- Kendon, A. (2004). *Gesture: Visible Action as Utterance*.  
<https://doi.org/10.1017/CBO9780511807572>
- Kita, S., Gijn, I. van, & Hulst, H. van der. (1997). Movement phases in signs and co-speech gestures, and their transcription by human coders. *Gesture and Sign Language in Human-Computer Interaction*, 23–35. <https://doi.org/10.1007/BFb0052986>
- Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, 48(1), 16–32.  
[https://doi.org/10.1016/S0749-596X\(02\)00505-3](https://doi.org/10.1016/S0749-596X(02)00505-3)
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.
- Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85(5), 363–394.  
<https://doi.org/10.1037/0033-295X.85.5.363>
- Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, 57(3), 396–414. <https://doi.org/10.1016/j.jml.2007.06.005>
- Krason, A., Fenton, R., Varley, R., & Vigliocco, G. (2021). The role of iconic gestures and mouth movements in face-to-face communication. *Psychonomic Bulletin & Review*.  
<https://doi.org/10.3758/s13423-021-02009-5>
- Krason, A., Varley, R., & Vigliocco, G. (2024, June 25). Multimodal discourse comprehension. Retrieved from [osf.io/6z xu w](https://osf.io/6z xu w)
- Krason, A., Vigliocco, G., Mailend, M.-L., Stoll, H., Varley, R., & Buxbaum, L. J. (2023). Benefit of visual speech information for word comprehension in post-stroke aphasia. *Cortex*. <https://doi.org/10.1016/j.cortex.2023.04.011>

- Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, 62, 621–647. <https://doi.org/10.1146/annurev.psych.093008.131123>
- Lachs, L., & Pisoni, D. B. (2004). Cross-Modal Source Information and Spoken Word Recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 30(2), 378–396. <https://doi.org/10.1037/0096-1523.30.2.378>
- Lüdecke, D., Bartel, A., Schwemmer, C., Powell, C., Djalovski, A., & Titz, J. (2022). *sjPlot: Data Visualization for Statistics in Social Science* (2.8.11). <https://CRAN.R-project.org/package=sjPlot>
- Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-Reading Aids Word Recognition Most in Moderate Noise: A Bayesian Explanation Using High-Dimensional Feature Space. *PLoS ONE*, 4(3), e4638. <https://doi.org/10.1371/journal.pone.0004638>
- Macoun, A., & Sweller, N. (2016). Listening and watching: The effects of observing gesture on preschoolers' narrative comprehension. *Cognitive Development*, 40, 68–81. <https://doi.org/10.1016/j.cogdev.2016.08.005>
- Malik, S., Mihm, B., & Reichelt, M. (2021). The impact of face masks on interpersonal trust in times of COVID-19. *Scientific Reports*, 11, 17369. <https://doi.org/10.1038/s41598-021-96500-7>
- McKern, N., Dargue, N., Sweller, N., Sekine, K., & Austin, E. (2021). Lending a hand to storytelling: Gesture's effects on narrative comprehension moderated by task difficulty and cognitive ability. *Quarterly Journal of Experimental Psychology*, 74(10), 1791–1805. <https://doi.org/10.1177/17470218211024913>

- McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of comprehension. In *The psychology of learning and motivation*, Vol. 51 (pp. 297–384). Elsevier Academic Press. [https://doi.org/10.1016/S0079-7421\(09\)51009-2](https://doi.org/10.1016/S0079-7421(09)51009-2)
- McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press.
- McNeill, D., Cassell, J., & McCullough, K.-E. (1994). Communicative Effects of Speech-Mismatched Gestures. *Research on Language & Social Interaction*, 27(3), 223–237. [https://doi.org/10.1207/s15327973rlsi2703\\_4](https://doi.org/10.1207/s15327973rlsi2703_4)
- McNeill, D., & Levy, E. (1982). *Conceptual Representations in Language Activity and Gesture*.
- Melinger, A., & Levelt, W. J. M. (2004). Gesture and the communicative intention of the speaker. *Gesture*, 4(2), 119–141. <https://doi.org/10.1075/gest.4.2.02mel>
- Mheidly, N., Fares, M. Y., Zalzale, H., & Fares, J. (2020). Effect of Face Masks on Interpersonal Communication During the COVID-19 Pandemic. *Frontiers in Public Health*, 8. <https://www.frontiersin.org/articles/10.3389/fpubh.2020.582191>
- Murgiano, M., Motamedi, Y., & Vigliocco, G. (2021). Situating Language in the Real-World: The Role of Multimodal Iconicity and Indexicality. *Journal of Cognition*, 4(1), 38. <https://doi.org/10.5334/joc.113>
- Obermeier, C., Dolk, T., & Gunter, T. C. (2012). The benefit of gestures during communication: Evidence from hearing and hearing-impaired individuals. *Cortex*, 48(7), 857–870. <https://doi.org/10.1016/j.cortex.2011.02.007>
- Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, 68, 169–181. <https://doi.org/10.1016/j.cortex.2015.03.006>
- Pycha, A., Cohn, M., & Zellou, G. (2022). Face-Masked Speech Intelligibility: The Influence

- of Speaking Style, Visual Information, and Background Noise. *Frontiers in Communication*, 7, 874215. <https://doi.org/10.3389/fcomm.2022.874215>
- Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In *Hearing by eye: The psychology of lip-reading* (pp. 97–113). Lawrence Erlbaum Associates, Inc.
- Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex (New York, N.Y.: 1991)*, 17(5), 1147–1153. <https://doi.org/10.1093/cercor/bhl024>
- Ross, P., & George, E. (2022). Are Face Masks a Problem for Emotion Recognition? Not When the Whole Body Is Visible. *Frontiers in Neuroscience*, 16. <https://www.frontiersin.org/articles/10.3389/fnins.2022.915927>
- Saunders, G. H., Jackson, I. R., & Visram, A. S. (2021). Impacts of face coverings on communication: An indirect impact of COVID-19. *International Journal of Audiology*, 60(7), 495–506. <https://doi.org/10.1080/14992027.2020.1851401>
- Schegloff, E. (1984). On some questions and ambiguities in conversation. *Structures of Social Action*, 28–52.
- Schwartz, J.-L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: Evidence for early audio-visual interactions in speech identification. *Cognition*, 93(2), B69–B78. <https://doi.org/10.1016/j.cognition.2004.01.006>
- Sekine, K., & Kita, S. (2015). Development of multimodal discourse comprehension: Cohesive use of space by gestures. *Language, Cognition and Neuroscience*, 30(10), 1245–1258. <https://doi.org/10.1080/23273798.2015.1053814>



- Sekine, K., & Kita, S. (2017). The listener automatically uses spatial story representations from the speaker's cohesive gestures when processing subsequent sentences without gestures. *Acta Psychologica*, 179, 89–95. <https://doi.org/10.1016/j.actpsy.2017.07.009>
- Skipper, J. I., Goldin-Meadow, S., Nusbaum, H. C., & Small, S. L. (2009). Gestures Orchestrate Brain Networks for Language Understanding. *Current Biology*, 19(8), 661–667. <https://doi.org/10.1016/j.cub.2009.02.051>
- Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, 26(2), 212–215. <https://doi.org/10.1121/1.1907309>
- Toscano, J. C., & Toscano, C. M. (2021). Effects of face masks on speech recognition in multi-talker babble noise. *PLOS ONE*, 16(2), e0246842. <https://doi.org/10.1371/journal.pone.0246842>
- Trujillo, J., Özyürek, A., Holler, J., & Drijvers, L. (2021). Speakers exhibit a multimodal Lombard effect in noise. *Scientific Reports*, 11(1), 16721. <https://doi.org/10.1038/s41598-021-95791-0>
- Tye-Murray, N., Sommers, M., & Spehar, B. (2007). Auditory and Visual Lexical Neighborhoods in Audiovisual Speech Perception. *Trends in Amplification*, 11(4), 233–241. <https://doi.org/10.1177/1084713807307409>
- van Dijk, T. A., & Kintsch, W. (1983). Strategies of discourse comprehension. Academic Press.
- Vigliocco, G., Convertino, L., Felice, S. D., Gregorians, L., Kewenig, V., Mueller, M. A. E., Veselic, S., Musolesi, M., Hudson-Smith, A., Tyler, N., Flouri, E., & Spiers, H. (2024). Ecological Brain: Reframing the Study of Human Behaviour and Cognition. <https://doi.org/10.31234/osf.io/zr4nm>

- Whitney Goodrich Smith, & Carla L. Hudson Kam. (2012). Knowing ‘who she is’ based on ‘where she is’: The effect of co-speech gesture on pronoun comprehension. *Language and Cognition*, 4(2), 75–98. <https://doi.org/10.1515/langcog-2012-0005>
- Wilms, V., Drijvers, L., & Brouwer, S. (2022). The Effects of Iconic Gestures and Babble Language on Word Intelligibility in Sentence Context. *Journal of Speech, Language, and Hearing Research: JSLHR*, 65(5), 1822–1838. [https://doi.org/10.1044/2022\\_JSLHR-21-00387](https://doi.org/10.1044/2022_JSLHR-21-00387)
- Zhang, Y., Frassinelli, D., Tuomainen, J., Skipper, J. I., & Vigliocco, G. (2021). More than words: Word predictability, prosody, gesture and mouth movements in natural language comprehension. *Proceedings of the Royal Society B: Biological Sciences*, 288(1955), 20210500. <https://doi.org/10.1098/rspb.2021.0500>
- Zwaan, R. A. (2003). *THE IMMERSED EXPERIENCER: TOWARD AN EMBODIED THEORY OF LANGUAGE COMPREHENSION*. 28.
- Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The Construction of Situation Models in Narrative Comprehension: An Event-Indexing Model. *Psychological Science*, 6(5), 292–297. <https://doi.org/10.1111/j.1467-9280.1995.tb00513.x>
- Zwaan, R. A., Magliano, J. P., & Graesser, A. C. (1995). Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(2), 386–397. <https://doi.org/10.1037/0278-7393.21.2.386>
- Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2), 162–185. <https://doi.org/10.1037/0033-2909.123.2.162>