Cognition 248 (2024) 105806

Contents lists available at ScienceDirect

Cognition

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

Visual bodily signals and conversational context benefit the anticipation of turn ends

Marlijn ter Bekke [a,b], Stephen C. Levinson [b], Lina van Otterdijk [a], Michelle Kühn [a], Judith Holler [a,b,*]

[a] Donders Institute for Brain, Cognition & Behaviour, Radboud University, Nijmegen, the Netherlands
[b] Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands

ARTICLE INFO

Keywords:
Turn-taking
Turn end anticipation
Visual bodily signals
Conversation
Discourse context

ABSTRACT

The typical pattern of alternating turns in conversation seems trivial at first sight. But a closer look quickly reveals the cognitive challenges involved, with much of it resulting from the fast-paced nature of conversation. One core ingredient to turn coordination is the anticipation of upcoming turn ends so as to be able to ready oneself for providing the next contribution. Across two experiments, we investigated two variables inherent to face-to-face conversation, the presence of visual bodily signals and preceding discourse context, in terms of their contribution to turn end anticipation. In a reaction time paradigm, participants anticipated conversational turn ends better when seeing the speaker and their visual bodily signals than when they did not, especially so for longer turns. Likewise, participants were better able to anticipate turn ends when they had access to the preceding discourse context than when they did not, and especially so for longer turns. Critically, the two variables did not interact, showing that visual bodily signals retain their influence even in the context of preceding discourse. In a pre-registered follow-up experiment, we manipulated the visibility of the speaker's head, eyes and upper body (i.e. torso + arms). Participants were better able to anticipate turn ends when the speaker's upper body was visible, suggesting a role for manual gestures in turn end anticipation. Together, these findings show that seeing the speaker during conversation may critically facilitate turn coordination in interaction.

1. Introduction

Conversation is the most common form of human linguistic exchange but still rather poorly understood in terms of the precise cognitive mechanisms that underpin it. A major conundrum is the fact that, on the surface, conversation is characterized by smooth transitions between speakers with very small latencies between speaking turns (Sacks, Schegloff, & Jefferson, 1974; Stivers et al., 2009) hardly perceivable to the human ear (Heldner, 2011). And yet, the psycholinguistic processes giving rise to conversational turn-taking are complex. Since the production of individual words takes around 600–1200 ms in picture naming tasks (Indefrey & Levelt, 2004) and initiating sentences describing simple actions around 1500 ms (Griffin & Bock, 2000), interlocutors must engage in parallel processing in order to produce the short turn transitions typical of conversation (Garrod & Pickering, 2015; Levinson, 2016): that is, next speakers must start to plan their turn while also processing the information from the turn that is still underway. By now, there is plenty of experimental evidence corroborating this assumption of parallel planning and comprehending (Barthel & Levinson, 2020; Barthel, Meyer, & Levinson, 2017; Barthel, Sauppe, Levinson, & Meyer, 2016; Bögels, 2020; Bögels, Casillas, & Levinson, 2018; Bögels, Magyari, & Levinson, 2015; Boiteau, Malone, Peters, & Almor, 2013; Corps, Crossley, Gambi, & Pickering, 2018; Magyari, de Ruiter, & Levinson, 2017; Sjerps & Meyer, 2015). Interlocutors are also able to plan (and in case planning finishes early are able to buffer) their planned next contribution before articulating it (Barthel et al., 2017; Bögels et al., 2015; Corps, Gambi, & Pickering, 2018), lending feasibility to the idea that next speakers start planning their turns even quite some time before uttering it.

One intriguing question is how people know when to launch their planned utterances as a next turn. The oscillator model of turn-taking timing (Wilson & Wilson, 2005) claims that interlocutors operate like oscillators coupled in antiphase. This means that the rhythmic structure of an on-going speaking turn (syllable rate) creates a cyclic pattern of

* Corresponding author at: Maria Montessori building, Thomas van Aquinostraat 4, 6525 GD, Donders Institute for Brain, Cognition, & Behaviour, Radboud University, Nijmegen, The Netherlands.
E-mail address: Judith.Holler@mpi.nl (J. Holler).

https://doi.org/10.1016/j.cognition.2024.105806
Received 27 April 2022; Received in revised form 4 March 2024; Accepted 24 April 2024
Available online 14 May 2024

possible turn transitions, with next speakers being maximally prepared to begin speaking when current speakers are least likely to produce their next syllable, thus reducing the likelihood of simultaneous turn beginnings. But the process requires more than this. Next speakers need to pinpoint a *particular* possible turn transition for choosing to launch their next turn. It has been suggested that current speakers make themselves predictable in a number of ways which allow a next speaker to foreshadow when a current turn comes to an end. In part, this is based on turn content and syntactic structure being projectible to a certain extent (Sacks et al., 1974). Being able to project turn content and syntactic structure also makes it possible to project, roughly, when it will end (de Ruiter, Mitterer, & Enfield, 2006; Magyari & de Ruiter, 2012; Riest, Jorschick, & de Ruiter, 2015). In addition, turn ends are marked by a wide range of prosodic cues (Barthel et al., 2016; Beattie, Cutler, & Pearson, 1982; Bögels & Torreira, 2015, 2021; Casillas & Frank, 2017; Lammertink, Casillas, Benders, Post, & Fikkert, 2015; Local, Kelly, & Wells, 1986; Turk & Shattuck-Hufnagel, 2014). The interplay of parallel processing, early next turn planning and the role of turn-final go signals are captured in a recent turn-taking model by Levinson & Torreira (2015).

Together, the above studies convincingly demonstrate that the vocal modality contains a significant number of cues that influence turn timing in conversation and the cognitive processes that underpin it. However, the modern human communication system has emerged as a product of face-to-face interaction. As part of this process, language has evolved into a multimodal phenomenon, where verbal utterances are embedded in a rich visual display of communicative signals (Goldin-Meadow, 2017; Kendon, 2017; Levinson & Holler, 2014; McNeill, 2012; Vigliocco, Perniss, & Vinson, 2014). However, most models of conversational turn-taking focus on the vocal and verbal sources of information alone (Garrod & Pickering, 2015; Levinson, 2016; Levinson & Torreira, 2015; Sacks et al., 1974; Wilson & Wilson, 2005). Duncan's work (Duncan, 1972, 1974; Duncan & Niederehe, 1974) marks an exception and makes a strong case for visual bodily signals feeding into the process of turn coordination. The claim is that in addition to prosodic signals (intonation contour, drawl, pitch, loudness), syntax, and certain stereotyped expressions commonly populating turn ends (e.g. 'you know', 'but uh', 'or something'), the termination of manual gestures (or even relaxation of hand poses) acts as a distinctive turn-yielding signal. Moreover, Duncan claims that these different signals act in a probabilistic fashion, with turn transition becoming more likely the more signals are displayed. Furthermore, engagement in manual gesturing apparently trumps all of these signals by acting as a clear attempt-suppressing signal. This prevents potential next speakers from trying to take the turn, no matter how many turn-yielding signals are displayed (critically, this does not include movement involved in retracting the hand once a gesture has been performed). Nevertheless, on the whole, the causal role of the body in signalling upcoming turn ends is not well understood and even less so the cognitive processes underpinning multimodal turn-taking, due to a paucity of experimental work in this domain. Moreover, we have not even begun to investigate which specific visual signals might underlie this causal role of the body in signalling and anticipating upcoming turn ends. Knowing which bodily signals are most implicated in this process is crucial for developing multimodal theories and models of turn end anticipation and conversational turn-taking more broadly.

Importantly, the general contribution of semantic, pragmatic and interactional information to speaker's utterances through visual bodily signals is well-evidenced (e.g., Bavelas, 2022; Kendon, 1967, 2004; McNeill, 1992; Nota, Trujillo, & Holler, 2023; Özyürek, 2014), thus providing the basis for assuming that they may shape recipients' anticipatory cognitive processes during turn coordination, too (such as by shaping predictions of upcoming turn content or foreshadowing upcoming turn completion). A small number of studies has begun to study the role of visual bodily signals in turn-timing experimentally, with mixed results. Barkhuysen et al. (Study 2, 2008) and Latif, Alsius, and Munhall (2017) asked participants to judge whether, at the end of a current turn, a turn transition would follow or not, using turns with and without visual bodily signals. Both studies found an advantage of the presence of visual signals in turn transition judgement tasks (and one study using similar list-based stimuli to Barkhuysen et al. even found high accuracy based on visual signals alone, see Bi & Swerts, 2017). However, a study by Mixdorff, Hönemann, Kim, and Davis (2015) failed to find a significant benefit of audio-visual over audio-only stimuli in turn transition judgement tasks, at least when language proficiency was high. All of these studies focused on the accuracy of judgements about the perceived likelihood of upcoming turn transitions, but not on *when* turns would end. Some studies have addressed the latter issue, but also with mixed results. In their first experiment, Barkhuysen, Krahmer, and Swerts (2008) used stimulus turns which consisted of word lists and asked participants to detect the end of the turns in vision-only, audio-only or audio-visual conditions. Their results showed no improvement in reaction times when seeing the speaker in addition to hearing the verbal content contained in the stimulus turns. Latif, Alsius, and Munhall (2018) used stimulus turns based on natural conversations combined with a task that required participants to press a button when the turns would end. Similar to Barkhuysen and colleagues, they found no improvement in synchronization of button presses and actual turn ends in the audio-visual over the audio-only condition. However, a recent study based on stimulus turns from Dutch sign language showed that even non-signers are able to anticipate upcoming turn ends in a way comparable to native signers when seeing the signed turns (de Vos, Casillas, Uittenbogert, Crasborn, & Levinson, 2021). This suggests that the visual modality appears to contain some turn end projecting signals that are 'globally accessible' and not specific to the linguistic content. This raises the possibility that similar features may be present in visual bodily signals accompanying spoken conversation. Moreover, previous studies had some limitations which may explain why they found no advantage of visual signals, such as the use of word lists rather than spontaneous conversation extracts (Barkhuysen et al., 2008) or showing the interactants together (putting the participant in an overhearer position) and from the side (thus possibly obscuring visual signals, see below) (Latif et al., 2018). In short, further research is needed to shed light on the extent to which the presence of visual bodily signals may affect end anticipation of speaking turns.

If the body plays a causal role in signalling upcoming turn ends, an open question is which visual signals are responsible for this effect. Knowing this can constrain our interpretations of the underlying mechanisms by which seeing the speaker might help interlocutors anticipate when their current turn is coming to an end. Three visual signals are especially interesting to consider in this context: manual gestures, gaze and head movements. Certain manual gestures can function in a turn-yielding fashion, namely by 'handing over' the turn (Bavelas, Chovil, Coates, & Roe, 1995; Kendon, 2004; Streeck & Hartge, 1992). For other gestures, instead their *termination* may be a turn-yielding signal (in line with Duncan; Duncan, 1972, 1974; Duncan & Niederehe, 1974). There is experimental evidence supporting the idea that the termination of gesture makes a turn transition more likely, at least if gestures occur just prior to turn boundaries (Zellers, House, & Alexanderson, 2016). Further, research has shown that turns accompanied by gestures are responded to faster than those not featuring a gestural component (Holler, Kendrick, & Levinson, 2018; Kendrick, Holler, & Levinson, 2023); at least a subset of these fast responses followed gestures that retracted before the end of the speaking turn, thus acting as an early turn completion signal. Trujillo, Levinson, & Holler (2021) manipulated the degree of visibility in free conversation and found that manual gestures are associated with smoother turn timing (less overlap and smaller gaps). However, due to the free conversation paradigm, many variables may have correlated with the presence of gestures, such as semantic speech content, specific syntactic structures, prosodic patterns and so forth, thus making it impossible to draw conclusions about the causal effect of gestures on turn-timing. In terms of other visual bodily signals, gaze (often in conjunction with head

direction) has been much discussed in terms of turn-taking, with some ascribing it an important turn end signalling function (Bavelas, Coates, & Johnson, 2002; Degutyte & Astell, 2021; Ho, Foulsham, & Kingstone, 2015; Kendon, 1967) (but see Duncan, 1972, 1974; Rossano, 2012; Streeck, 2014). An empirical study using virtual avatars showed that overall, the presence, direction and timing of gaze shifts did not impact turn end anticipation (Gambi, Jachmann, & Staudte, 2015), but to what extent these results generalise to human-human interaction is not clear. Finally, head movements may function as a turn-holding or turn-yielding signal as they are used increasingly in turn transition contexts involving overlap (Danner, Krivokapić, & Byrd, 2021), and predictive models use head movements to predict turn ends (De Kok & Heylen, 2009). However, whether recipients use these head movements to anticipate turn ends is currently unknown. Thus, amongst the myriad of visual signals, manual gestures, gaze and head movements have the potential to contribute to the coordination of speaking turns. However, the causal role of the body in signalling upcoming turn ends is not well understood, and it is unknown which specific visual signals do play a role. The two present studies fill this gap.

2. Present study

The conflicting results on the role of visual bodily signals in the context of turn timing, and the lack of research into which visual signals specifically may be important, call for further empirical investigations. The present study aims to address this issue with a turn end anticipation paradigm (response times) in combination with turn stimuli that were based on unscripted, casual conversations between friends. Moreover, it builds on previous studies in this domain by using a novel set-up that allowed for the generation of video stimuli featuring all of the visual bodily signals from the hands, head, face and eyes that were available to interlocutors in the original conversational interaction. Thus, the studies contrasts with those that showed recordings of two participants engaged in a conversation from a lateral perspective (de Vos et al., 2021; Hirvenkari et al., 2013; Latif et al., 2018; Preisig et al., 2016), which can obscure part of the face, as well as gaze movements when they are not discernable from head direction. Considering that especially gaze is often deemed an important turn-taking signal (Kendon, 1967), the visual unavailability of such signals may be significant. Moreover, the present paradigm allows participants to see current speakers from a frontal view, just as the respective interlocutor at the time did. This evokes a second person rather than a third person perspective, which can have a marked effect on cognitive and neural processes, amongst others due to greater involvement of the 'mentalizing network' (Redcay & Schilbach, 2019). Although the present study did not involve fully reciprocal social interaction, the second person perspective might help to simulate judgements about when a turn comes to an end since participants might feel somewhat more involved in the conversation, at least more so than when making those judgements from a third person observer perspective. The present studies test the hypothesis that, under these presentation conditions and in the presence of naturalistic conversational audio-visual turn stimuli, we may see an effect of visual signals on turn end anticipation after all.

Experiment 1 combines the turn end anticipation paradigm with a manipulation of conversational context, where turns are either shown in their chronological order (with previous turns thus providing context for a current turn) or in random order. A previous study has investigated the role of intonational cues in spoken turn end projection combined with a manipulation of available preceding discourse (10–20s versus none) (Bögels & Torreira, 2021). This study found only a trend of an effect for very short turns, a hampering effect of discourse context in the case of longer turns, and no effect at all when considering a wider response window (of 500 ms). Two other studies manipulated whether a spoken one-sentence context constrained the predictability of the to-be-anticipated turn and found that a constraining context made turn end anticipation earlier but not more precise than an unconstraining context (Corps, Crossley, et al., 2018; Corps, Pickering, & Gambi, 2019). The present study tests whether conversational context improves participants' ability to anticipate turn ends in conjunction with stimuli from face-to-face dialogue where interlocutors may be relying on the various information sources they have at their disposal in ways different from telephone conversations or scripted discourse context.

The presence of conversational context may also modulate the influence of visual signals, depending on the level at which these are functioning: if visual bodily signals do have an effect on turn end anticipation and this effect stems primarily from an improved prediction of the semantic content (at the lexical or message level), then conversational context may reduce the potential benefit of visual signals (since preceding context itself may improve content prediction). However, if visual bodily signals act as part of a portfolio of turn final cues (Levinson & Torreira, 2015) and exert their influence primarily 'locally' (i.e. within the boundaries of that turn and independently of any information prior to it), then preceding conversational context may not significantly attenuate the benefit of visual bodily signals in the process of turn end anticipation.

Experiment 2 uses the same turn end anticipation paradigm to zoom into the role that specific visual articulators play in signalling upcoming turn ends. Rather than an overall manipulation of speaker visibility (audiovisual vs. audio-only) as in Experiment 1, Experiment 2 manipulates the visibility of specific visual articulators. Based on the literature, we investigated the role of visibility of the eyes (Gambi et al., 2015; Kendon, 1967), the head (Danner et al., 2021) and the upper body (i.e. torso and arms, including manual gestures; Bavelas et al., 1995; Duncan, 1972; Duncan & Niederehe, 1974; Kendon, 2004). By manipulating their visibility, we can begin to investigate the causal role that each of these articulators has in turn end anticipation. While these are arguably still coarse categories, with e.g. the head including many different visual signals such as facial expressions, visible speech, and head nods, this provides an important first step into understanding the impact of different visual articulators in coordinating turn-taking behaviour.

Finally, both experiments take into consideration turn duration as an important variable influencing response times. This is based on past studies, some of which have shown that longer turns tend to lead to shorter response times (Corps et al., 2019; Corps, Crossley, et al., 2018; de Ruiter et al., 2006; Gambi et al., 2015), and others which have shown that both very short and long turns lead to longer response times (Roberts, Torreira, & Levinson, 2015). Research on general relationships between stimulus onset and reaction times outside of the domain of turn end anticipation has long suggested that this relationship is complex, with the variation of stimulus duration and participants' expectations playing an important role (Näätänen, 1970). These findings alone mandate a consideration of turn duration in the statistical models we apply to the analyses of the response time data. Another reason is that differences in turn duration may interact with the presence of visual bodily signal. This is because longer turns often involve more than one point of possible completion. These are points of syntactical, semantic, or pragmatic completion which may make turn transition to a next speaker relevant (Ford & Thompson, 1996; Sacks et al., 1974). Thus, visual bodily signals may influence turn end anticipation particularly for longer turns by preventing participants from being 'gardenpathed' by such early points of possible completion which were not the actual end of the turn. A similar effect was observed by de Ruiter et al. (2006) when comparing turn anticipation for turns that were stripped of their intonational contours to turns where the intonation contour was preserved. Another possibility is that longer turns are accompanied by more visual signals, thus exerting a stronger influence in terms of facilitating the processing of semantic or pragmatic information. Whichever of these two explanations may hold, we would expect to observe an interaction between turn duration and speaker visibility. Furthermore, conversational context may also prevent participants from being 'gardenpathed' by early points of possible completion especially for longer turns. Hence we may expect a stronger effect of conversational context for longer

turns (i.e. an interaction between conversational context and turn duration). And, of course, the presence of visual signals, conversational context and turn duration may interact in that conversational context may weaken the effect of visual bodily signals being present, especially so for longer turns.

3. Experiment 1

In Experiment 1, we investigated two variables inherent to face-to-face conversation, the presence of visual bodily signals and preceding discourse context in terms of their contribution to turn end anticipation. Participants were presented with turns from naturalistic conversations, with the task to anticipate the moment at which they thought the speaker would be finished speaking and to press the button as close as possible to this moment. We manipulated speaker visibility (audio-only vs. audiovisual) and conversational context (turns in random vs. chronological order) to investigate effects on turn end anticipation.

3.1. Methods

The analysis script can be found on OSF: https://osf.io/4gtw5/.

3.1.1. Participants

Thirty-two native speakers of Dutch (23 female) participated in the experiment. Participants' age ranged from 18 to 48 years (M = 24.4). Four additional participants were tested but not included due to incomplete data resulting from technical malfunction/experimenter error. Participants were recruited via the participant database of the Max Planck Institute for Psycholinguistics in Nijmegen and were paid for their participation. The study was approved by the Social Sciences Faculty Ethics Committee of the Radboud University, Nijmegen.

3.1.2. Materials

The stimulus materials were taken from recordings of natural dyadic conversation between friends. For these recordings, participants were recruited via the participant database of the Max Planck Institute for Psycholinguistics in Nijmegen and were paid for their participation. All participants were native Dutch speakers. The study was approved by the Social Sciences Faculty Ethics Committee of the Radboud University Nijmegen. Informed consent was obtained before and after the recordings, including consent for using the recordings as stimuli materials in future studies.

The speakers were instructed to converse freely about whatever they wanted, but not about the experiment. The audio and video of these conversations were recorded. To obtain clean audio tracks from the individual speakers, the recordings were made using an apparatus similar to a video-call set-up (such as with Zoom). The speakers sat in different rooms in front of a computer screen on which they saw the other person. Audio and video were recorded using sensitive microphones clipped to the bottom of the screen and a high-definition camera placed inside a box which projected the frontal recording of the participant onto the other the computer screen in the other participant's room. Critically, the recording apparatus involved a mirror construction which meant participants did not see the camera recording them but the projection of the other participant instead. Participants could hear each other via earphones. The audio that was recorded in this way always contained only the speech from one speaker, even if the speakers were talking simultaneously. Each dyad was recorded for 40 min.

The turns used as stimuli were extracted from four of eleven recorded dyads to be able to show longer stretches from the selected conversations (thus constituting the variable 'conversational context'). The other seven conversations were not used because they either included talk about the recording set-up itself, featured the visitor badge of the speakers or other distracting visual features, or ended early due to a technical error. Also, the selected conversations included enough turn transition stretches not containing any information deemed too personal or person-identifying in terms of the content of talk. Turns from all eight speakers in the four dyads (all female-female) were selected. First, for each dyad one excerpt of a few minutes (4m33s; 3 m43; 4m15s; 3m45s) was chosen that was balanced in terms of speaking time across interlocutors, to maximize the number of turn transitions. Turns were extracted from the excerpts, and were used as stimuli. Turns were excluded only if they occurred in complete overlap with the other speaker's turn and thus may not have been heard and was not overtly responded to by the other (n = 7; for example, in one case A said "Maar goed, ga je zelf ontdekken" ["But okay, that's something you'll find out yourself"] in complete overlap with B saying "Okay nou ik ben benieuwd" ["Okay well I am curious"]). Turns shorter than 5 words were not used in the analysis (following de Ruiter et al., 2006; n = 24) but were presented in the experiment to retain the flow and cohesion of the conversation (however, due to their high frequency, continuers [e.g. "yeah", "mhm" (Schegloff, 1982)] were not coded as turns and were thus not included, nor were non-verbal sounds such as laughter, coughs or sighs). Retaining the flow of the conversation was important for our manipulation of conversational context (see below). Ultimately, 157 turns were presented in the experiment (per dyad: 51, 21, 35, 50), of which 133 were analysed (per dyad: 42, 19, 32, 40).

The onset, offset and duration of each turn were annotated in ELAN 4.9.3 (Lausberg & Sloetjes, 2009) based on the acoustic signal, to create the clips. In principle, a turn was presented from speech onset to offset. However, if this meant that a visual signal was partly cut off, the duration of the clip was lengthened to include this visual signal. For more precise analyses, we later measured the offset of each acoustic turn with millisecond precision in Praat (Boersma, 2001) and used this offset to calculate the response times. This meant that sixty-seven video clips were somewhat longer than the acoustic turn (M = 101 ms), some due to lengthened clips to include the visual signals, and some due to more precise acoustic measurements made in Praat. Importantly, the offset measure used in our analyses was always based on the end of the acoustic turn measured in Praat, not the end of the video clip. The resulting turn durations varied across items, ranging from 748 ms to 36,140 ms (M = 6525, SD = 7255; Fig. 1). Moreover, the clips never contained speech from the next speaker, due to our video-call set-up described above.

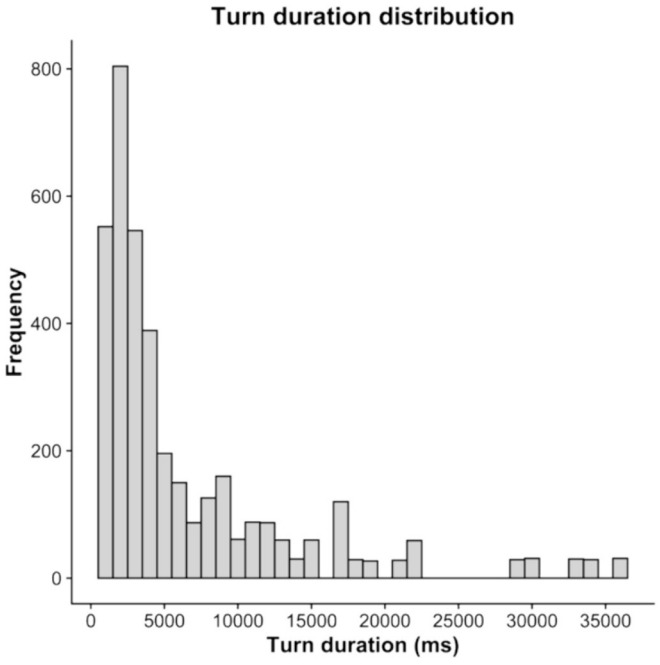

Fig. 1. Histogram indicating the distribution of turn durations. Each bin represents 1000 ms.

Next, for each turn the following turn transition type as it occurred in the conversation was coded. The rationale was that the design of turns produced in overlap or with a gap may be different and that such potential differences may be something that participants pick up on when making their judgements. We therefore added 'transition type' as a control variable. Turns were thus coded as gap turns, if they were followed by a turn transition gap of longer than 200 ms ($n = 53$), as smooth transition turns, if they were followed by a gap shorter than 200 ms ($n = 19$), and as overlap turns, if the onset of the following turn preceded the offset of the turn in question ($n = 50$). Finally, turns were coded as being associated with 'no transition' ($n = 11$) if they were the last turn of the conversation excerpt or if that turn was followed by another turn from the same speaker.

3.1.3. Design

We manipulated speaker visibility (audio-only vs. audiovisual) and conversational context (random vs. chronological order) to investigate effects on turn end anticipation. By crossing visibility and context, four conditions were created (audio-only stimuli in random order, audio-only stimuli in chronological order, audiovisual stimuli in random order, audiovisual stimuli in chronological order). For the audio-only items, participants were shown a black screen with a white fixation circle in the middle. For the audiovisual items, participants were shown the speaker from around 30 cm above the head to the knees in a frontal perspective (Fig. 2). In the chronological condition, turns were shown in the original order of the conversation. In the random condition, turns were shown in a random order, with the exception that turns were not immediately followed (or preceded) by the turns that immediately followed (or preceded) them in the conversation.

Participants were presented with four blocks of stimuli, with a different condition in each block. In each block, they would see turns from one dyad. Conditions and dyads were randomized according to a Latin square design (Saville & Wood, 1991). The two audio-only blocks and the two audiovisual blocks were always presented after each other, e.g., audio-audio-audiovisual-audiovisual or audiovisual-audiovisual-audio-audio, with the conversational context variables counterbalanced within the audio-only and audiovisual blocks.

3.1.4. Procedure

Participants were seated in front of a computer (24-in. screen) and received written instructions, explaining that short fragments from conversations would be presented in the four different conditions. Participants were asked to ignore the differences between these conditions as much as possible and to focus on their task: anticipating the moment at which they thought the speaker would be finished speaking and to press the button as close as possible to this moment. Participants were explicitly told not to wait for the end of the fragment and then press the button. These instructions were based on an earlier turn anticipation study (de Ruiter et al., 2006). In addition, participants were instructed to look at the computer screen, during both the audio-only and audiovisual blocks.

Each trial started with a visual countdown from three to one, with each number appearing for 1000 ms (see Fig. 2). Next, a black screen was shown for 200 ms. Then, in the audio-only conditions, a white circle was presented, while the audio fragment played. In the audiovisual conditions, video and audio of the speaker were presented. Participants heard or saw each fragment only once. When participants pressed the button, the audio or video would stop immediately, to prevent giving feedback about the turn end anticipation accuracy. If the button was not pressed before the turn end, a black screen appeared. Button press time relative to stimulus onset was recorded by the computer. The next trial would appear only after a (further) button press, followed by a 1000 ms black screen. After six practice trials, the experimental trials were presented in four blocks. In between blocks, participants took breaks which terminated at points of their own choice (button press).

3.1.5. Analysis

We tested whether turn end anticipation was affected by visibility, context, transition type and turn duration. Due to a mistake in the stimulus lists, four participants accidentally saw the same item twice. Responses to these repetitions were excluded from analysis. In addition, responses to two items were excluded (from all conditions). Due to experimenter error, in these videos the listener was shown instead of the speaker. Furthermore, following de Ruiter et al. (2006), we excluded 379 responses that occurred >2000 ms after the turn end (9.0% of the data). These excluded responses were distributed similarly across the four conditions (audio-random: 89, audio-chronological: 96, audiovisual-random: 93, audiovisual-chronological: 101).

We fitted linear mixed effects models using the lme4 package (version 1.1.26; Bates, Mächler, Bolker, & Walker, 2015) in R (version 3.5.3.; The R Core Team, 2018). *p*-Values were obtained with the package lmerTest (version 3.0.1; Kuznetsova, Brockhoff, & Christensen, 2017). An lmer model was used to test the effects of Visibility (Audio, Audiovisual), Context (Random order, Chronological order), Transition Type (Gap, Overlap, Smooth, No Transition) and Turn Duration (continuous) on the dependent variable (Response Time). The dependent variable Response Time was created by subtracting the turn duration from the reaction time, such that negative response time values indicated that the participant pressed the button before turn end, while positive response time values indicated that the participant pressed the button after turn end. Please note that the turn duration sometimes differed slightly from the clip length, as clips were lengthened if that prevented a visual signal from being partly cut off (but response times were still calculated from speech offset, see Materials). The continuous variables Turn Duration and Response Time were z-scaled. The factors Visibility and Context were sum-to-zero contrast coded (Visibility: audio-only −1, audiovisual +1; Context: random order −1, chronological order +1). The factor Transition Type was also sum-to-zero contrast coded, where turns followed by smooth transitions were the reference level (−1) and other types (gap turns, overlap turns, no transition turns) were each compared to the reference level (all +1).

We started off with an intercept-only model and added possible

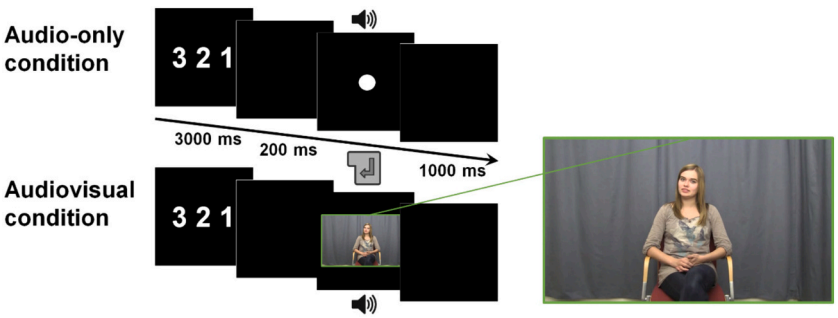

Fig. 2. Trial timeline.

factors and interactions in a stepwise manner. Factors or interactions were added to the model only if they significantly improved the model fit, as tested using the R function anova (Crawley, 2007). This way, the most parsimonious model was obtained. Factors and interactions were added in the following order: Visibility, Context, Visibility*Context, Transition Type, Visibility*Transition type, Turn Duration, Visibility*Turn Duration, Context*Turn Duration, Visibility*Context*Turn Duration. Finally, random effects were added to the model if they significantly improved the model. There were no convergence issues when modelling the predictors, but if a random slope resulted in convergence issues, it was removed from the model.

After the final model was obtained, we completed two analysis checks. First, the lmer function assumes that model residuals are normal. When this assumption was violated, we also ran the model using the function rlmer from the package robustlmm (Koller, 2016). This function also performs linear mixed effects analyses, but is more robust against deviations of residual normality. The second follow-up concerned the issue of pseudoreplication (Arnqvist, 2020). When the random effects structure is misspecified, *p*-values are inappropriately deflated. Because we could not always implement the full random effects structure, contrary to what is recommended (Arnqvist, 2020; Barr, Levy, Scheepers, & Tily, 2013), our analysis was at risk for deflated *p*-values. Following Arnqvist (2020), we therefore tested whether the results from the lmer could be replicated using simpler models of group means. For this, we averaged the (non-transformed) data points across participants, leaving each stimulus item with a single mean response time value (the average across 32 participants). Using the R function aov (Crawley, 2007), we ran an ANOVA that tested for the effects found with the lmer. The ANOVA was done by-items, rather than by-participants, because Turn Duration was a crucial variable in our model and when averaging across items, this variable is lost.

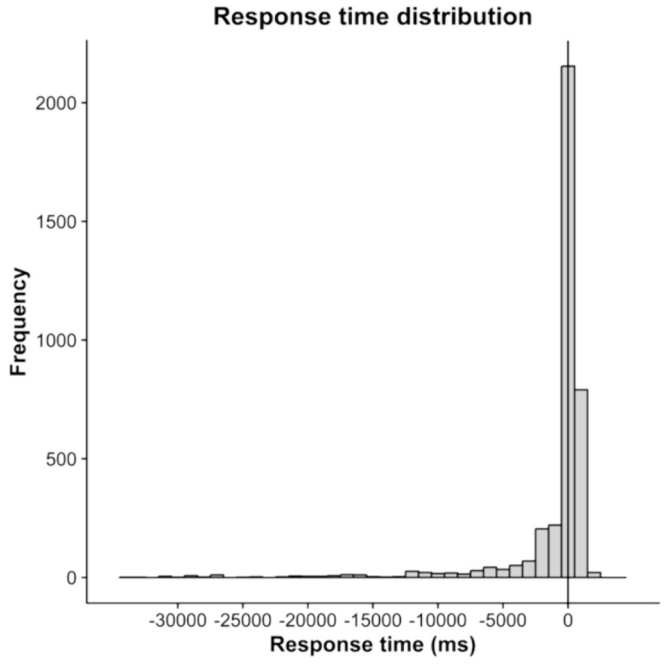

Fig. 3. Histogram indicating the distribution of button presses with respect to turn end (i.e. Response Time). The plot includes data from all the Visibility and Context conditions. Negative response time values indicate that the button was pressed before the turn end, positive response time values indicate the button was pressed after the turn end. Response time values above 2000 ms were considered outliers and removed. Each bin represents 1000 ms.

3.2. Results

On average, participants pressed the button 901.7 ms before the end of the turn (SD = 3965.1 ms) (see Fig. 3). Response times were most frequently on the order of 300 ms after turn end (and median = 322 ms).

We tested whether turn end anticipation was affected by visibility, context, transition type and turn duration. Using the model building procedure described above, we obtained the following model: Response Time ~ (Visibility * Turn Duration) + (Context * Turn Duration), with random intercepts for Item and Participant, as well as a random slope of Visibility by Participant (Table 1).

This model revealed a main effect of Turn Duration, an interaction between Turn Duration and Visibility, and an interaction between Turn Duration and Context. The main effect of Turn Duration indicated that for longer turns, button presses preceded the turn end more. This is not surprising, given that for longer turns there are more possible points of completion that could be misinterpreted as the turn end. The interaction between Turn Duration and Visibility indicated that Turn Duration had a smaller effect on Response Time when the speaker was visible, compared to when the speaker was not (Fig. 4). This means that when the speaker was visible, participants' responses were closer to the actual turn end, especially for longer turns (see Fig. A1 in the Appendix for visualisations of response precision). Similarly, the interaction between Turn Duration and Context indicated that Turn Duration had a smaller effect on Response Time when turns were presented in their chronological order, compared to random order (Fig. 4). Thus, when participants could use the conversational context, button presses were closer to the actual turn end, especially for longer turns (see Fig. A1 in the Appendix for visualisations of response precision). There were no other significant effects. Notably, during the model building procedure, we found no interactions between visibility and context, nor a three-way interaction between visibility, context and turn duration. Visibility and context were not significant as main effects (see Figs. A2 and A3 in the Appendix for visualisations). Moreover, there were no effects of transition type, showing that in our data, turn end anticipation was not affected by whether the turn resulted in overlap, a smooth transition or a gap in the original conversation.

To verify these model results, we ran a few follow-up analyses. First, we used the function rlmer, which is more robust against violations of residual normality, as well as against outliers (median was 322 ms versus a mean of −902 ms). The results described above were replicated (see Table A1 in the Appendix). Thus, the results were not driven by violation of the assumption of residual normality or outliers.

Second, we ran a by-items ANOVA to verify whether the results found above were perhaps driven by misspecified random effects. The ANOVA tested for the effects of Visibility, Context and their respective interactions with Turn Duration on mean Response Time. Visibility and Context were added as crossed random effects by item. The ANOVA revealed a significant main effect of Visibility, as well as an interaction between Visibility and Turn Duration, $F(1, 129) = 18.325$, $p < 0.001$. Moreover, there was a significant interaction between Context and Turn Duration, $F(1, 128) = 6.282$, $p = 0.01$. Visualisations of these interactions showed the same patterns as in Fig. 2. Thus, the results from the lmer were replicated. Overall, these results show that both visibility of the speaker and the conversational context of the turn lead to more accurate turn end anticipation, especially for longer turns.

3.3. Discussion

Experiment 1 set out to investigate whether the visual signals that accompany spoken turns in conversational face-to-face interaction help interlocutors anticipate the point when a current turn comes to an end. We also tested the effect of conversational context and the way in which it may allow us to better predict turn ends through the potentially constraining effect of preceding discourse on turn content prediction. Finally, we added turn duration as a predictor in our models since this is

Table 1
Details for the model predicting turn end anticipation (Response Time).

Fixed effects	β	*SE*	*df*	*t*	*p*
Intercept	−0.02	0.06	56.30	−0.28	0.78
$\text{Visibility}_{\text{audiovisual}}$	0.04	0.02	30.81	1.84	0.07
$\text{Context}_{\text{chronological}}$	0.02	0.01	3619.40	1.73	0.08
Turn duration	−0.49	0.03	130.79	−14.80	< 0.001***
$\text{Visibility}_{\text{audiovisual}}$ * Turn duration	0.06	0.01	3631.30	4.77	< 0.001***
$\text{Context}_{\text{chronological}}$ * Turn duration	0.04	0.01	3644.20	3.23	< 0.01**
Random effects		Var	SD		
Item	Intercept	0.12	0.35		
Participant	Intercept	0.08	0.28		
	$\text{Visibility}_{\text{audiovisual}}$	0.01	0.09		
Residual		0.55	0.74		

For the fixed effects, estimates (β), standard errors (SE), degrees of freedom (df), t-values (t) and *p*-values (*p*) are given. For the random effects, variance (var) and standard deviations (SD) are reported. The subscript 'audiovisual' indicates that these values are for the audiovisual condition with the audio-only condition as reference level (contrast coding: audiovisual 1, audio-only − 1). The subscript 'chronological' indicates that these values are for the chronological order condition with the random order condition as reference level (contrast coding: chronological 1, random − 1).
Formula in R: lmer (Response Time ~ Visibility * Turn duration + Context * Turn duration + (1 + Visibility | Participant) + (1 | Item)).

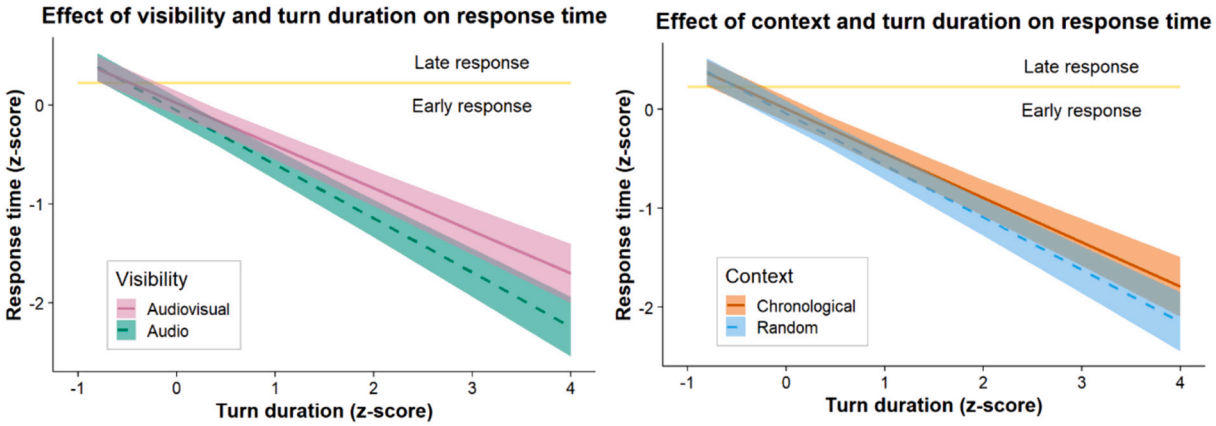

Fig. 4. Effects of visibility (left) and context (right) on turn end anticipation (response time). Lines indicate the linear model fit, the shaded areas around the lines represent their confidence intervals. The yellow line indicates the target button press exactly at the turn end (this line is the z-score [0.22] corresponding to the response time value of 0 ms). Areas separated by yellow lines indicate which responses were late (i.e. button press after turn end) or early (i.e. button press before turn end). *Left*: Turn end anticipation was more accurate for audiovisual than audio-only (i.e. closer to target button press), especially for longer turns. *Right*: Turn end anticipation was more accurate for turns presented in chronological than random order (i.e. closer to target button press), especially for longer turns. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

a factor that has been shown to influence response times (see Introduction).

Our results show that the manipulation of speaker visibility per se affected turn end anticipation to some extent (in terms of a strong trend in the mixed effects model, and a significant main effect in the anova), and significantly so when taking into account turn duration. This is in line with our expectations and means that visual signals may be safeguarding participants from responding to early points of possible completion (based on semantic, pragmatic and syntactic grounds) when current speakers did not intend them to be treated that way. We return to this finding in more detail in the General discussion.

Similarly to the results on the influence of visual bodily signals, our results showed that conversational context by itself only had some effect on turn end anticipation (again in terms of a trend), but in conjunction with turn duration the effect was significant. This means that participants benefited from the preceding conversation constraining the anticipation of turn ends, especially for longer turns. Current theories agree that predicting the content of incoming turns is key for achieving rapid turn-taking (Garrod & Pickering, 2015; Levinson & Torreira, 2015), but how exactly these content predictions aid successful turn-taking remains unclear. Corps, Crossley, et al. (2018) proposed two hypotheses: content predictability may facilitate turn-taking because listeners can prepare their response earlier when the turn's content is more predictable, or because listeners are more precise at predicting the speaker's turn end when the turn's content is more predictable. In line with the second hypothesis, turn end anticipation benefited from preceding context in our study, especially for longer turns.

This finding contrasts with previous studies using carefully constructed stimuli, showing that the presence of constraining context makes turn end anticipation responses earlier, but not more precise (Corps et al., 2019; Corps, Crossley, et al., 2018). Moreover, our results contrast with findings by Bögels and Torreira (2021) who used stimuli from natural conversations, as in the present study. They found no significant effect of discourse context, except that responses to longer turns were hampered. In order for discourse context from naturalistic conversations to have an effect (in combination with turn duration), a fair amount of context may have to be available: the amount of context included in their study (10–20 s) compared to the present study (up to between 20 turns (3m29s) to 50 turns (4m16s), depending on the conversation) differed considerably. In short, the present study is the first to demonstrate that preceding naturalistic conversational context can facilitate turn end anticipation, especially for longer turns, even when visual signals are present. Thus, preceding constraining context may aid successful turn-taking both by allowing for earlier response planning (Corps, Crossley, et al., 2018) and by improving turn end predictions. Models of turn end anticipation and language processing during

conversational turn-taking more broadly should thus include conversational context as a factor. Future research is necessary to specify when context improves turn end prediction and when it may not, depending on features of the context, the turn, and the listener.

Finally, the two variables that we manipulated in this experiment, visual bodily signals and conversational context, did not interact significantly, nor was the interaction significant when adding turn duration to the interaction. This suggests that preceding discourse does not diminish the influence of visual signals, thus underlining the independent and robust contribution visual bodily signals make to turn end anticipation. Importantly, it remains an open question which visual signals from the speaker's body exactly facilitated turn end anticipation in Experiment 1. Experiment 2 addresses this issue.

4. Experiment 2

Experiment 1 showed that seeing visual bodily signals of the speaker impacted turn end anticipation, possibly safeguarding participants from responding to early points of possible completion. However, it is unclear to what extent different visual signals (coming from the eyes, head, or upper body) played a role in this effect. Therefore, in Experiment 2, we created different versions of each clip, with either the eyes masked, the head masked, the upper body masked (i.e. torso + arms, everything below the neck) or with the full view of the speaker. We investigated whether masking of each of these visual articulators impacted turn end anticipation compared to seeing the full view of the speaker.

4.1. Methods

The preregistration of this experiment is available here: https://aspredicted.org/V3W_PNQ. The data and (power) analysis scripts can be found on OSF: https://osf.io/4gtw5/.

4.1.1. Participants

Sixty native speakers of Dutch (47 female) participated in the experiment. Participants' age ranged from 18 to 37 years (M = 23.0). One additional participant was tested but not included due to technical malfunction. Participants were recruited via the Radboud University Research Participation System and were paid or compensated with study credits for their participation. This study was approved by the Social Sciences Faculty Ethics Committee of the Radboud University, Nijmegen.

In our pre-registration, we stated that we would first test 60 participants and run our planned analyses. If there would be no significant effect of speaker visibility, we would then test another 20 participants. Although the effect of visibility was not apparent in the overall analysis, it was in the pre-registered follow-ups (for details, see Analysis). Therefore, we stopped testing at 60 participants.

4.1.2. Materials

We used the same stimulus materials as in Experiment 1, but we created four visibility conditions for the video with (1) full view of the speaker, (2) eyes masked, (3) head masked, and (4) upper body masked.

For masking, all stimuli were loaded into Adobe After Effects (version 2020; Christiansen, 2013). They were put in separate pre-compositions and contained an Adjustment layer that made up the mask. This layer is the same for all three blurred conditions – it contains a mask with a solid Fill effect added to it. This effect had a colour similar to the background (for the head and upper body condition) or the skin colour of the participants (for the eyes condition) to make the mask look less distracting. The mask in the head and upper body condition was static, but adjusted in scale to fit each speakers' head or body proportions. To reliably mask the eye region of each face, the eyes were motion tracked using the After Effects Tracker, and manually adjusted where necessary.

4.1.3. Design

We manipulated the level of speaker visibility to further investigate effects on turn end anticipation (see Fig. 5). The first condition remained the same as in Experiment 1, showing the speaker in full view. The second condition showed the speaker with the eyes masked. The third condition showed the speaker's upper body and hand movements, but the head was fully masked. The fourth condition is the opposite of the previous one with the head visible but the upper body completely masked.

Participants were presented with four blocks of stimuli, with a different visibility condition in each block. In each block, they would see turns from one dyad in chronological order. Conditions and dyads were randomized according to a Latin square design[1] (Saville & Wood, 1991).

4.1.4. Procedure

Participants were seated in front of a computer (24-in. screen) and received written and verbal instructions. These instructions were the same as for Experiment 1.

As in Experiment 1, each trial started with a visual countdown from three to one, with each number appearing for 1000 ms. Next, a black screen was shown for 200 ms. Then video and audio of the speaker were presented, in one of the four visibility conditions. Participants heard and saw the fragment only once. The procedure of the button press was identical to Experiment 1: after a press the video and audio stopped immediately. If the button was not pressed before the turn end, a black screen appeared and the next trial would appear only after a (further) button press, followed by a 1000 ms black screen. Button press time relative to stimulus onset time was recorded by the computer. Participants were given eight practice items first and then proceeded to the four blocks of experimental trials. In between blocks, there was a self-paced break and after the experiment, participants were given a few background questions.

4.1.5. Analysis

We tested whether turn end anticipation was affected by visibility and turn duration. Following de Ruiter et al. (2006), we excluded 234 responses that occurred >2000 ms after the turn end (2.9% of the data; Full body: 45, Head masked: 66, Eyes masked: 81, Upper body masked: 42).

We fitted linear mixed effects models using the lme4 package (version 1.1.30; Bates et al., 2015) in R (version 4.2.1; The R Core Team, 2018). p-Values were obtained with the package lmerTest (version 3.1.3; Kuznetsova et al., 2017). An lmer model was used to test the effects of Visibility (Full body, Head masked, Eyes masked, Upper body masked) and Turn Duration (continuous) on the dependent variable (Response Time). The dependent variable Response Time was created as in Experiment 1. The continuous variables Turn Duration and Response Time were z-scaled. The factor Visibility was sum-to-zero contrast coded, where each condition was compared to Full body (Full body: −1; Other condition: 1).

We used the following model: Response Time ~ (Visibility * Turn Duration). As in Experiment 1, we added random intercepts by item and subject. We added random slopes to the model if they significantly improved the model fit and did not result in convergence issues, resulting in a random slope of Visibility by Participant, exactly as in Experiment 1. After the final model was obtained, we completed the two analysis checks as in Experiment 1 (for details, see above).

Moreover, we also ran the above-described analyses using two subsets of the data. The reason for this is that for some of the clips the hands were still visible in the upper body masked condition. Sometimes, this happened for grooming movements (e.g., touching the hair; n = 8 items), but for other clips part of a meaningful gesture was visible when

[1] Due to experimenter error, two participants accidentally saw the same randomization.

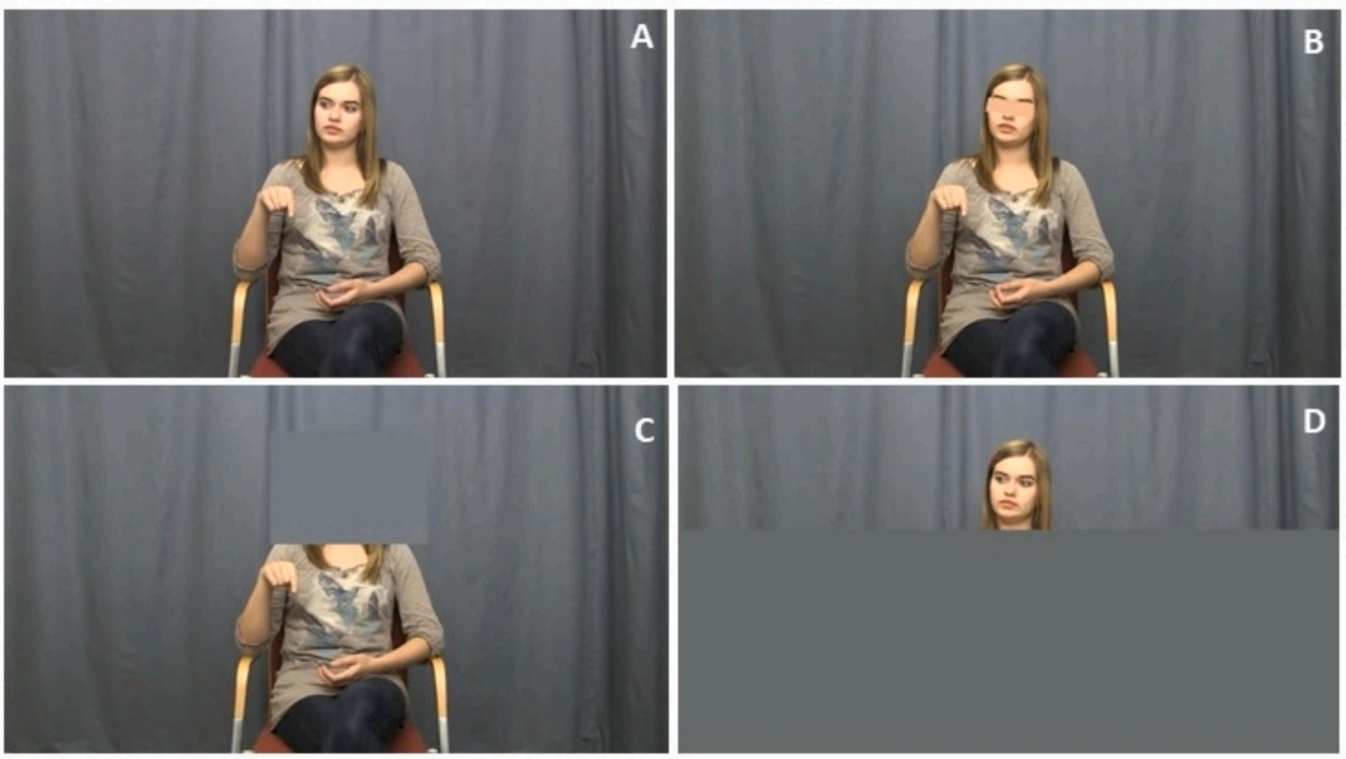

Fig. 5. The four visibility conditions: A) Full view, B) Eyes masked, C) Head masked, D) Upper body masked. For the trial timeline see Fig. 1.

it was performed at head height (n = 16 items). To make sure these instances did not impact the results, we reran the analyses on 1) the subset of the data with the visible gestures excluded (n = 117 items); 2) the subset of the data with all the visible hands excluded (n = 109 items).

4.2. Results

On average, participants pressed the button 1652 ms before the end of the turn (SD = 4822.8 ms) (see Fig. 6). Response times were most frequently on the order of 400 ms after turn end (and median = 257 ms).

We tested whether turn end anticipation was affected by visibility and turn duration (Table 2). The pre-registered model revealed only a main effect of Turn Duration, indicating that for longer turns, button presses preceded the turn end more (Fig. 7, left panel). Notably, there was no main effect of Visibility nor an interaction between Visibility and Turn Duration, showing that turn end anticipation was not affected by masking the speaker's eyes, head or upper body compared to seeing the full speaker.

Next, we performed the pre-registered follow-up analyses as done for Experiment 1. Running the analysis using the function rlmer did not replicate the results (instead there was a significant interaction between Turn duration and Visibility for the contrast between the Upper body masked condition and the Full view condition; see Table A2 in the Appendix). This suggests that perhaps the lmer results were affected by violation of the assumption of residual normality or outliers. Running the analysis with an ANOVA to verify whether the lmer results found above were perhaps driven by misspecified random effects (Arnqvist, 2020) replicated the results, with no significant interaction between Visibility and Turn Duration, $F(3, 393) = 1.722$, $p = 0.162$.

Finally, to make sure the clips with the hands still visible in the upper body masked condition did not impact the results, we reran the above-described analyses on 1) the subset of the data with the visible gestures excluded; 2) the subset of the data with all the visible hands excluded (as pre-registered). For both subsets, we found the expected

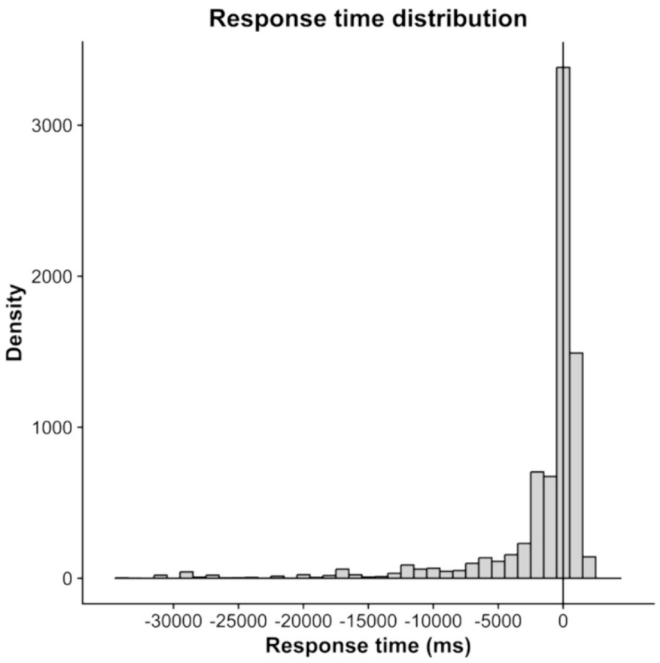

Fig. 6. Density plot indicating the distribution of button presses with respect to turn end (i.e. Response Time). The plot includes data from all the Visibility conditions. Negative response time values indicate that the button was pressed before the turn end, positive response time values indicate the button was pressed after the turn end. Response time values above 2000 ms were considered outliers and removed. Each bin represents 1000 ms.

Table 2
Details for the model predicting turn end anticipation (Response Time).

Fixed effects	β	*SE*	*df*	*t*	*p*
Intercept	0.01	0.05	183.08	0.16	0.88
$\text{Visibility}_{\text{eyes\_vs\_full}}$	−0.00	0.01	7583.34	−0.02	0.99
$\text{Visibility}_{\text{head\_vs\_full}}$	0.00	0.01	7584.36	0.22	0.83
$\text{Visibility}_{\text{upperbody\_vs\_full}}$	−0.02	0.01	7581.06	−1.59	0.11
Turn duration	−0.60	0.04	132.80	−15.34	<0.001***
$\text{Visibility}_{\text{eyes\_vs\_full}}$ * Turn duration	−0.00	0.01	7578.64	−0.26	0.79
$\text{Visibility}_{\text{head\_vs\_full}}$ * Turn duration	−0.01	0.01	7579.04	−0.50	0.62
$\text{Visibility}_{\text{upperbody\_vs\_full}}$ * Turn duration	−0.01	0.01	7580.36	−0.92	0.36
Random effects		**Var**	**SD**		
Item	Intercept	0.20	0.45		
Participant	Intercept	0.04	0.19		
Residual		0.39	0.63		

For the fixed effects, estimates (β), standard errors (SE), degrees of freedom (df), t-values (t) and *p*-values (*p*) are given. For the random effects, variance (var) and standard deviations (SD) are reported. The subscripts indicate the (sum-to-zero) contrasts, where each condition was compared to Full body (Full body: −1; Other condition: 1).
Formula in R: lmer (Response Time ~ Visibility * Turn duration + (1 | Participant) + (1 | Item)).

interaction between Visibility and Turn Duration, specifically for the contrast between the Upper body masked condition and the Full view condition (Fig. 7, right panel). This was the case for the subset without gestures (Table 3; $\beta = -0.04$, $SE = 0.01$, $t = -2.81$, $p < 0.01$), as well as the subset without hands visible in the Upper body masked condition at all ($\beta = -0.04$, $SE = 0.01$, $t = -3.06$, $p < 0.01$). These findings were replicated using the robust rlmer (see Tables A2 and A3 in the Appendix) and the ANOVA (without gestures: $F(3, 345) = 5.609$, $p < 0.001$; without hands: $F(3,321) = 4.084$, $p < 0.01$). There were no significant differences between the Eyes masked condition and the Full view condition, nor between the Head masked and the Full view condition. Thus, these results using a cleaner comparison between the Full body and Upper body masked condition suggest that seeing the upper body leads to more accurate turn end anticipation, especially for longer turns.

4.3. Discussion

Experiment 2 set out to investigate to what extent different visual signals (coming from the eyes, head, or upper body) help interlocutors anticipate the point when a current turn comes to an end. To do so, we used the turn-end anticipation paradigm and manipulated the visibility of the speaker's head, eyes and upper body (i.e. torso + arms). Our results show that the manipulation of upper body visibility affected turn end anticipation in interaction with turn duration. Crucially, this was only the case when in the clips with the upper body masked, hand gestures were not visible at all, also not at the height of the head. This finding suggests that especially signals from the upper body (i.e. hand gestures, since hardly any other movements occurred in this area) may be safeguarding participants from responding to early possible completion points.

Although Experiment 2 was not a direct replication of Experiment 1, two conditions in Experiment 2 together masked the full body (head masked and upper body masked, see Fig. 5). If neither of these masks would have had any impact on turn end anticipation, this would have created serious doubts about the replicability of the findings of Experiment 1. However, the visibility of the upper body was important for turn end anticipation, in line with the finding that speaker visibility overall (audiovisual vs. audio-only) benefits turn end anticipation. It suggests that the overall effect of speaker visibility may be due to the role of visual signals coming from the torso, arms or hands. We discuss this in more detail in the General discussion below.

5. General discussion

Precise coordination is at the very heart of turn-taking in interaction and necessary for achieving the minimal gaps and overlaps characteristic of human conversation (Sacks et al., 1974). Previous research has shown that interlocutors draw on a number of information sources to be able to make turn end predictions with as much precision as possible, including lexical information, syntactical information, and intonation (Barthel et al., 2017; Bögels & Torreira, 2015, 2021; Corps, Crossley, et al., 2018; Corps, Gambi, & Pickering, 2020; de Ruiter et al., 2006; Magyari et al., 2017; Magyari & de Ruiter, 2012; Riest et al., 2015;

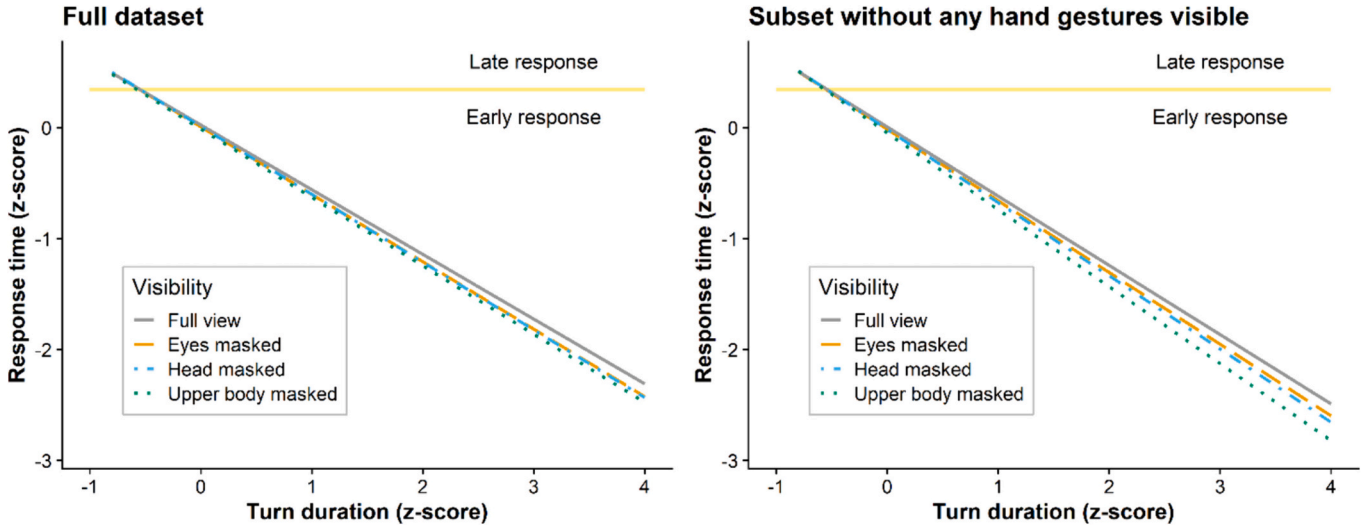

Fig. 7. Effects of visibility and turn duration on turn end anticipation (response time) for the full dataset ($n = 133$ items; left) and the subset of the data where in the upper body masked condition hand gestures were never visible ($n = 117$ items; right). Lines indicate the linear model fit. The yellow line indicates the target button press exactly at the turn end (this line is the z-score [0.34] corresponding to the response time value of 0 ms). Areas separated by yellow lines indicate which responses were late (i.e. button press after turn end) or early (i.e. button press before turn end). Turn end anticipation (response time) was more accurate (i.e. closer to target button press) when the turn duration was shorter. Moreover, for the subset (right panel), turn end anticipation was more accurate for the full view compared to when the upper body was masked, especially for longer turns. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Details for the model predicting turn end anticipation (Response Time) in the subset with no hand gestures visible in the Upper body block condition.

Fixed effects	β	*SE*	*df*	*t*	*p*
Intercept	−0.02	0.04	165.65	−0.43	0.67
$\text{Visibility}_{\text{eyes\_vs\_full}}$	0.00	0.01	6671.90	0.34	0.74
$\text{Visibility}_{\text{head\_vs\_full}}$	0.00	0.01	6673.10	0.17	0.87
$\text{Visibility}_{\text{upperbody\_vs\_full}}$	−0.03	0.01	6670.13	−2.60	<0.01**
Turn duration	−0.66	0.04	116.64	−15.87	<0.001***
$\text{Visibility}_{\text{eyes\_vs\_full}}$ * Turn duration	0.01	0.01	6653.19	0.74	0.46
$\text{Visibility}_{\text{head\_vs\_full}}$ * Turn duration	−0.00	0.01	6654.55	−0.31	0.75
$\text{Visibility}_{\text{upperbody\_vs\_full}}$ * Turn duration	−0.04	0.01	6655.58	−2.81	<0.01**
Random effects		**Var**	**SD**		
Item	Intercept	0.13	0.36		
Participant	Intercept	0.03	0.16		
Residual		0.27	0.52		

For the fixed effects, estimates (β), standard errors (SE), degrees of freedom (df), t-values (t) and *p*-values (*p*) are given. For the random effects, variance (var) and standard deviations (SD) are reported. The subscripts indicate the (sum-to-zero) contrasts, where each condition was compared to Full body (Full body: −1; Other condition: 1).
Formula in R: lmer (Response Time ~ Visibility * Turn duration + (1 | Participant) + (1 | Item)).

Roberts et al., 2015).

However, in face-to-face conversation, visual bodily signals are ubiquitous and provide semantic and pragmatic information interwoven with information coming from speech. It has been argued that these signals may be important devices for the coordination of conversational turns (Duncan, 1972, 1974; Duncan & Niederehe, 1974; Holler et al., 2018; Kendon, 1967; Kendrick et al., 2023; Mondada, 2007; Stivers et al., 2009; Streeck & Hartge, 1992; Zellers et al., 2016). Experimental studies have corroborated these observations by showing that in the presence of visual signals participants are significantly more accurate at judging whether turn transition will occur or whether the same speaker will continue to talk (Barkhuysen et al., 2008; Latif et al., 2017). However, the question as to whether visual bodily signals accompanying speaking turns are advantageous for participants in predicting when the end of a currently unfolding speaking turn will occur (rather than making judgements about their continuation after having seen the complete fragment) has hardly been addressed by extant studies. This is a judgement that is close to the process interactants engage in during conversation and thus merits investigation. A study by Latif et al. (2018) showed that turn end anticipation appears to be unaffected by the presence of visual bodily signals. Their study used stimuli that captured the dyadic interaction between interlocutors to test for signals produced by both speaker and addressee. However, the lateral view meant that, in making their judgements, participants may not have had the full set of signals at their disposal, since the interlocutors were orienting towards one another, thus making signals from the face potentially less accessible. The only other study on turn end anticipation with visual signals is a recent one that focused on signed rather than spoken conversational turns (de Vos et al., 2021). This study showed that some visual signals used by signers seem to be signals that can benefit turn-end anticipation judgements also in hearing individuals when seeing signed turns (thus pointing towards some globally accessible cues). This begs the question whether visual bodily signals used in spoken conversation also benefit turn end anticipation, and if so, which visual articulators may play a role in this process.

The present study set out to investigate just this. To do so, it used a novel apparatus providing participants with a frontal view of the current speaker, thus capturing all manual and head gestures as well as facial signals, including gaze movements. In Experiment 1, we tested whether seeing the speaker and knowing the conversational context helps interlocutors anticipate the point when a current turn comes to an end. In Experiment 2, we zoomed in on the effect of speaker visibility and investigated the role that various visual articulators (eyes, head, upper body/hand gestures) play in turn end anticipation.

Our results show that the manipulation of speaker visibility affected turn end anticipation especially for longer turns. This means that visual signals may prevent listeners from being 'gardenpathed' by early points of possible completions which do not represent the actual end of the turn. Visual signals had this effect even when conversational context was present, underlining the independent and robust contribution visual bodily signals make to turn end anticipation. Importantly, visual signals from the upper body appear to be driving this effect, most likely manual gestures, since hardly any other movements in this area occurred. Torso movements that did occur are an unlikely candidate, because they almost always co-occurred with movement of the head (e.g., leaning the torso forward or backward also involves moving the head) and masking the head had no impact on turn end anticipation. Our findings pointing towards a role of manual gestures in turn end anticipation is in line with an earlier study by Trujillo et al. (2021) which manipulated visibility in free conversation. However, the visibility manipulation in this earlier study did not allow for a clear mapping onto the visibility of different bodily articulators, and due to the free conversational paradigm, only correlations between turn-timing and manual gestures could be measured. By using controlled stimuli, the present study significantly builds on this work by suggesting a *causal* effect of manual gestures on turn end judgements.

Manual gestures may facilitate turn end anticipation through a variety of means. For instance, a gesture that has just been launched or has unfolded only halfway may indicate that the current speaker is far from 'being done', with more talk to come (Duncan, 1972; Duncan & Niederehe, 1974). Thus, the kinematic profile of gestures and their specific movement phases may override the verbal or vocal cues that may point towards possible completion. In fact, a recent corpus study showed that turn transition was less likely when points of possible completion in the verbal utterance coincided with gestures during their preparation or stroke phase, i.e. the most meaning bearing parts of the gestural movement (Kendrick et al., 2023). Alternatively (or even additionally), due to their typical early timing, gestural signals may depict information which foreshadows that corresponding semantic or pragmatic information in the speech stream is still to come (Ferré, 2010; Holler & Levinson, 2019; ter Bekke, Drijvers, & Holler, 2024) and thus that the end has not yet been reached. Such visual information may even constrain the predictions participants made about the content, possibly even including prediction of the number of lexical items (Magyari & de Ruiter, 2012), thus making projection of the turn end more precise. And similar effects may operate at the pragmatic level as visual signals may facilitate comprehension, and perhaps even prediction, of the speech act that a current turn performs (Holler & Levinson, 2019). In a paradigm using verbal turns only, Corps, Crossley, et al. (2018) found no effects of semantic content prediction on turn end anticipation, which may be due to a number of methodological reasons as they point out. However, it may mean that also in the present study an influence of visual signals on semantic or pragmatic content prediction may not be the most likely explanation for the turn end anticipation effect we found. This would be in line with our finding that preceding conversational context (which may itself improve content predictions) did not attenuate the benefit of visual signals. More likely may be the first possibility presented, i.e., that visual bodily signals guide participants in considering only those possible completion points as relevant for transition that current speakers intend them to, which may involve overriding vocal or verbal possible completion signals.

Although null effects should always be interpreted with caution, an interesting finding was that masking the head did not impact turn end anticipation. We masked the head because research has shown that speakers who are approaching an overlapping turn exchange show an increase in their head movements, which could function as a turn-

holding or turn-yielding signal (Danner et al., 2021). However, our head mask also rendered visible speech movements from the mouth invisible. The fact that this mask had no impact suggests that the benefits of seeing the speaker for turn end anticipation were not due to improved speech perception (e.g., Bernstein, Auer, & Takayanagi, 2004; Grant & Seitz, 2000; McGurk & MacDonald, 1976). Therefore, the mechanisms underlying how visual signals improve turn end anticipation may differ from how visual signals improve other aspects of language processing. Future research may be able to throw more light on the precise mechanisms that underpin the contribution of different visual bodily signals to turn end anticipation, especially studies using online measurements (e.g., EEG or eye tracking) without explicit judgement tasks and based on more interactional paradigms suitable for dipping into the cognitive processes during turn-taking.

The present findings inform us about the influence of speaker visibility on how precisely people can anticipate an upcoming turn end. What they do not tell us is how they may influence interlocutors' turn-taking behaviour. For example, Holler et al. (2018) found that in casual conversation, the presence of manual and head gestures during questions was associated with next speakers responding faster than when the questions were unaccompanied by gestures. This does not mean that they anticipated the end of the turn less well when gestures were present. In conversation, many factors influence turn transition times, such as pragmatics (Kendrick & Torreira, 2015; Roberts et al., 2015) or group size (Holler et al., 2021), to name but two examples. That is, interlocutors may be able to anticipate a turn end rather precisely, but may sometimes choose to hold off answering to not make their answer appear too prompt, and other times may respond particularly fast to claim the right to a turn when there is competition, amongst other reasons. Thus, the present study taps into one of the cognitive mechanisms core to turn-taking coordination—the ability to anticipate turn ends as precisely as possible—a fundamental prerequisite for timing one's next contribution in a way that best fits the dynamics and pragmatics of conversation at a given moment (which may involve purposefully coming in early, with a delay and so forth).

An important question for further research is how the role of a certain visual signal in turn end anticipation may differ depending on which other signals are present, since past work has demonstrated the interaction with visible speech cues, for example (see e.g. Drijvers & Özyürek, 2017; Krason, Fenton, Varley, & Vigliocco, 2022; Zhang, Frassinelli, Tuomainen, Skipper, & Vigliocco, 2021). It has been argued that multimodal language processing may involve the binding of different signals (verbal and visual) into holistic 'gestalts' which are perceived differently from the sum of their parts (Holler & Levinson, 2019; Trujillo & Holler, 2023; see also Lücking and Ginzburg (2023) and Mondada (2014) on the notion of multimodal gestalts). Following this idea, it is possible that for example a gaze shift to the other interlocutor only impacts turn end anticipation if it co-occurs with a head tilt. The experiments presented here cannot tell us about such interactions, but future research could for example use virtual avatars to manipulate the presence of individual visual signals as well as their combinations.

Moreover, the present study builds on previous work by employing a second person rather than a third person perspective, which may have meant that participants' cognitive processing relied more on the 'mentalizing network' (Redcay & Schilbach, 2019), potentially facilitating judgements about when another speaker's turn will come to an end than when doing so from a third person observer perspective. However, since this is not a variable that was systematically manipulated here and thus cannot be disentangled from the visibility of signals, investigating the contribution of the second person perspective and the simulation of turn-taking processes this might evoke requires future investigation.

In conclusion, the present study is the first to show that speaker visibility helps to anticipate the end of speaking turns, most likely driven by the effect of manual gestures. Conversational context helps, too, but does not mitigate the weight carried by visual bodily signals. Both of these factors benefit turn end anticipation with increasing turn duration, making a significant difference especially for longer turns. These findings advance our understanding of the role of visual signals regarding the cognitive processes that underpin conversational turn-taking. Future research with more interactive, situated paradigms is needed to further our insights into how those processes interact with the demands of speech planning and preparation as well as the social processes governing turn-taking behaviour in interaction.

Author credits

Marlijn ter Bekke was involved in study design and in the statistical analysis of the data reported, as well as in drafting the manuscript. Judith Holler and Stephen Levinson conceived of the idea for the study; Judith Holler was responsible for study design, write-up, and lead the data collection, analysis, and revision process. Lina van Otterdijk and Michelle Kühn have substantially contributed to testing the many participants needed for Study 2 and were involved in discussions of its design and planning the implementation. All authors listed have agreed to being author and to the author roles assigned to them.

CRediT authorship contribution statement

Marlijn ter Bekke: Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Stephen C. Levinson:** Conceptualization, Funding acquisition, Supervision. **Lina van Otterdijk:** Data curation, Project administration. **Michelle Kühn:** Data curation, Project administration. **Judith Holler:** Conceptualization, Data curation, Funding acquisition, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Data availability

The manuscript includes a link to the data and code for the statistical analyses used.

Acknowledgements

This research was financially supported through grants from the European Research Council (#269484 awarded to SCL, and #773079 awarded to JH). We would also like to acknowledge the help of two student assistants in acquiring the data (Madelief Lenders and Katharina Menn).

Appendices

A.1. Experiment 1

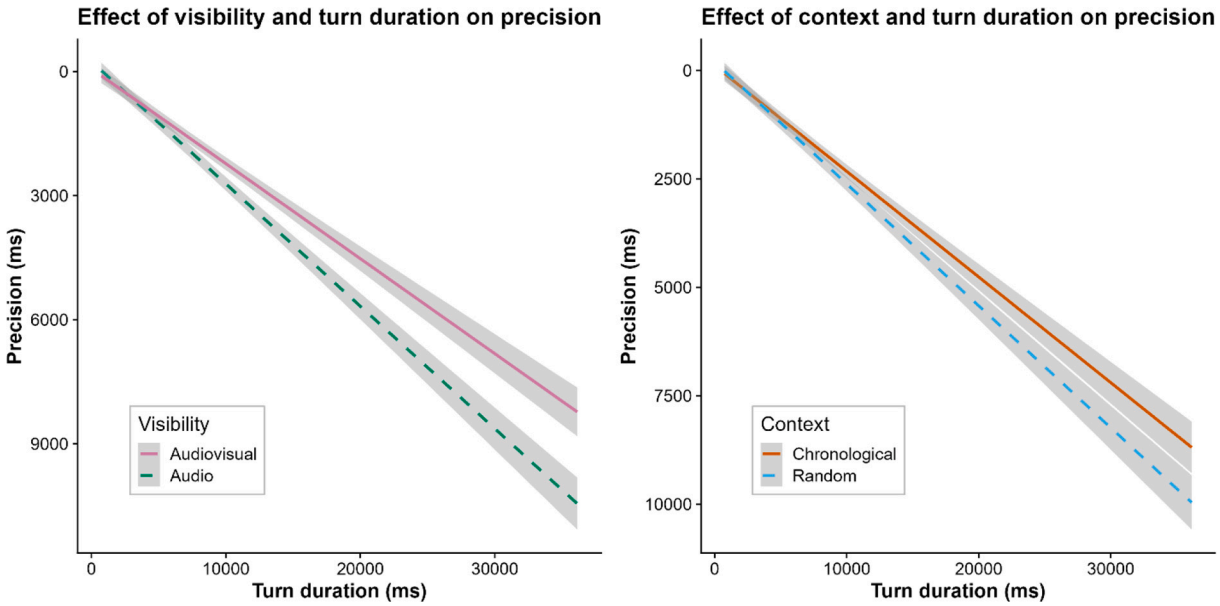

Fig. A1. Effects of visibility (left) and context (right) on the precision of turn end anticipation (absolute values of response time). Precision of 0 ms indicates the target button press exactly at the turn end. Lines indicate the linear fit through the data (plotted with geom_smooth), the shaded areas around the lines represent their confidence intervals. *Left*: Turn end anticipation was more precise for audiovisual than audio-only (i.e. closer to target button press), especially for longer turns. *Right*: Turn end anticipation was more precise for turns presented in chronological than random order (i.e. closer to target button press), especially for longer turns.

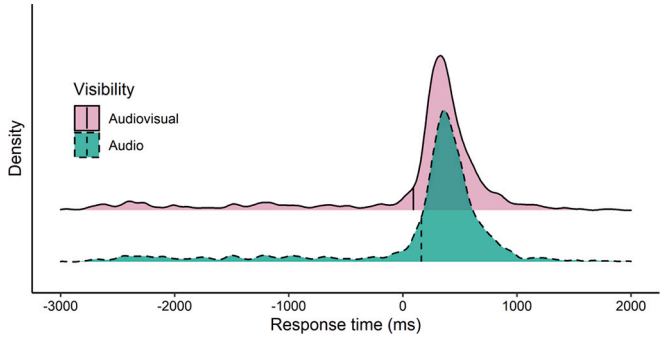

Fig. A2. Response times by visibility. Response time of 0 ms indicates the target button press being exactly at turn end. The vertical line indicates the mean. For visualisation purposes, response times earlier than −3000 ms are not displayed.

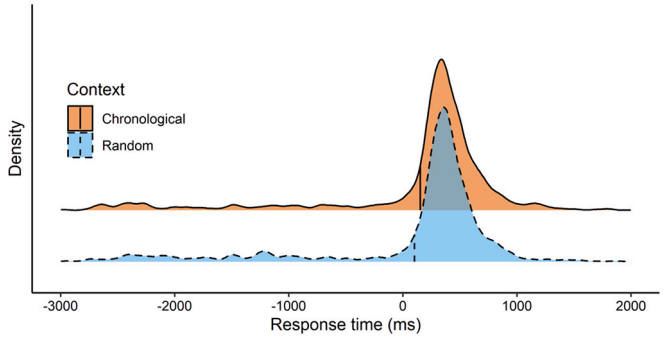

Fig. A3. Response times by context. Response time of 0 ms indicates the target button press exactly at the turn end. The vertical line indicates the mean. For visualisation purposes, response times earlier than −3000 ms are not displayed.

Table A1
Details for the robust model predicting turn end anticipation.

Fixed effects	β	*SE*	*df*	*t*
Intercept	0.21	0.02	3808	10.05***
$\text{Visibility}_{\text{audiovisual}}$	0.01	0.01	1	1.01
$\text{Context}_{\text{chronological}}$	0.01	0.00	1	3.39
Turn duration	−0.12	0.01	129	−11.01***
$\text{Visibility}_{\text{audiovisual}}$ * Turn duration	0.01	0.00	129	4.98***
$\text{Context}_{\text{chronological}}$ * Turn duration	0.02	0.00	129	5.94***
Random effects		Var	SD	
Item	Intercept	0.01	0.11	
Participant	Intercept	0.01	0.10	
	$\text{Visibility}_{\text{audiovisual}}$	0.00	0.03	
Residual		0.03	0.17	

For the fixed effects, estimates (β), standard errors (SE), degrees of freedom (df), t-values (t) are given. For the random effects, variance (var) and standard deviations (SD) are reported. Statistical significance is indicated with asterisks and was determined based on the T-distribution with that degrees of freedom. The subscript audiovisual indicates that these values are for the audiovisual condition with the audio-only condition as reference level (contrast coding: audiovisual 1, audio-only − 1). The subscript chronological indicates that these values are for the chronological order condition with the random order condition as reference level (contrast coding: chronological 1, random − 1).
Formula in R: rlmer (Response Time ~ Visibility * Turn duration + Context * Turn duration + (1 + Visibility | Participant) + (1 | Item)).
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

A.1.1. Entropy analyses

To explore whether there was higher agreement or consistency in participants' button presses when visual information or conversational context was available, we carried out an entropy analysis following de Ruiter et al. (2006), using a bin-width of 250 ms.

When we grouped the data by Visibility (Audio, Audiovisual) to calculate entropy, we found lower entropy for the Audiovisual condition ($M = 1.89$) compared to the Audio-only condition ($M = 2.02$). A linear mixed effects model with Entropy as outcome variable, Visibility as predictor and a random intercept by item confirmed that entropy was lower in the Audiovisual compared to the Audio-only condition (Fig. A4; $\beta = -0.07$, $SE = 0.02$, $t = -2.76$, $p < 0.01$).

When we grouped the data by Context (Random order, Chronological order) to calculate entropy, we found similar entopy in both conditions (Random order: $M = 1.98$; Chronological order: $M = 1.97$). A linear mixed effects model with Entropy as outcome variable, Context as predictor and a random intercept by item confirmed that entropy did not differ between the conditions ($\beta = 0.00$, $SE = 0.02$, $t = -0.10$, $p = 0.92$).

When we grouped the data by both Visibility and Context (Audio-only-Random order, Audio-only-Chronological order, Audiovisual-Random order, Audiovisual-Chronological order) to calculate the entropy, we found no effect of Context ($\beta = 0.03$, $SE = 0.02$, $t = 1.33$, $p = 0.18$). Moreover, there was a trend towards an effect of Visibility on entropy ($\beta = -0.04$, $SE = 0.02$, $t = -1.89$, $p = 0.059$). Finally, there was no interaction between Context and Visibility ($\beta = 0.01$, $SE = 0.02$, $t = 0.25$, $p = 0.80$). Overall, it appears that participants' responses may be more consistent with each other when visual information is available, but not when conversational context is available.

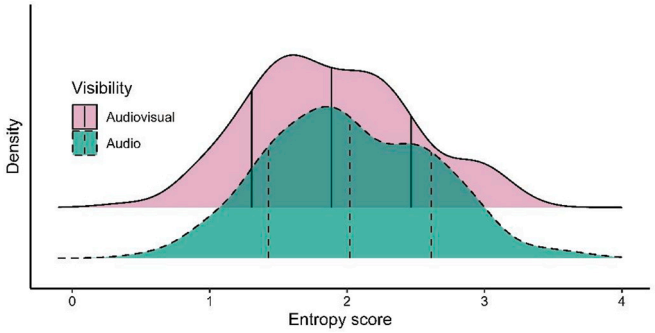

Fig. A4. Entropy of participants' responses by visibility. Entropy of 0 indicates that for each item, all participants responded within the same 250 ms bin. The middle vertical lines indicate the mean, the outer vertical lines indicate one standard deviation from the mean. Turn end anticipation responses may be more consistent with each other (i.e. lower entropy) when visual information is available.

A.2. Experiment 2

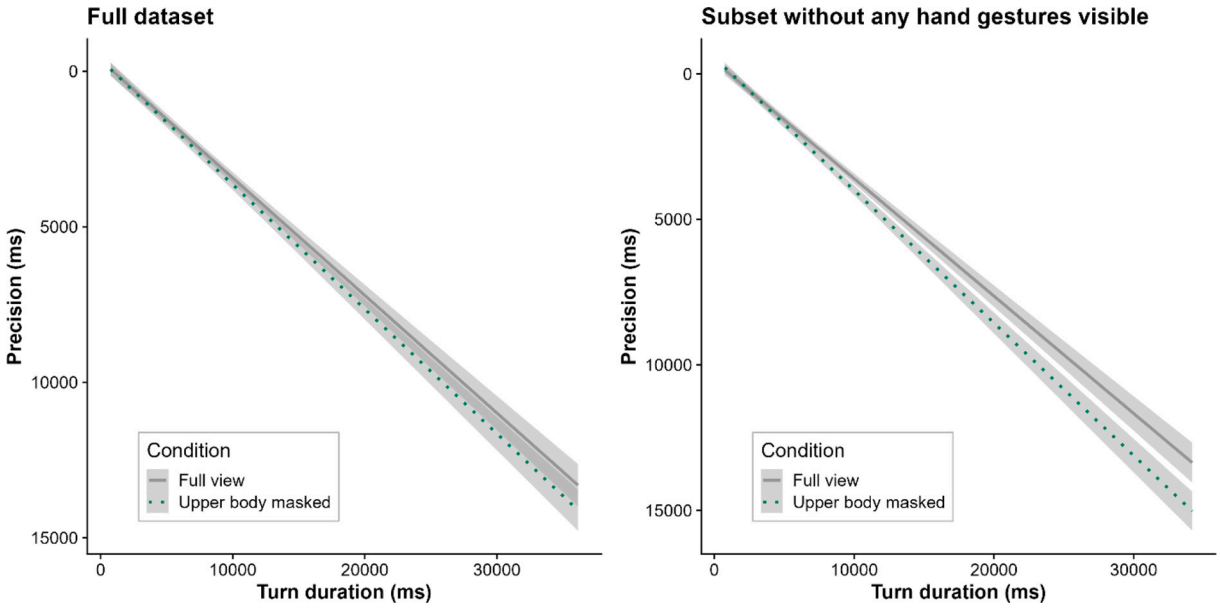

Fig. A5. Effects of visibility and turn duration on the precision of turn end anticipation (absolute values of response time) for the full dataset ($n = 133$ items; left) and the subset of the data where in the upper body masked condition hand gestures were never visible ($n = 117$ items; right). Precision of 0 ms indicates the target button press exactly at the turn end. Lines indicate the linear fit through the data (plotted with geom_smooth), the shaded areas around the lines represent their confidence intervals. Turn end anticipation (response time) was more precise (i.e. closer to target button press) when the turn duration was shorter. Moreover, for the subset (right panel), turn end anticipation was more precise (i.e closer to target button press) for the full view compared to when the upper body was masked, especially for longer turns.

A.2.1. Whole dataset

Table A2
Details for the robust model predicting turn end anticipation (Response Time) based on the whole dataset.

Fixed effects	*β*	*SE*	*df*	*t*
Intercept	0.11	0.02	7745	3.88***
$\text{Visibility}_{\text{eyes\_vs\_full}}$	0.00	0.00	1	0.16
$\text{Visibility}_{\text{head\_vs\_full}}$	0.01	0.00	1	2.37
$\text{Visibility}_{\text{upperbody\_vs\_full}}$	−0.01	0.00	1	−4.96
Turn duration	−0.40	0.03	132	−14.90***
$\text{Visibility}_{\text{eyes\_vs\_full}}$ * Turn duration	−0.00	0.00	132	−1.13
$\text{Visibility}_{\text{head\_vs\_full}}$ * Turn duration	−0.00	0.00	132	−1.09
$\text{Visibility}_{\text{upperbody\_vs\_full}}$ * Turn duration	−0.01	0.00	132	−3.05**
Random effects		Var	SD	
Item	Intercept	0.09	0.30	
Participant	Intercept	0.01	0.07	
Residual		0.02	0.14	

For the fixed effects, estimates (β), standard errors (SE), degrees of freedom (df), t-values (t) are given. For the random effects, variance (var) and standard deviations (SD) are reported. Statistical significance is indicated with asterisks and was determined based on the T-distribution with that degrees of freedom. The subscripts indicate the (sum-to-zero) contrasts, where each condition was compared to Full body (Full body: −1; Other condition: 1).
Formula in R: rlmer (Response Time ~ Visibility * Turn duration + (1 | Participant) + (1 | Item)).
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

A.2.2. Subset without any hand gestures visible in the Upper body mask condition

Table A3
Details for the robust model predicting turn end anticipation (Response Time) based on the subset with items excluded where hand gestures were visible in the Upper body mask condition.

Fixed effects	*β*	*SE*	*df*	*t*
Intercept	0.04	0.03	6812	1.40
$\text{Visibility}_{\text{eyes\_vs\_full}}$	0.00	0.00	1	0.93
$\text{Visibility}_{\text{head\_vs\_full}}$	0.00	0.00	1	1.74
$\text{Visibility}_{\text{upperbody\_vs\_full}}$	−0.02	0.00	1	−6.04

(*continued on next page*)

Table A3 (*continued*)

Fixed effects	β	*SE*	*df*	*t*
Turn duration	−0.54	0.03	115	−18.37***
$\text{Visibility}_{\text{eyes\_vs\_full}}$ * Turn duration	0.00	0.00	115	0.55
$\text{Visibility}_{\text{head\_vs\_full}}$ * Turn duration	−0.00	0.00	115	−1.58
$\text{Visibility}_{\text{upperbody\_vs\_full}}$ * Turn duration	−0.02	0.00	115	−5.09***
Random effects		Var	SD	
Item	Intercept	0.07	0.26	
Participant	Intercept	0.00	0.07	
Residual		0.02	0.13	

For the fixed effects, estimates (β), standard errors (SE), degrees of freedom (df), t-values (t) are given. For the random effects, variance (var) and standard deviations (SD) are reported. Statistical significance is indicated with asterisks and was determined based on the T-distribution with that degrees of freedom. The subscripts indicate the (sum-to-zero) contrasts, where each condition was compared to Full body (Full body: −1; Other condition: 1).
Formula in R: rlmer (Response Time ~ Visibility * Turn duration + (1 | Participant) + (1 | Item)).
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

A.2.3. Subset without any hands visible in the Upper body mask condition

Table A4
Details for the robust model predicting turn end anticipation (Response Time) based on the subset with items excluded where hands were visible in the Upper body mask condition.

Fixed effects	β	*SE*	*df*	*t*
Intercept	0.07	0.03	6351	2.53
$\text{Visibility}_{\text{eyes\_vs\_full}}$	0.00	0.00	1	0.58
$\text{Visibility}_{\text{head\_vs\_full}}$	0.00	0.00	1	1.27
$\text{Visibility}_{\text{upperbody\_vs\_full}}$	−0.02	0.00	1	−5.11
Turn duration	−0.47	0.03	108	−14.44
$\text{Visibility}_{\text{eyes\_vs\_full}}$ * Turn duration	−0.00	0.00	108	−0.06
$\text{Visibility}_{\text{head\_vs\_full}}$ * Turn duration	−0.00	0.00	108	−1.08
$\text{Visibility}_{\text{upperbody\_vs\_full}}$ * Turn duration	−0.02	0.00	108	−4.42
Random effects		Var	SD	
Item		0.07	0.26	
Participant	Intercept	0.00	0.07	
	$\text{Visibility}_{\text{eyes\_vs\_full}}$	0.00	0.03	
	$\text{Visibility}_{\text{head\_vs\_full}}$	0.00	0.02	
	$\text{Visibility}_{\text{upperbody\_vs\_full}}$	0.00	0.01	
Residual		0.02	0.12	

For the fixed effects, estimates (β), standard errors (SE), degrees of freedom (df), t-values (t) are given. For the random effects, variance (var) and standard deviations (SD) are reported. Statistical significance is indicated with asterisks and was determined based on the T-distribution with that degrees of freedom. The subscripts indicate the (sum-to-zero) contrasts, where each condition was compared to Full body (Full body: −1; Other condition: 1).
Formula in R: rlmer (Response Time ~ Visibility * Turn duration + (1 + Visibility | Participant) + (1 | Item)).
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

References

Arnqvist, G. (2020). Mixed models offer no freedom from degrees of freedom. *Trends in Ecology & Evolution, 35*(4), 329–335. https://doi.org/10.1016/j.tree.2019.12.004

Barkhuysen, P., Krahmer, E., & Swerts, M. (2008). The interplay between the auditory and visual modality for end-of-utterance detection. *The Journal of the Acoustical Society of America, 123*(1), 354–365. https://doi.org/10.1121/1.2816561

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

Barthel, M., & Levinson, S. C. (2020). Next speakers plan word forms in overlap with the incoming turn: Evidence from gaze-contingent switch task performance. *Language, Cognition and Neuroscience, 35*(9), 1183–1202. https://doi.org/10.1080/23273798.2020.1716030

Barthel, M., Meyer, A. S., & Levinson, S. C. (2017). Next speakers plan their turn early and speak after turn-final "go-signals". *Frontiers in Psychology, 8*. https://doi.org/10.3389/fpsyg.2017.00393

Barthel, M., Sauppe, S., Levinson, S. C., & Meyer, A. S. (2016). The timing of utterance planning in task-oriented dialogue: Evidence from a novel list-completion paradigm. *Frontiers in Psychology, 7*. https://www.frontiersin.org/article/10.3389/fpsyg.2016.01858.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Bavelas, J. B. (2022). *Face-to-face dialogue: Theory, research, and applications*. Oxford University Press.

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., Coates, L., & Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. *Journal of Communication, 52*(3), 566–580. https://doi.org/10.1111/j.1460-2466.2002.tb02562.x

Beattie, G. W., Cutler, A., & Pearson, M. (1982). Why is Mrs Thatcher interrupted so often? *Nature, 300*(5894). https://doi.org/10.1038/300744a0. Article 5894.

ter Bekke, M., Drijvers, L., & Holler, J. (2024). Hand gestures have predictive potential during conversation: An investigation of the timing of gestures in relation to speech. *Cognitive Science, 48*(1), Article e13407. https://doi.org/10.1111/cogs.13407

Bernstein, L. E., Auer, E. T., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication, 44*(1–4), 5–18. https://doi.org/10.1016/j.specom.2004.10.011

Bi, R., & Swerts, M. (2017). A perceptual study of how rapidly and accurately audiovisual cues to utterance-final boundaries can be interpreted in Chinese and English. *Speech Communication, 95*, 68–77. https://doi.org/10.1016/j.specom.2017.07.002

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*(9/10), 341–345.

Bögels, S. (2020). Neural correlates of turn-taking in the wild: Response planning starts early in free interviews. *Cognition, 203*, Article 104347. https://doi.org/10.1016/j.cognition.2020.104347

Bögels, S., Casillas, M., & Levinson, S. C. (2018). Planning versus comprehension in turn-taking: Fast responders show reduced anticipatory processing of the question. *Neuropsychologia, 109*, 295–310. https://doi.org/10.1016/j.neuropsychologia.2017.12.028

Bögels, S., Magyari, L., & Levinson, S. C. (2015). Neural signatures of response planning occur midway through an incoming question in conversation. *Scientific Reports, 5*, 12881. https://doi.org/10.1038/srep12881

Bögels, S., & Torreira, F. (2015). Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics, 52*, 46–57. https://doi.org/10.1016/j.wocn.2015.04.004

Bögels, S., & Torreira, F. (2021). Turn-end estimation in conversational turn-taking: The roles of context and prosody. *Discourse Processes, 58*(10), 903–924. https://doi.org/10.1080/0163853X.2021.1986664

Boiteau, T. W., Malone, P. S., Peters, S. A., & Almor, A. (2013). Interference between conversation and a concurrent visuomotor task. *Journal of Experimental Psychology: General, 143*(1), 295. https://doi.org/10.1037/a0031858

Casillas, M., & Frank, M. C. (2017). The development of children's ability to track and predict turn structure in conversation. *Journal of Memory and Language, 92*, 234–253. https://doi.org/10.1016/j.jml.2016.06.013

Christiansen, M. (2013). *Adobe after effects CC visual effects and compositing studio techniques*. Adobe Press.

Corps, R. E., Crossley, A., Gambi, C., & Pickering, M. J. (2018). Early preparation during turn-taking: Listeners use content predictions to determine what to say but not when to say it. *Cognition, 175*, 77–95. https://doi.org/10.1016/j.cognition.2018.01.015

Corps, R. E., Gambi, C., & Pickering, M. J. (2018). Coordinating utterances during turn-taking: The role of prediction, response preparation, and articulation. *Discourse Processes, 55*(2), 230–240. https://doi.org/10.1080/0163853X.2017.1330031

Corps, R. E., Gambi, C., & Pickering, M. J. (2020). How do listeners time response articulation when answering questions? The role of speech rate. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*(4), 781–802. https://doi.org/10.1037/xlm0000759

Corps, R. E., Pickering, M. J., & Gambi, C. (2019). Predicting turn-ends in discourse context. *Language, Cognition and Neuroscience, 34*(5), 615–627. https://doi.org/10.1080/23273798.2018.1552008

Crawley, M. J. (2007). *The R book*. John Wiley & Sons Ltd.

Danner, S. G., Krivokapić, J., & Byrd, D. (2021). Co-speech movement in conversational turn-taking. *Frontiers in Communication, 6*, Article 779814. https://doi.org/10.3389/fcomm.2021.779814

De Kok, I., & Heylen, D. (2009). Multimodal end-of-turn prediction in multi-party meetings. In *Proceedings of the 2009 international conference on multimodal interfaces* (pp. 91–98). https://doi.org/10.1145/1647314.1647332

de Vos, C., Casillas, M., Uittenbogert, T., Crasborn, O., & Levinson, S. C. (2021). Predicting conversational turns: Signers' and non-signers' sensitivity to language-specific and globally accessible cues. *Language, 98*(1), 35–62. https://doi.org/10.1353/lan.2021.0085

Degutyte, Z., & Astell, A. (2021). The role of eye gaze in regulating turn taking in conversations: A systematized review of methods and findings. *Frontiers in Psychology, 12*. https://www.frontiersin.org/articles/10.3389/fpsyg.2021.616471.

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

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology, 23*(2), 283–292. https://doi.org/10.1037/h0033031

Duncan, S. (1974). On the structure of speaker–auditor interaction during speaking turns. *Language in Society, 3*(2), 161–180. https://doi.org/10.1017/S0047404500004322

Duncan, S., & Niederehe, G. (1974). On signalling that it's your turn to speak. *Journal of Experimental Social Psychology, 10*(3), 234–247. https://doi.org/10.1016/0022-1031(74)90070-5

Ferré, G.. Timing relationships between speech and co-verbal gestures in spontaneous French. https://hal.archives-ouvertes.fr/hal-00485797.

Ford, C. E., & Thompson, S. A. (1996). Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. In E. Ochs, E. A. Schegloff, & S. A. Thompson (Eds.), *Interaction and grammar* (pp. 134–184). Cambridge University Press. https://doi.org/10.1017/CBO9780511620874.003.

Gambi, C., Jachmann, T., & Staudte, M. (2015). The role of prosody and gaze in turn-end anticipation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 764–769.

Garrod, S., & Pickering, M. J. (2015). The use of content and timing to predict turn transitions. *Frontiers in Psychology, 6*, 751. https://doi.org/10.3389/fpsyg.2015.00751

Goldin-Meadow, S. (2017). What the hands can tell us about language emergence. *Psychonomic Bulletin & Review, 24*(1), 213–218. https://doi.org/10.3758/s13423-016-1074-x

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), 1197–1208. https://doi.org/10.1121/1.1288668

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science, 11*(4), 274–279. https://doi.org/10.1111/1467-9280.00255

Heldner, M. (2011). Detection thresholds for gaps, overlaps, and no-gap-no-overlaps. *The Journal of the Acoustical Society of America, 130*(1), 508–513. https://doi.org/10.1121/1.3598457

Hirvenkari, L., Ruusuvuori, J., Saarinen, V.-M., Kivioja, M., Peräkylä, A., & Hari, R. (2013). Influence of turn-taking in a two-person conversation on the gaze of a viewer. *PLoS One, 8*(8), Article e71569. https://doi.org/10.1371/journal.pone.0071569

Ho, S., Foulsham, T., & Kingstone, A. (2015). Speaking and listening with the eyes: Gaze signaling during dyadic interactions. *PLoS One, 10*(8), Article e0136905. https://doi.org/10.1371/journal.pone.0136905

Holler, J., Alday, P. M., Decuyper, C., Geiger, M., Kendrick, K. H., & Meyer, A. S. (2021). Competition reduces response times in multiparty conversation. *Frontiers in Psychology, 12*, 693124. https://doi.org/10.3389/fpsyg.2021.693124

Holler, J., Kendrick, K. H., & Levinson, S. C. (2018). Processing language in face-to-face conversation: Questions with gestures get faster responses. *Psychonomic Bulletin & Review, 25*(5), 1900–1908. https://doi.org/10.3758/s13423-017-1363-z

Holler, J., & Levinson, S. C. (2019). Multimodal language processing in human communication. *Trends in Cognitive Sciences, 23*(8). https://doi.org/10.1016/j.tics.2019.05.006

Indefrey, P., & Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition, 92*(1–2), 101–144. https://doi.org/10.1016/j.cognition.2002.06.001

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica, 26*, 22–63. https://doi.org/10.1016/0001-6918(67)90005-4

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

Kendon, A. (2017). Reflections on the "gesture-first" hypothesis of language origins. *Psychonomic Bulletin & Review, 24*(1), 163–170. https://doi.org/10.3758/s13423-016-1117-3

Kendrick, K. H., Holler, J., & Levinson, S. C. (2023). Turn-taking in human face-to-face interaction is multimodal: Gaze direction and manual gestures aid the coordination of turn transitions. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 378*(1875), 20210473. https://doi.org/10.1098/rstb.2021.0473

Kendrick, K. H., & Torreira, F. (2015). The timing and construction of preference: A quantitative study. *Discourse Processes, 52*(4), 255–289. https://doi.org/10.1080/0163853X.2014.955997

Koller, M. (2016). robustlmm: An R package for robust estimation of linear mixed-effects models. *Journal of Statistical Software, 75*, 1–24. https://doi.org/10.18637/jss.v075.i06

Krason, A., Fenton, R., Varley, R., & Vigliocco, G. (2022). The role of iconic gestures and mouth movements in face-to-face communication. *Psychonomic Bulletin & Review, 29*(2), 600–612. https://doi.org/10.3758/s13423-021-02009-5

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*, 1–26. https://doi.org/10.18637/jss.v082.i13

Lammertink, I., Casillas, M., Benders, T., Post, B., & Fikkert, P. (2015). Dutch and English toddlers' use of linguistic cues in predicting upcoming turn transitions. *Frontiers in Psychology, 6*, 495.doi. https://doi.org/10.3389/fpsyg.2015.00495

Latif, N., Alsius, A., & Munhall, K. G. (2017). Seeing the way: The role of vision in conversation turn exchange perception. *Multisensory Research, 30*(7–8), 653–679. https://doi.org/10.1163/22134808-00002582

Latif, N., Alsius, A., & Munhall, K. G. (2018). Knowing when to respond: The role of visual information in conversational turn exchanges. *Attention, Perception, & Psychophysics, 80*(1), 27–41. https://doi.org/10.3758/s13414-017-1428-0

Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods, 41*(3). https://doi.org/10.3758/BRM.41.3.841

Levinson, S. C. (2016). Turn-taking in human communication – Origins and implications for language processing. *Trends in Cognitive Sciences, 20*(1), 6–14. https://doi.org/10.1016/j.tics.2015.10.010

Levinson, S. C., & Holler, J. (2014). The origin of human multi-modal communication. *Philosophical Transactions of the Royal Society B, 369*(1651), 20130302. https://doi.org/10.1098/rstb.2013.0302

Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology, 6*, 731. https://doi.org/10.3389/fpsyg.2015.00731

Local, J. K., Kelly, J., & Wells, W. H. G. (1986). Towards a phonology of conversation: Turn-taking in Tyneside English1. *Journal of Linguistics, 22*(2), 411–437. https://doi.org/10.1017/S0022226700010859

Lücking, A., & Ginzburg, J. (2023). Leading voices: Dialogue semantics, cognitive science and the polyphonic structure of multimodal interaction. *Language and Cognition, 15*(1), 148–172. https://doi.org/10.1017/langcog.2022.30

Magyari, L., & de Ruiter, J. P. (2012). Prediction of turn-ends based on anticipation of upcoming words. *Frontiers in Psychology, 3*, 376. https://doi.org/10.3389/fpsyg.2012.00376

Magyari, L., de Ruiter, J. P., & Levinson, S. C. (2017). Temporal preparation for speaking in question-answer sequences. *Frontiers in Psychology, 8*, 211. https://doi.org/10.3389/fpsyg.2017.00211

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(4), 158–161. https://doi.org/10.1080/00357529.1980.11764651

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press.

McNeill, D. (2012). *How language began: Gesture and speech in human evolution*. Cambridge University Press. https://doi.org/10.1017/CBO9781139108669

Mixdorff, H., Hönemann, A., Kim, J., & Davis, C. (2015). *Anticipation of turn-switching in auditory-visual dialogs*. Proceedings of FAAVSP.

Mondada, L. (2007). Multimodal resources for turn-taking: Pointing and the emergence of possible next speakers. *Discourse Studies, 9*(2), 194–225. 10.1177/1461445607075346.

Mondada, L. (2014). The local constitution of multimodal resources for social interaction. *Journal of Pragmatics, 65*, 137–156. https://doi.org/10.1016/j.pragma.2014.04.004

Näätänen, R. (1970). The diminishing time-uncertainty with the lapse of time after the warning signal in reaction-time experiments with varying fore-periods. *Acta Psychologica, 34*, 399–419. https://doi.org/10.1016/0001-6918(70)90035-1

Nota, N., Trujillo, J. P., & Holler, J. (2023). Specific facial signals associate with categories of social actions conveyed through questions. *PLoS One, 18*(7), Article e0288104. https://doi.org/10.1371/journal.pone.0288104

Özyürek, A. (2014). Hearing and seeing meaning in speech and gesture: Insights from brain and behaviour. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 369*(1651), 20130296. https://doi.org/10.1098/rstb.2013.0296

Preisig, B. C., Eggenberger, N., Zito, G., Vanbellingen, T., Schumacher, R., Hopfner, S., … Müri, R. M. (2016). Eye gaze behavior at turn transition: How aphasic patients process speakers' turns during video observation. *Journal of Cognitive Neuroscience, 28*(10), 1613–1624. https://doi.org/10.1162/jocn_a_00983

Redcay, E., & Schilbach, L. (2019). Using second-person neuroscience to elucidate the mechanisms of social interaction. *Nature Reviews Neuroscience, 20*(8), 495–505. https://doi.org/10.1038/s41583-019-0179-4

Riest, C., Jorschick, A. B., & de Ruiter, J. P. (2015). Anticipation in turn-taking: Mechanisms and information sources. *Frontiers in Psychology, 6*, 89. https://doi.org/10.3389/fpsyg.2015.00089

Roberts, S. G., Torreira, F., & Levinson, S. C. (2015). The effects of processing and sequence organization on the timing of turn taking: A corpus study. *Frontiers in Psychology, 6*, 509. https://doi.org/10.3389/fpsyg.2015.00509

Rossano, F. (2012). *Gaze behavior in face-to-face interaction*. Nijmegen, The Netherlands: Radboud University Nijmegen. http://hdl.handle.net/2066/99151.

de Ruiter, J. P., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language, 82*(3), 515–535. https://doi.org/10.1353/lan.2006.0130

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*(4), 696–735. https://doi.org/10.2307/412243

Saville, D. J., & Wood, G. R. (1991). Latin square design. In D. J. Saville, & G. R. Wood (Eds.), *Statistical methods: The geometric approach* (pp. 340–353). Springer. https://doi.org/10.1007/978-1-4612-0971-3_13.

Schegloff, E. A. (1982). Discourse as an interactional achievement: Some uses of "uh huh" and other things that come between sentences. In D. Tannen (Ed.), *Analyzing discourse: Text and talk* (pp. 71–93). Georgetown University Press.

Sjerps, M. J., & Meyer, A. S. (2015). Variation in dual-task performance reveals late initiation of speech planning in turn-taking. *Cognition, 136*, 304–324. https://doi.org/10.1016/j.cognition.2014.10.008

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., … Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences, 106*(26), 10587–10592. https://doi.org/10.1073/pnas.0903616106

Streeck, J. (2014). Mutual gaze and recognition: Revisiting Kendon's "Gaze direction in two-person conversation". In *In M. Seyfeddinipur & M. Gullberg (Eds.), From gesture in conversation to visible vction as utterance: Essays in honor of Adam Kendon* (pp. 35–56). John Benjamins Publishing Company. https://benjamins.com/catalog/z.188.03str.

Streeck, J., & Hartge, U. (1992). Previews: Gestures at the transition place. In P. Auer, & A. D. Luzio (Eds.), *The contextualization of language* (pp. 135–157). John Benjamins Publishing Company.

The R Core Team. (2018). *R: A language and environment for statistical computing*. R Core Team. https://www.R-project.org/.

Trujillo, J. P., & Holler, J. (2023). Interactionally embedded gestalt principles of multimodal human communication. *Perspectives on Psychological Science, 18*(5), 1136–1159. https://doi.org/10.1177/17456916221141422

Trujillo, J. P., Levinson, S. C., & Holler, J. (2021). Visual information in computer-mediated interaction matters: Investigating the association between the availability of gesture and turn transition timing in conversation. In M. Kurosu (Ed.), *Human-computer interaction. Design and user experience case studies* (pp. 643–657). Springer International Publishing. https://doi.org/10.1007/978-3-030-78468-3_44.

Turk, A., & Shattuck-Hufnagel, S. (2014). Timing in talking: What is it used for, and how is it controlled? *Philosophical Transactions of the Royal Society, B: Biological Sciences, 369*(1658), 20130395. https://doi.org/10.1098/rstb.2013.0395

Vigliocco, G., Perniss, P., & Vinson, D. (2014). Language as a multimodal phenomenon: Implications for language learning, processing and evolution. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 369*(1651), 20130292. https://doi.org/10.1098/rstb.2013.0292

Wilson, M., & Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review, 12*(6), 957–968. https://doi.org/10.3758/BF03206432

Zellers, M., House, D., & Alexanderson, S. (2016). Prosody and hand gesture at turn boundaries in Swedish. *Speech Prosody*, 831–835. https://doi.org/10.21437/SpeechProsody.2016-170

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