Visual Information in Computer-Mediated Interaction Matters: Investigating the Association Between the Availability of Gesture and Turn Transition Timing in Conversation

James P. Trujillo[1,2(✉)], Stephen C. Levinson[2], and Judith Holler[1,2]

[1] Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, The Netherlands
j.trujillo@donders.ru.nl

[2] Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525XD Nijmegen, The Netherlands

Abstract. Natural human interaction involves the fast-paced exchange of speaker turns. Crucially, if a next speaker waited with planning their turn until the current speaker was finished, language production models would predict much longer turn transition times than what we observe. Next speakers must therefore prepare their turn in parallel to listening. Visual signals likely play a role in this process, for example by helping the next speaker to process the ongoing utterance and thus prepare an appropriately-timed response.

To understand how visual signals contribute to the timing of turn-taking, and to move beyond the mostly qualitative studies of gesture in conversation, we examined unconstrained, computer-mediated conversations between 20 pairs of participants while systematically manipulating speaker visibility. Using motion tracking and manual gesture annotation, we assessed 1) how visibility affected the timing of turn transitions, and 2) whether use of co-speech gestures and 3) the communicative kinematic features of these gestures were associated with changes in turn transition timing.

We found that 1) decreased visibility was associated with less tightly timed turn transitions, and 2) the presence of gestures was associated with more tightly timed turn transitions across visibility conditions. Finally, 3) structural and salient kinematics contributed to gesture's facilitatory effect on turn transition times.

Our findings suggest that speaker visibility--and especially the presence and kinematic form of gestures--during conversation contributes to the temporal coordination of conversational turns in computer-mediated settings. Furthermore, our study demonstrates that it is possible to use naturalistic conversation and still obtain controlled results.

Keywords: Multimodality · Kinematics · Adaptation

1 Introduction

Natural human interaction involves the fast-paced exchange of speaker turns. Remarkably, the gap between when one person stops speaking and the other begins speaking

M. Kurosu (Ed.): HCII 2021, LNCS 12764, pp. 643–657, 2021.
https://doi.org/10.1007/978-3-030-78468-3_44

is quite small, on the order of 200 ms (Stivers et al. 2009). This has posed an interesting challenge to models of language production and comprehension because if a next speaker were to wait with planning their turn until the end of the current speaker's turn, production models would predict inter-speaker gaps to be on the order of 500–600 ms, minimally, even for very short responses (Levinson 2016; Levinson and Torreira 2015). In order for this tight temporal coordination to be possible, next speakers must therefore prepare their turn in parallel to listening (Heldner and Edlund 2010; Levinson and Torreira 2015). Sensitivity to turn-final cues allows next speakers to then launch their pre-planned turn on time.

Human interaction is not, however, limited only to speech. In face-to-face interaction, the natural environment of language use, visual signals such as hand gestures are tightly integrated with speech ('co-speech gestures'), forming a multimodal communicative system (Bavelas and Chovil 2000; Holler and Levinson 2019; Kendon 2004; McNeill 1992). This multimodality also plays a role in the timing of speaker turn transitions it appears, as evidenced by the finding of questions receiving faster responses (i.e., shorter turn transition times) when they were uttered with gestures compared with responses to questions without a gestural component (Holler et al. 2018; ter Bekke 2020). These results therefore suggest that visual signals play an important role in efficient turn-taking behavior. However, the process by which this happens it not well understood.

The relation between gesture and inter-speaker turn transitions can be further elucidated by looking at how turn transition times are affected by the visibility of co-speech gestures, as well as how they are related to the communicative kinematic features of gestures. Mostly in gesture studies we have to be content with correlational accounts, but if it is possible to control precisely what an interlocutor sees, then it may be possible to give a stronger account of what features of gesture contribute to turn-taking efficiency. For example, using a computer-mediated channel it is possible to explore more precisely how gestures may contribute to fast turn transitions, by successively downgrading the visibility of visual signals. If very small, subtle finger movements and handshapes are most critical, then a beneficial effect of gestures on turn timing should mainly be observed during very low blur grades since this information disappears with more degrees of blur. However, if slightly more coarse-grained gestural features, such as larger handshapes and arm movements are most beneficial, then the effect of the presence of gestures may be most pronounced during medium blur grades. Here, the gestural information may also stand out more since other, non-gestural signals that can facilitate turn-taking (such as eye gaze) are not available anymore. However, if very coarse-grained gestural movements are most beneficial, then the effect of gesture on turn-transition should still be evident during stronger blur grades. Here, it may also be that the termination of gestures (e.g., the onset of the retraction) may stand out more than other information. This issue of granularity may be particularly relevant for the domain of human-computer-interaction, as the granularity of motion-tracking or motion recognition algorithms should reflect the way visual signals are used and understood in natural interaction.

Beyond the facilitatory presence of gestures, it is currently unclear how exactly gestures contribute to turn-taking timing. According to the seminal turn-taking model by Sacks et al. (1974), speakers try to minimize the turn-transition time or in other words, aim to reduce both gaps and overlaps as much as possible (Levinson and Torreira 2015).

Two questions arise. First, is it possible to replicate the finding from Holler et al. (2018) that showed a correlation between gestures and faster response times, and to do so in a computer-mediated setting which allows more systematic investigation? Second, is it possible to show that gestures not only speed responses but also – through indicating forthcoming speaker termination – lower the amount of overlap?

While manipulating visibility can be informative about the granularity of visual information that is contributing to turn-taking timing, investigating how different gesture types contribute to turn-taking can be informative on a more functional level. Specifically, while some types of gestures contribute largely to semantic meaning (e.g., iconics, metaphorics, McNeill 1992), others contribute more to the temporal structural and emphasis of an utterance (e.g., beats; McNeill 1992), and yet others more to the dialogic process between the interlocutors (e.g., interactive gestures, Bavelas et al. 1995). Note that these 'type categories' should not be considered as mutually exclusive, but rather as dimensions, with the foregrounded dimension serving to class a given gesture as one or the other type for the purpose of analysis. However, one such type of gesture may contribute more than others to turn-taking timing, shedding light on how gestures facilitate turn-transition times.

Further, examining the specific kinematic features of a gesture can provide insights into whether the form of the gesture, beyond how visible it is, also influences its effect on turn timing. This is relevant because kinematic features of gestures (such as their size, location in gesture space, etc.) have previously been linked to semantic and pragmatic functions (e.g., Campisi and Özyürek 2013; Gerwing and Bavelas 2004; Holler and Wilkin 2011; Trujillo et al. 2019, 2020). For example, subtle differences in kinematics such as the position of the gesture in space can signal communicative intention (Trujillo et al. 2018), while temporal features, such as how clearly segmented the gesture is into smaller constituent movements aids in the processing of the gesture's meaning (Trujillo et al. 2019). If, as discussed in Holler et al. (2018), utterances with co-occurring gestures lead to faster responses due to a facilitation of processing, then kinematically salient and clear gestures (e.g., those that are larger, more centrally located, or with clearly segmented structures) should be associated with even shorter turn-transition times when compared to gestures with less salient kinematics. Therefore, understanding which kinematic features relate to the duration of turn transitions, and how this interacts with the visibility of gestures, can provide further insights into how gestures may be influencing turn-taking behavior.

In order to assess how gesture contributes to the timing of turn-taking, we quantified the kinematics of co-speech gestures performed during unscripted, naturalistic, casual conversation. In order to manipulate the visual availability of the gestural signal, we employed a special computer-mediated set-up: Participants in each dyad were located in separate rooms and saw each other via a live feed, displayed on a large screen directly in front of them. The size of the screen ensured that the displayed image was close to life-sized. This set-up provides a crucial advantage to our study, as most studies of gesture use in conversation are mostly qualitative or correlational in nature. Utilizing computer-mediated interaction of this type allows an extra degree of experimental control, allowing us to more systematically examine the relationship between visual signals and turn-taking behavior.

Critically, the visual quality changed in ten separate steps (blur grades), with each step occurring after four minutes. This was performed in order to test whether the visibility of gestures, as well as any kinematic changes in their production (e.g., larger versus smaller movements) would affect turn transitions. Because gestures have been shown to be associated with a decrease in overall turn-transition time in face-to-face interaction, we suspected their gradual disappearance as a consequence of the increase in blur grade to also affect turn transition times in the present computer-mediated setting (i.e., to result in an increase of turn transition time). As visibility was systematically reduced over ten blur grades, we were also able to investigate the general shape of gesture's contribution to turn-transition times. We expected gestures to facilitate turn-transition timing the most during moderate visibility conditions, as turn-taking timing may be negatively affected by the decreased general visibility of the speaker, but gestures are expected to be visible enough to still facilitate. As visibility becomes more severely degraded, the gestural benefit will likely decrease again. Similarly, in low blur grades (i.e., high visibility), additional, potentially more spatially fine-grained cues to turn-taking, such as eye-gaze (Kendon 1967), are still visible. Gestures are therefore expected to stand out most in moderate compared to low blur grades.

To summarise, we first tested whether (RQ1) visibility of the speaker, and thus the visibility of visual signals, affected turn transition time. We expected turn-transition times to become less tight, with more gaps and overlaps. Next, we tested whether we could replicate the finding of (RQ2) gestures relating to faster responses, and whether this facilitation is specifically related to overall faster responses (i.e., shorter gaps, longer overlaps) or to more tightly coordinated turn-transitions (i.e., shorter gaps, shorter overlaps). As a secondary test, we also assess whether certain gesture types influence turn-transition times more than other types. Finally, we tested whether (RQ3) communicatively relevant gesture kinematics are associated with turn transition time. We expected increased kinematic salience to be associated with a decrease in gap and overlap duration.

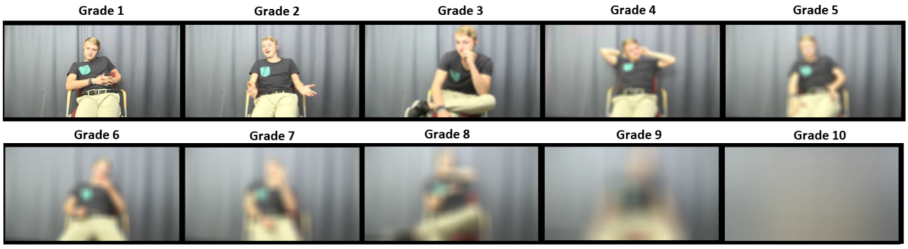

Fig. 1. Visual depiction of the actual blur grades used in this study

2 Methods

2.1 Participants

Data were initially collected from 38 dyads (74 participants), recruited from the participant database of the Max Planck Institute for Psycholinguistics. Dyads were excluded if

there were technical errors in the audio or visual recordings of either participant (often caused by problems with the script implementing the blur grades), leaving 20 dyads (40 participants total; 31 females, mean age: 24.23 years) for analysis. Participants were all native speakers of Dutch and were familiar with the respective other person in the dyad.

2.2 Task

Dyads engaged in a natural, unscripted conversation, mediated by an audio-video based telecommunication interface, for 40 min each. Participants in each dyad were located in separate rooms and saw each other via a live feed, displayed on a large screen (27″, 16:9, 1920 × 1080 px, full HD) directly in front of them. The size of the screen ensured that the displayed image was close to life-sized. Visual quality changed in ten separate steps, with each step occurring after four minutes (see Fig. 1). The direction of quality change (clear to full blur or full blur to clear) was randomized across participants. Ten dyads went from blurry to clear, and ten dyads went from clear to blurry. Conversations were not constrained, and participants were instructed to simply talk as they would if they met at home or in a café for the duration of the experiment.

2.3 Audio Segmentation and Turn Transition Time

We performed automatic segmentation of the audio stream of each participant into speech and silence, and further segmented speech into discrete utterances following the methodology of Heldner and Edlund (2010) and Roberts et al. (2015). These analyses were performed using custom Python (version 3.7) scripts.

In order to segment the audio stream into speech and silence, we first calculated the amplitude of the audio file of each separate participant using the Python package Parselmouth (Jadoul et al. 2018) which allows direct interfacing of Python with PRAAT (Boersma and Weenink 2007) functions. Next, we used a threshold of 35 dB below the maximum value to separate the file into speech and silence. In order to exclude backchannels, we removed any speech segment of less than 700 ms, as up to 75% of speech segments less than 700 ms are likely to be backchannels (Roberts et al. 2015). We then merged any speech segments that were separated by less than 180 ms in order to reduce the risk of stop closures being mistaken for pauses between utterances (Heldner and Edlund 2010).

Based on the resulting speech annotations we calculated turn transition time. This was done by taking the offset (i.e., end time) of each speech utterance and finding the nearest speech onset time. For the calculation of nearest utterance, we used a window of 2 s, following the methodology of Heldner and Edlund (2010), who similarly used data from Dutch speakers (Heldner and Edlund 2010). The nearest onset could be either a successive speech utterance by the same speaker, or a new speech utterance by the other speaker. In the case of the nearest onset being produced by the other speaker, the time difference between the offset of the first utterance (speaker A) and onset of the second utterance (speaker B) was taken as the turn transition time. Therefore, negative values indicate an overlap, where speaker B started their utterance before speaker A stopped speaking, while positive values indicate gaps. Analyses were carried out on

the total turn-transition time, consisting of both negative and positive values, as well as separately on gaps and overlaps (described below).

2.4 Gesture Annotation and Selection

For gesture, we utilized the SPUDNIG application (Ripperda et al. 2020) to detect all movements produced by each participant. These movements were then manually checked and non-gesture movements were removed. A second coder blind to the hypotheses manually identified gestures in 10% of the data (four minutes per participant). This subset of the data contained 12.7% of all gestures. Agreement reached 81% (Cohen's kappa could not be calculated as annotation categories were not included in this stage of the reliability testing). After removal of non-gesture movements we annotated each gesture as being representational (Alibali et al. 2001; e.g., iconics, metaphorics), pragmatic ((Kendon 2004); e.g., beats, emphatics, mood and stance modifiers), emblem (Ekman and Friesen 1969; e.g., "thumbs-up", "ok sign"), or interactive (Bavelas et al. 1995, e.g., palm-up open-hand gesture handing over the turn, addressee-directed deictics). In total we annotated 3,587 unique gestures (Representational: 980, Pragmatic: 627, Interactive: 1,897, Emblem: 83). Single gesture annotations could include multiple "beats" or movements, but were split when a continuous series of movements changed in either form or function. Annotations included only the preparation and main (stroke) phase of the gesture (Kita et al. 1998), but not the retraction. This was to ensure any abrupt retraction movements did not influence the speech-gesture alignment analyses. Reliability for gesture categorization was again performed on (a different set of) 10% of the data, which contained 18.5% of all gestures. Gestures identified by the primary coder were given to the second coder, without labels, and this second coder categorized each gesture. We observed a modified Cohen's kappa of 0.74, indicating substantial agreement (Landis and Koch 1977).

For each turn from speaker A with a corresponding next turn from speaker B (i.e., those for which turn transition times were calculated), we extracted any gestures that occurred during Speaker A's utterance. For these gestures, we calculated kinematic features, as described below.

2.5 Kinematic Feature Calculation

We calculated kinematic features that are related to visual salience and thus the communicative import of gestures. These features were the *peak velocity*, defined as the maximum velocity value during a gesture, number of *submovements*, defined as the number of individual ballistic movements (e.g., strokes, preparatory movements), *hold-time*, defined as the amount of time during gesture execution where the hands are still but depicting gestural information (i.e. excluding rest before initial movement and after final movement), *volume*, which is defined as the maximum geometric space utilized, *max distance,* which defines the maximum extension of the gesturing hand(s), and *McNeillian space*, which is based on McNeill's (1992) delineation of gesture space into center-center, center, periphery, and extra-periphery. Our McNeillian Space calculation is operationalized as the space category that is used most (i.e., the mode) during gesture production.

2.6 Analysis

All statistical tests were performed in R (R Core Team 2019) using the lme4 package (Bates et al. 2015), implementing mixed effects regression models. Models described below are performed separately for overlap data and gap data.

For all mixed models, we compared our model of interest, described below for each research question, against a null model that only contained the dependent variable, utterance duration as a covariate, and any random terms. Utterance duration was included as it has been demonstrated to affect subsequent turn-transition time (Roberts et al. 2015). This was done using chi-square tests of model of comparison.

To assess whether speaker visibility affected gap duration (RQ1), we tested a linear mixed model with transition time as dependent variable, and blur grade (i.e., visibility), together with utterance duration as independent variable. Dyad and participant were modeled as a nested random effect, with random slopes included when this did not lead to singular model fits. To assess whether the presence of gestures led to shorter turn transition times (RQ2), we first tested a model containing turn transition time as dependent variable, and gesture presence (together with utterance duration and blur grade) as independent variables. We additionally included an interaction term between blur grade and gesture presence, as well as a nested random effect of dyad and participant. As we were interested in how gesture visibility interacts with turn timing, we additionally tested whether a second-order polynomial (i.e., quadratic) model was a better fit than the linear model. As a final step, if gesture presence was a significant predictor of turn transition time, we tested this model against a 'gesture type model', in which switched out gesture presence as a predictor for gesture type as a predictor. Gesture type included the five types described above, as well as 'None', for the case of no gesture present. If this model was a better fit to the data, then this would indicate that turn transition time is differentially affected by different gesture types. To assess whether gesture kinematics influenced turn transition times (RQ3), we started with a maximal model including all kinematic features and their interaction with blur grade. Using chi-square tests, we systematically removed model terms until we found the best fit model of interest. This model was then compared against the null model. Due to differences in scale between kinematic features, all features were normalized (i.e., converted to a 0-1 scale) to facilitate model fitting.

3 Results

3.1 Effect of Visibility on Turn Transition Time

For total turn-transition time, we found a significant effect of reduced visibility leading to longer turn-transition time, $\chi 2(6) = 6.433$, $p = 0.011$. Splitting the data into gaps and overlaps, we found that decreasing visibility was associated with significantly higher inter-speaker overlaps, $\chi 2(3) = 120.58$, $p < 0.001$. However, visibility was not associated with any differences in inter-speaker turn gaps, $\chi 2(1) = 0.7663$, $p = 0.381$.

3.2 Effect of Gesture Presence on Turn Transition Time

For total turn-transition time, the best fit model included utterance duration, a quadratic effect of visibility, and gesture presence (χ^2 (2) = 23.157, $p < 0.001$; gesture type did not lead to a better model fit, $\chi^2(4) = 2.676$, $p = 0.614$). Figure 2A shows that, when including the non-linear effect of visibility, utterances with gestures seem to have more overlaps (i.e., faster responses; see Fig. 2A). However, the highly overlapping confidence intervals make it difficult to draw firm conclusions. Splitting the data into gaps and overlaps provides more insights into what is happening. When utterances were accompanied by co-speech gestures, we observed a reduction in subsequent inter-speaker overlaps, $\chi^2(1) = 4.608$, p = 0.032, as well as a reduction in inter-speaker turn gaps, χ^2 (7) = 57.446, p < 0.001 (see Fig. 2B and C). In other words, in the split analysis we see turn transition times moving towards zero when utterances are accompanied by a gesture. We additionally found that the quadratic model was a better fit than the linear model, χ^2 (2) = 46.539, $p < 0.001$, with gesture presence being associated with a reduction in overlaps during the first five grades of visibility, but falling off as visibility was reduced. We found a similar quadratic effect for gaps, $\chi^2(2) = 11.555$, $p = 0.003$, with gestures reducing gap duration primarily during the first 5 blur grades. See Fig. 2

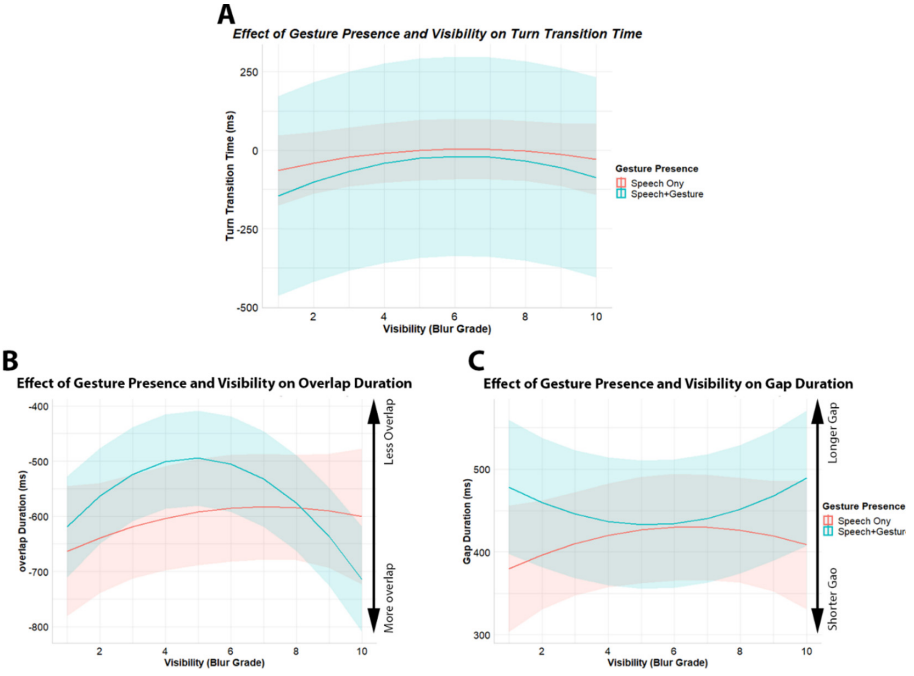

Fig. 2. Effects of gesture presence on turn transition times. Panel **A** depicts total turn transition time, panel **B** depicts overlaps, and panel **C** depicts gaps. Blur grade (i.e., visibility) is given on the x-axis. The y-axis gives overlap and gap duration. Speech only utterances are depicted by the red line, while speech-gesture utterances are depicted by the blue line. The shaded areas give the confidence interval. (Color figure online)

for an overview of the quadratic results. Similar to the total turn-transition time analysis, adding gesture type did not lead to a better model fit for overlaps, $\chi^2(4) = 3.772$, $p = 0.438$, or for gaps, $\chi^2(4) = 9.077$, $p = 0.059$.

3.3 Effect of Gesture Kinematics on Turn Transition Time

As the results of the gesture presence analysis indicated that the direction of association between gesture and turn transition time was dependent on whether we are considering gaps or overlaps, we focus our kinematic analyses on the split data. For utterances accompanied by gestures, we observed a significant association between the kinematic features of these gestures and both inter-speaker overlaps, $\chi^2(8) = 19.809$, $p = 0.011$, as well as inter-speaker gaps, $\chi^2(8) = 24.215$, $p = 0.002$. While the individual kinematic parameters were somewhat non-homogenous in their effects, we see at least several consistent findings. Gesture size, measured by maximum distance, as well as segmentation, measured by holdtime, both led to longer turn transition times. Peak velocity and submovements, on the other hand, led to shorter turn transition times.

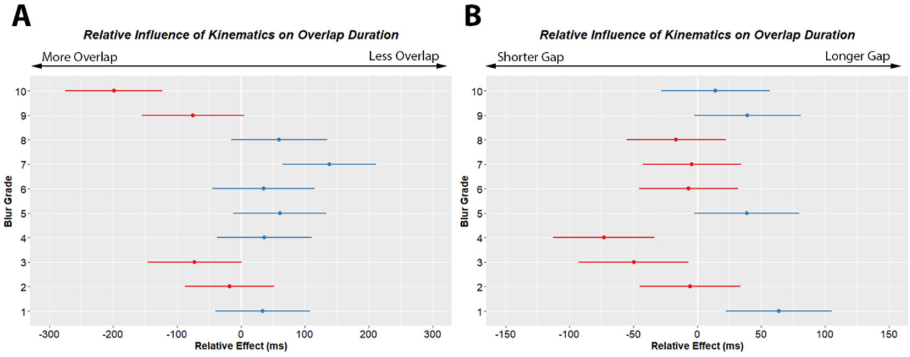

Fig. 3. Relative effects of gesture kinematics on turn transition times, over and beyond the main effects of visibility and gesture presence. The panel on the left depicts overlaps, while the panel on the right depicts gaps. Blur grade (i.e., visibility) is given on the y-axis. The x-axis gives net effect of all kinematic features relative to the intercept. In other words, positive values (given in blue) indicate longer turn transition times, while negative values (given in red) indicate shorter turn transition times. Mean effect, correcting for random effects and utterance duration, are indicated by circles, while the standard deviation is depicted as the lines behind the circles. Effect estimates are given for each of the blur grades. (Color figure online)

Overall, as can be seen in Fig. 3, the combination of these positive and negative effects is still able to explain a significant portion of the variation associated with reduced gaps and overlaps for utterances that are paired with gestures. Specifically, Fig. 3 shows how kinematics, as one synergistic set of features, can shift the extent of overlaps and gaps beyond the main effects of visibility and gesture presence. These main effects are thus the zero-point on the graphs, with the effect of increased kinematic salience or communicative import represented by the plotted values. Specific model terms can be seen in Appendix 1, and the effects on turn transition times across the visibility conditions can be seen in Fig. 3.

4 Discussion

We found that (1) turn transition time was significantly related to speaker visibility overall, 2) as well as to the presence of co-speech gestures (when they were visible) and 3) their kinematic properties. Examining the turn transition times across the spectrum of both gaps and overlaps, it seemed that decreased speaker visibility led to overall longer turn transition times. Looking at the presence of gestures and splitting the data into overlaps and gaps, we observed that the use of co-speech gestures was associated with a reduction in subsequent overlaps, but also in inter-speaker gaps. Furthermore, more kinematic salience and complexity correlated with a reduction in overlaps and gaps. This suggests that the way a gesture is performed may mediate a gesture's influence on upcoming turn transitions.

The finding of turn transition times being influenced by speaker visibility is at least partially in line with our hypotheses for RQ1. Specifically, that visual signals likely play a role in the timing of speaker turn transitions. Interestingly, decreased visibility was mostly associated with an increase in overlaps. This suggests that participants erred towards faster responses, at the risk of overlapping with the current speaker, rather than delaying their own speech onset. The increase in overlaps is in line with a previous report of increased overlaps in telephone conversations compared to face-to-face (ten Bosch et al. 2005). However, it is interesting to note that although turn transition times were affected in moderately reduced visibility, timing returned to the initial baseline (i.e., clear visibility) levels as visibility was removed completely. One possibility is that this difference arises because in ten Bosch et al. (2005), backchannels were not removed from the analysis, whereas we removed all short utterances from analysis. An alternative explanation is that, at least in computer-mediated interaction, full visibility versus fully blurred is not treated as face-to-face versus telephone in terms of turn-taking strategy. To put this another way, participants may employ a different strategy to coordinating turn-taking behavior when engaging in computer-mediated interaction. Gaining more insight into this issue will require future research.

When investigating the effect of co-speech gestures on the pooled turn-transition times, we found that utterances with gestures seemed to be followed by faster turn transitions. When splitting these analyses into gaps and overlaps, we specifically found more overlap in utterances without gestures. This finding would be in line with the findings of Holler et al. (2018) who found that questions accompanied by gestures get faster responses than those without a co-occurring gesture (Holler et al. 2018). However, splitting the data into gaps and overlaps shows a more nuanced picture: the presence of gestures is actually associated with a reduction in both overlaps and gaps. The reason for the pooled analysis showing a slightly different picture is likely because the effect of gesture on overlaps and gaps go in opposite directions, with the reduction in overlap being slightly stronger. This can be seen in the extent of the curves in Fig. 2, and could lead to a net effect, in the pooled data, of an overall increase in turn transition time. This finding provides more evidence for the idea that speakers may use a different strategy to coordinate turn-taking behavior during computer-mediated interaction. Rather than using the potential processing advantages of multimodal utterances to respond faster while overlapping with the other speaker (as they appeared to be doing in Holler et al. 2018), speakers in computer-mediated interaction seem to strive to minimize both gaps

and overlaps. As we observe in these data, the presence of gestures seems to facilitate this coordination strategy.

The non-linear effect of gesture presence on turn transition times across visibility conditions suggests that these effects are most pronounced in moderate visibility reductions. We interpret this as meaning that moderately coarse-grained visual information is playing a prominent role in turn timing. In other words, rather than fine-grained information in finger configurations or subtle movements, the larger movements and configurations of the hands likely influence the turn-transition times. At the same time, the fact that the gestural contribution to turn timing begins to decrease before visibility is completely removed suggests that this is not just a low-level effect of seeing any visual motion. Another important advancement from this study is that the interaction between visibility and gesture presence strongly suggests that the mechanism by which gestures support tighter turn-taking is indeed through the visual channel. In other words, if the turn-taking effect was mainly driven by prosodic changes induced by the biomechanical forces of gesture production (Pouw et al. 2020), then gesture presence should have led to tighter turn-taking, regardless of visibility.

The finding of gesture playing the strongest role in moderately affected visibility conditions is also in line with a study by Drijvers and colleagues who found that, when assessing the contribution of co-speech gestures to understanding degraded speech, gestures provide the greatest enhancement to perception when presented in moderately degraded speech compared to severely degraded or clear speech (Drijvers and Özyürek 2017). While highly speculative to suggest any similarity between these findings, it is interesting to note that co-speech gestures seem to be most beneficial in moderately disrupted communicative settings. Future studies will be needed to corroborate this idea. However, the current findings offer support for the notion that gestures facilitate the tight temporal coordination of turns, while showing that this is through an overall smoother transition, affecting both overlaps and gaps.

We additionally found that enhancement of gesture kinematics was associated with both overlap and gap durations, across visibility conditions. Importantly, the increase in some kinematic features may have contributed to temporally more coordinated turn transitions, while the reduction in others may have had the same effect. In other words, it is not a simple case of larger, more complex, or faster movements facilitating turn coordination. Instead, there seems to be a complex relationship between gesture kinematics and turn transition times. As subtle differences in gesture kinematics can signal communicative intent (Trujillo et al. 2018, 2020) and make the meaning of silent gestures easier to recognize (Trujillo et al. 2019), it may be that certain kinematic features make the overall content of the utterance easier to predict. This could contribute to the turn-end being easier to predict, leading to temporally more tightly coordinated transitions between speakers (Levinson 2016), or it may facilitate turn content prediction (Holler and Levinson 2019). Together with the interaction effect between visibility and gesture presence (discussed above), these results suggest that gestures support tight turn-timing through the spatio-temporal characteristics of the more coarse-grained movements of the arms and hands.

4.1 Relevance

The current findings provide new insights into the role of visibility, and more specifically, manual gestures, in how individuals coordinate turn-taking during computer-mediated interaction. Manual co-speech gestures may provide a way for an addressee, whether human or machine, to more easily process the meaning and intention behind an utterance, thus allowing this addressee to time their turn in a more coordinated manner. In terms of kinematic features, it may be that particular combinations of kinematic qualities lead to gestures that are easiest to process. We speculate, based on previous kinematic studies on gesture production and comprehension, that gestures should have an optimal balance of communicative salience and clear structure, without being so large, fast, or overly complex so as to slow down processing. Importantly, as gestures were associated with the greatest reduction in gaps and overlaps during moderate visibility conditions, it seems that fine-grained information, such as relating to subtle movements and configuration of the individual fingers, may be less necessary for maintaining the general coordination-facilitatory effect of gestures when compared to the larger, more coarse-grained information in the movements and configurations of the arms and hands. Overall, these results show that speakers may be partly employing a different strategy for turn-taking (especially relating to overlap) during computer-mediated interaction when compared to face-to-face or telephone conversations, but further investigations are needed in this domain.

4.2 Strengths and Limitations

This study utilized an experimental manipulation of speaker visibility, while also allowing unconstrained, naturalistic conversation. This design contributes to the ecological validity of our findings. However, the unconstrained nature of the experiment also provided many degrees of freedom for participant behavior. Our processing and modeling choices for the data were aimed at controlling for as much of this variability as was feasible. Without directly manipulating more task parameters, however, we cannot draw strong conclusions about the causal relations between variables. Instead, these findings should serve as ecologically valid observations about the associations between communicative signals, visual availability, and turn-taking performance. Future studies should use experimental manipulation to verify these findings and tease apart the complex web of associations that characterizes multimodal conversational behaviour.

4.3 Conclusions

Overall, our findings provide new insights into the role of gestures in rapid turn-taking in computer-mediated human interaction, particularly related to visibility. These findings suggest that both the presence and form of manual gestures contribute to the temporal coordination of speaker turn-transitions. More generally, our study demonstrates that computer-mediated interaction may in part elicit different turn-taking behavior compared to face-to-face or telephone-based interactions. At the same time, our findings highlight the resilience and flexibility of human communication in response to different, potentially difficult communicative contexts, as well as the core contributions of the visual modality.

Appendix 1

Parameter overview of kinematic analyses

Model parameter	*Parameter estimate**	*Standard deviation*	*t-value*
Overlaps			
Formula: turn_transition_time ~ utterance_duration + maxdist + peakvel*grade + MN_mode + submovements*grade + holdtime + volume + (1 | participant + (1 | utterance)			
(Intercept)	−442.20	3318.00	−0.13
utterance duration	0.00	0.00	−1.25
blur grade	898.10	331.60	2.71
maximum distance	184.80	126.70	1.46
peak velocity	−20.71	227.50	−0.09
McNeillian mode	49.83	40.27	1.24
submovements	−366.10	295.20	−1.24
holdtime	136.80	235.60	0.58
volume	−176.70	91.83	−1.92
peak velocity * grade	−60.62	22.12	−2.74
submovements * grade	88.97	39.00	2.28
Gaps			
Formula: turn_transition_time ~ utterance_duration + maxdist + peakvel + MN_mode * grade + submovements + holdtime* grade + volume + (1 | dyad/participant) + (1 | utterance)			
(Intercept)	1454.00	2044.00	0.71
utterance duration	0.00	0.00	−6.55
blur grade	−40.13	11.99	−3.35
maximum distance	79.07	81.79	0.97
peak velocity	−53.15	127.30	−0.42
McNeillian mode	−123.30	50.80	−2.43
Submovements	−58.75	120.50	−0.49
holdtime	224.10	220.20	1.02
volume	106.60	56.58	1.88
MCcNeillian Mode * grade	28.78	8.08	3.56
holdtime * grade	−34.44	31.19	−1.10

*Note that because kinematic values were normalized, parameter estimates cannot be directly interpreted in terms of effect on turn transition time, but rather should be interpreted in terms of effects relative to other kinematic features

References

Alibali, M.W., Heath, D.C., Myers, H.J.: Effects of visibility between speaker and listener on gesture production: some gestures are meant to be seen. J. Mem. Lang. **44**, 169–188 (2001)

Bates, D., Maechler, M., Bolker, B., Walker, S.: Fitting linear mixed-effects models using lme4. J. Stat. Softw. **67**(1), 1–48 (2015). https://doi.org/10.18637/jss.v067.i01

Bavelas, J.B., Chovil, N., Coates, L., Roe, L.: Gestures specialized for dialogue: Pers. Soc. Psychol. Bull. (1995). https://doi.org/10.1177/0146167295214010

Bavelas, J.B., Chovil, N.: Visible acts of meaning: an integrated message model of language in face-to-face dialogue. J. Lang. Soc. Psychol. **19**(2), 163–194 (2000). https://doi.org/10.1177/0261927X00019002001

ter Bekke, M., Drijvers, L., Holler, J.: The predictive potential of hand gestures during conversation: an investigation of the timing of gestures in relation to speech. (2020). https://doi.org/10.31234/osf.io/b5zq7

Boersma, P., Weenink, D.: Praat (Version 4.5. 25) (2007)

ten Bosch, L., Oostdijk, N., Boves, L.: On temporal aspects of turn taking in conversational dialogues. Speech Commun. **47**(1), 80–86 (2005). https://doi.org/10.1016/j.specom.2005.05.009

Campisi, E., Özyürek, A.: Iconicity as a communicative strategy: recipient design in multimodal demonstrations for adults and children. J. Pragmat. **47**(1), 14–27 (2013). https://doi.org/10.1016/j.pragma.2012.12.007

Drijvers, L., Özyürek, A.: Visual context enhanced: the joint contribution of iconic gestures and visible speech to degraded speech comprehension. J. Speech Lang. Hearing Res. **60**(1), 212–222 (2017). https://doi.org/10.1044/2016_JSLHR-H-16-0101

Gerwing, J., Bavelas, J.: Linguistic influences on gesture's form. Gesture **4**, 157–195 (2004)

Heldner, M., Edlund, J.: Pauses, gaps and overlaps in conversations. J. Phon. **38**(4), 555–568 (2010). https://doi.org/10.1016/j.wocn.2010.08.002

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

Holler, J., Levinson, S.C.: Multimodal language processing in human communication. Trends Cogn. Sci. **23**(8), 639–652 (2019). https://doi.org/10.1016/j.tics.2019.05.006

Holler, J., Wilkin, K.: An experimental investigation of how addressee feedback affects co-speech gestures accompanying speakers' responses. J. Pragmat. **43**(14), 3522–3536 (2011). https://doi.org/10.1016/j.pragma.2011.08.002

Jadoul, Y., Thompson, B., de Boer, B.: Introducing parselmouth: a python interface to praat. J. Phon. **71**, 1–15 (2018). https://doi.org/10.1016/j.wocn.2018.07.001

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

Kendon, A.: Gesture: Visible Action as Utterance. Cambridge University Press, Cambridge (2004)

Kita, S., van Gijn, I., van der Hulst, H.: Movement phases in signs and co-speech gestures, and their transcription by human coders. In: Wachsmuth, I., Fröhlich, M. (eds) Gesture and Sign Language in Human-Computer Interaction. GW 1997. Lecture Notes in Computer Science, vol. 1371, pp. 23–35. Springer, Berlin, Heidelberg (1998). https://doi.org/10.1007/BFb0052986

Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**, 159–174 (1977)

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

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

McNeill, D.: Hand and Mind: What Gestures Reveal about Thought. University of Chicago Press, Chicago (1992)

Pouw, W., Harrison, S.J., Dixon, J.A.: Gesture–speech physics: the biomechanical basis for the emergence of gesture–speech synchrony. J. Exp. Psychol. Gen. **149**(2), 391–404 (2020). https://doi.org/10.1037/xge0000646

R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing (2019). https://www.R-project.org/

Ripperda, J., Drijvers, L. Holler, J.: Speeding up the detection of non-iconic and iconic gestures (SPUDNIG): A toolkit for the automatic detection of hand movements and gestures in video data. Behav. Res. **52**, 1783–1794 (2020). https://doi.org/10.3758/s13428-020-01350-2

Roberts, S.G., Torreira, F., Levinson, S.C.: The effects of processing and sequence organization on the timing of turn taking: a corpus study, 509th edn, vol. 5. Frontiers in Psychology (2015). https://doi.org/10.3389/978-2-88919-825-2

Sacks, H., Schegloff, E. A., Jefferson, G.: A simplest systematics for the organization of turn-taking for conversation. **50**(40) (1974)

Stivers, T. et al.: Universals and cultural variation in turn-taking in conversation. PNAS. **106**, 10587–10592 (2009)

Trujillo, J.P., Simanova, I., Bekkering, H., Özyürek, A.: Communicative intent modulates production and comprehension of actions and gestures: a Kinect study. Cognition **180**, 38–51 (2018). https://doi.org/10.1016/j.cognition.2018.04.003

Trujillo, J.P., Simanova, I., Bekkering, H., Özyürek, A.: The communicative advantage: how kinematic signaling supports semantic comprehension. Psychol. Res. **84**(7), 1897–1911 (2019). https://doi.org/10.1007/s00426-019-01198-y

Trujillo, J.P., Simanova, I., Özyürek, A., Bekkering, H.: Seeing the unexpected: how brains read communicative intent through kinematics. Cereb. Cortex **30**(3), 1056–1067 (2020). https://doi.org/10.1093/cercor/bhz148