Head and hand movements during turn transitions: data-based multimodal analysis using the Frankfurt VR Gesture–Speech Alignment Corpus (FraGA)

Andy Lücking, Felix Voll, Daniel Rott, Alexander Henlein, Alexander Mehler
Goethe University Frankfurt
Text Technology Lab

Abstract

We introduce FraGA, a VR-based corpus of direction giving dialogues following the model of the SaGA corpus. The tracking data of FraGA are used to carry out multimodal computing: we look at turn transitions and re-evaluate findings from the literature on real-world dialogues and compare them with avatar-mediated VR dialogues. Interestingly, the established temporal overlap patterns of turns could not be replicated, and no significant amount of partner-directed head orientation (approximating gaze) was observed around turn transition points. However, the special status of hesitations is evidenced by the co-occurring head movements, but not by hand movements. In addition, we apply pink noise distribution fitting to the dialogue data, in particular to the hand movements. Here we find that hand movements indeed follow ¹⁄f fluctuations, a property of "meta-stable" dynamic systems.

1 Introduction

The interactive turn in cognitive science emphasizes face-to-face conversation (in the following called *interaction*) as the nucleus of cognition and communication (e.g., Spivey, 2023). At the same time, due to virtual reality (VR) technology, even face-to-face interactions can be technically mediated (e.g., in educational settings; Bagci et al., 2025). The question therefore arises as to what the differences are between human–human and VR augmented human–human dialogue? Interactions already have a comparatively long history of research in Conversation Analysis (Goodwin, 1981) and formal dialogue theory (Ginzburg et al., 2020), among others. In order to make empirical generalizations, interaction studies require conversational data. Data analyses are hampered by at least two difficulties: interactions can be quite complex, and interactions are usually multimodal (i.e., interlocutors use the full range of body signals in addition to speech), necessitating the integrated use of multiple analysis techniques (minimum for audio and video data) (Gregori et al., 2023). An early example of data analysis of a multimodal corpus focusing on speech and manual gestures is the *Speech and Gesture Alignment Corpus* (SaGA; Lücking et al., 2013), which consists of human–human direction dialogues. Accordingly, here we introduce an updated VR version of SaGA, namely FraGA (the ***Fra****nkfurt VR* ***G****esture–Speech* ***A****lignment Corpus*). FraGA is completely based on VR: the directions are given by the participants as avatars in a virtual environment, and speech, gaze and body behaviour are captured by the means of the tracking devices of the VR glasses—see section 3 for details. The raw data is accessible for largely automated analysis; we use it here to gain a better understanding of the interplay of turn taking, mutual attention and hand movements from a multimodal perspective. Our research questions and directions are outlined in section 2. In particular, we focus on a comparison of VR with *Real-World* (RW) dialogues. To this end, we replicate past studies and suggest a new methodology for quantitative conversation analysis, namely pink noise. FraGA is introduced in section 3. Results are presented in section 4 and discussed in section 5.

2 Research Questions and Topics

Conversation Analysis is beginning to abandon its old dogma that turns are primarily speech-based units and is recognizing that transition-relevance places are multimodal in nature (Duncan, 1974; Rohlfing et al., 2020; Kendrick et al., 2023; Mondada, 2007). Gaze takes on a central role: Kendon (1967) found that the speaker did not look at the a listener at the beginning of the utterance, but looked at the addressee at the end of the utterance in about 70 % of cases. However, gaze pattern differ across different types of interaction (Degutyte and Astell, 2021). Moreover, and in contrast to Kendon's results, the averted gaze of a virtual avatar is not interpreted as meaning that the avatar has not yet finished speaking (Gambi et al., 2015).

As earlier VR glasses did not offer eye tracking, the orientation of the head was often used as an approximation. (Murray and Roberts, 2006). This is of course not ideal (e.g., in high immersion conditions gaze is better for teamwork experience in collaborative games than head movement; Špakov et al., 2019). However, head movement seems to be sufficient for addressing and turn organisation in avatar-based interactions (Kothe et al., 2025), which is most relevant for the present study. Accordingly, for the time being we assume the approximation of gaze via head orientation and hope to be able to analyse gaze tracking data in the future (cf. section 6). We look at gaze respectively head orientation in avatar-mediated directions givings, a setting that is characterized by asymmetric roles (i.e., router and follower), and with task-oriented structure. In particular, we assess gazing patterns (i) at hesitation markers, (ii) at verbal turn transition points, and (iii) over the whole time course of a directions dialogue. Accordingly, the first research topic (**T**; we speak of research topic rather than research questions because there are no precise expectations regarding VR interactions yet) is

T1 Comparison of gaze/head orientation pattern between VR and RW.

From a dynamic systems perspective on human interaction (Dale et al., 2013), the fluctuation patterns of signals exchanged in dialogues are often "meta-stable" (Mayo and Gordon, 2020; Van Orden et al., 2011). Meta-stability is characterized by "pink noise", or $1/f$ distribution. Pink noise is a property of time series in between white noise and Brownian motion: white noise is completely random and future values cannot be predicted from past values, whereas Brownian motion is predictable from preceding values (Delignieres and Marmelat, 2012). $1/f$ fluctuation in data is characterized by the fact that the amplitude is inversely related to frequency. That is, fluctuation in data approximates $f^{-\alpha}$, where $\alpha \approx 1$. This means that pink noise, when transposed to a logarithmic axis, is a straight line with a slope of -1. $1/f$ scaling has been found, for instance, in the acoustic profile of repeated pronunciations of words (Kello et al., 2008). The study of Mayo and Gordon (2020) found that pink noise characterizes the synchrony of the gaze pattern of interlocutors. Less is known about the fluctuation of hand and arm movements, however. We hypothesize that it shows $1/f$ scaling, too. We compare the fluctuation of hand movements over the time course of a whole dialogue, and within turn transition time windows.

T2 Does multimodal behaviour data follow a pink noise pattern?

Moreover, speech–gesture production studies found that gestures facilitate word elicitation (Krauss and Hadar, 1999). Speech dysfluencies and hesitations in turn indicate that utterance production is faltering for some reason. Putting both observations together, we expect that increased hand movement occurs at the same time as or shortly after the dysfluency (indicated by transcription labels "[UH]" and "[UM]").

T3 Associations of verbal production difficulties and non-verbal behaviour.

Detailed studies in conversation analysis examined the temporal pattern observed in the verbal parts of turns, i.e. how quickly turn transitions occur in speech. A stable result is that the average time between turns is around 200 ms (Levinson and Torreira, 2015). Considering that utterance production and comprehension take a certain amount of time, this result can be understood as the fastest psycholinguistic threshold, *modulo* speech pauses, interruptions and predictive processing (Levinson, 2024). However, the results of previous studies were mainly gained in RW settings. Accordingly, the question arises whether they generalize to VR settings.

T4 Comparison of timing and turn-tranisition between VR and RW.

To address these research topics in a data-based way, appropriate VR data is needed. FRAGA collects such data in an avatar-mediated, VR-based setting resting on direction-giving dialogues.

3 Experimental Setting

In the FRAGA-experiments, two participants are engaged in a route description-giving task. One—called *Router* in the following—was sent on a virtual bus ride through a virtual town (fig. 1), which lasted approximately 7 minutes. Afterward, they met the other participant—called *Follower*—in a "living room" and described the route through the virtual town. The route descriptions involve giving directions and describing salient buildings or features of the landscape—so-called *landmarks*—, as studied in the psychology of route descriptions (Denis, 1997) (see also fig. 1). Accordingly, the paradigm is well-known for eliciting gestures (cf. the original SaGA corpus; Lücking et al., 2010). After the directions dialogue, the Follower has to find the way through the virtual town (in contrast to SaGA, where this step was skipped). After the experiment, additional data was collected using a small questionnaire, such as demographic data, primary languages, dominant hand, acquaintance with the interlocutor, and previous experiences with VR.

The entire experimental setup was implemented using Va.Si.Li-Lab (*VR-Lab for Simulation-based Learning*; Mehler et al., 2023), a VR platform that supports immersive simulation-based learning by using a graph-theoretic model that manages multimodal data and facilitates dynamic event processing in adaptive virtual environments (see (Abrami et al., 2023) for a comparison of Va.Si.Li-Lab with related approaches). As a multi-user collaborative tool, Va.Si.Li-Lab enables detailed exploration of virtual spaces through features such as multi-user annotation (Bagci et al., 2025) and real-time environment modification (Abrami et al., 2024). The participants are represented by Meta Avatars[1] that reflects the tracked hand, face and eye data, thus enabling multimodal interactions in VR. The glasses we used were the Meta Quest Pro, which not only cover the hands, but also the lower and upper face and eyes. Va.Si.Li-Lab tracks all data locally on the glasses and then sends it to our database. Both the local timestamps of the glasses and the timestamps when they reach the server are stored to enable the most accurate tracking possible. A third-person view on an avatar-mediated directions dialogue is shown in fig. 2.

In the case of FraGA, we experiment with an instantiation of Va.Si.Li-Lab which includes the virtual landscape of FraGA alongside a virtual meeting room in which the probands can interact with each other. Both participants meet in the virtual meeting room, and the Router explains the way through the city to the Follower.

4 Data-based Analysis

In the following, the research questions (see section 2) are addressed. We look first at verbal turn overlap. We then analyze gaze approximated by head orientation. Finally, pink noise fluctuation is addressed.

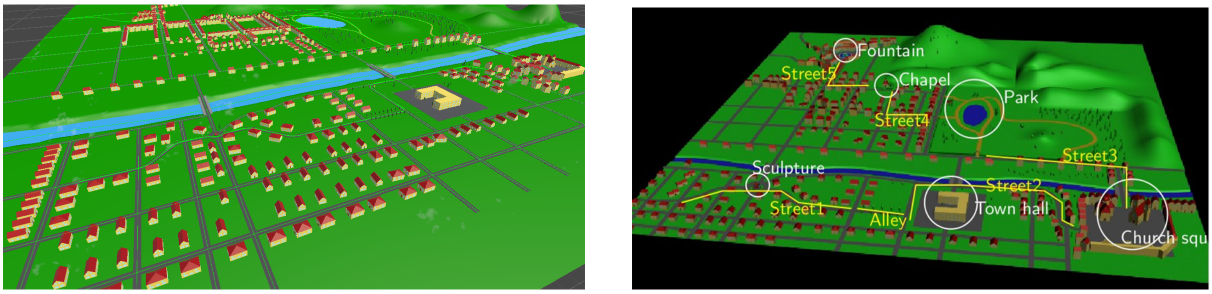
All audio recordings from the experiment were transcribed using CrisperWhisper (Zusag et al., 2024), a variant of OpenAI's Whisper (Radford et al., 2023), which provides more precise word-level timestamps and preserves disfluencies rather than omitting them. Since each participant was recorded using an individual headset equipped with its own microphone, no additional speaker diarization was required. 73 dialogues involving 146 speakers (73 Router and 73 Follower) have been analyzed so far.[2] Demographic information about the participants is collected in table 1. The speaking times and number of word tokens of the 73 dialogues of FraGA are summarized in table 2.

Age	
<18	2
18–20	23
21–23	72
24–26	26
27–29	13
≥ 30	10
Gender	
Female	39
Male	102
Diverse	1
No response	4
Dominant Hand	
Right	133
Left	12
Not applicable	1
VR Experience (1 = low . . . 5 = high)	
1	86
2	42
3	11
4	6
5	1
Acquaintance (1 = low . . . 5 = high)	
1	43
2	10
3	3
4	3
5	87

Table 1: Demographic characteristics of participants

[1] https://developers.meta.com/horizon/documentation/unity/meta-avatars-overview/

[2] There are actually more recorded dialogues. However, due to recording problems – especially WLAN interruptions that lead to a loss of synchronization –, they have to be corrected manually, which is still being worked on.



(a) Panorama view of FRAGA

(b) Route and landmarks in bird's eye view (taken from Lücking et al. 2010)

Figure 1: Experimental stimulus: Virtual town

Figure 2: Directions dialog between two participants as avatars, third-person perspective

	Speaking time	# Tokens
total:	12:44:37	92,923
Router:	8:17:19	70,517
Follower:	4:27:18	22,406
Avg. Router:	0:06:49	1,273
Avg. Follower:	0:03:40	966
Avg. Dialogue:	0:10:28	307

Table 2: Speaking times (hours:minutes:seconds) and number of tokens

4.1 Turn overlap (T4)

We calculated the temporal relation between adjacent utterances of different speakers. For that we take the transcribed audio from all participants, add time ranges for each utterance and combine transcriptions from Person 1 and 2 into a dialogue format, ordered by starting time of their utterances. All dialogue files are then processed, iterating over each utterance and checking if a switch from Person 1 to Person 2 or vice versa occurred. If so, the end time of the previous utterance is compared to the starting time of the following and the difference is saved if it falls within a window of −2500 to 2500 milliseconds. The 200.00 ms turn break time can not be confirmed – see fig. 3. Surprisingly, we found that overlap is the rule. An average overlap of −214.90 ms is observed, with a minimum value of −2,498.00 ms and a maximum value of 2,491.00 ms within the chosen window. This result is certainly due to the VR environment and the audio transmission.

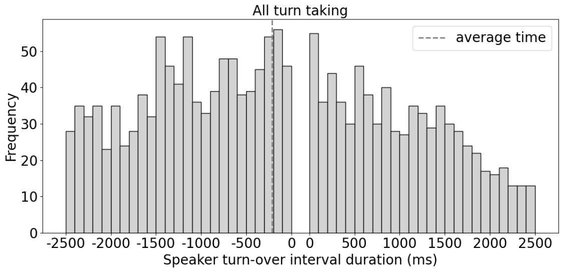


Figure 3: Time gap between adjacent utterances of different speakers.

Does the pattern emerge regardless of the sentence types at turn transition? To answer this question, we have divided utterances in question–answer pairs (i.e., pairs of utterances where speaker 1's utterance ends with a question mark and the following utterance of speaker 2 ends with a period), and others. Here an average value of −234.90 ms was observed, with a minimum value of −2,493.00 ms and a maximum value of 2,259.00 ms The results are shown in fig. 4, and confirm overlap as a rule, but with more variance.

4.2 Hesitations (T3)

If the production of at least some gestures is bound up with difficulties in word elicitation (cf. section 2), then we expect to find increased hand movement around verbal hesitation markers. Hesitations are tagged as "[UH] or [UM]" tokens and are part of the CripserWhisper transcription. We assessed the amount of movement within a time window of 1 second before and following these tokens. We

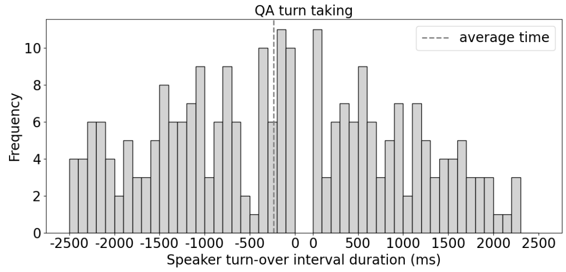


Figure 4: Time gap difference between syntactically determined question–answer pairs and other pairs of utterance

compared it to (i) the average amount of movement over the span of the whole dialogue, and (ii) the close temporal context. The amount of movemvent was assessed in terms of distances, where distances were computed for each subject by first calculating the Euclidean length of the frame-to-frame difference in three-dimensional hand coordinates. These individual step lengths were then summed to obtain the total distance traveled within each segment, and the corresponding number of steps was recorded. The mean step length for a given segment was defined as the total distance divided by the number of steps. Finally, the aggregated mean-step value across all subjects is calculated. Figure 5 shows the mean step distance (hand movement) per frame (± *standard error of mean*, SEM) for both left and right hands during the three token-aligned phases "Before," (i.e., start of hesitation token until minus 1 sec.) "During," (duration of token) and "After" (end of token until plus 1 sec.). Each bar shows the average distance per frame (±SEM) in that phase of the corresponding phase, while the two dashed horizontal lines mark the mean frame distance for each hand during the "During" phase of the dialogue timestamps (the reference baselines, nearly identical for both hands). Within every phase, left and right hands (see table 3) did not differ significantly (paired t-tests all non-significant), indicating that both hands moved to a similar extent around hesitation markers (p-value: Before: 0.78, During: 0.72, After: 0.30). Comparing "Before" and "After" also revealed no significant change for either hand (p-value: Left hand: 0.21, Right hand: 0.23). Finally, none of the three token-aligned phases differed significantly from the dialogue-based reference lines (p-values between 0.17 and 0.80), indicating that hand motion around tokens neither exceeds nor falls below the dialogue baseline. For the individual statistical charateristics see table 4.

Phase	**Left Hand**	**Right Hand**
Before	0.063748	0.063482
During	0.042608	0.042137
After	0.025926	0.027156

Table 3: Means for Token Bars

Test	Paired t-Tests	
N	146	
df (degrees of freedom)	145	
Left vs Right		
Before	$t(145) = 0.28$	$p = 0.78$
During	$t(145) = 0.36$	$p = 0.72$
After	$t(145) = -1.04$	$p = 0.30$
Before vs After		
Left Hand	$t(145) = 1.26$	$p = 0.21$
Right Hand	$t(145) = 1.21$	$p = 0.23$
Dialogue During Refline vs Token Phases		
Left, before	$t(145) = -1.29$	$p = 0.20$
Left, during	$t(145) = -1.37$	$p = 0.17$
Left, after	$t(145) = -0.25$	$p = 0.80$
Right, before	$t(145) = -1.24$	$p = 0.22$
Right, during	$t(145) = -1.24$	$p = 0.22$
Right, after	$t(145) = -0.27$	$p = 0.79$

Table 4: Statistical test results on hand movement (paired t-tests, $N = 146$, df=145)

Additionally, we looked at the amount of partner-directed gaze at hesitations. Gaze direction was approximated by head orientation. We found that partner-directed head orientation is likely to be avoided when production difficulties arise—see the comparison in fig. 6. Moreover, no difference in the amount of partner-directed gaze for different types of speech acts is found (when assigning all turns to either assertion or question).

Is there a difference in the amount of partner-directed gaze at turn taking intervals compared to mid-turn gaze? To answer this question, we compared head orientation at the ±500.00 ms time interval at the beginning and the end of an utterance ("Sentence Start" and "Sentence End", respectively) with head orientation during the time course of the utterance ("Sentence Mid"), see fig. 7. While a difference in head orientation patterns at turn-sensitive points is to be expected (see section 2), no significant difference in head orientations could be found.

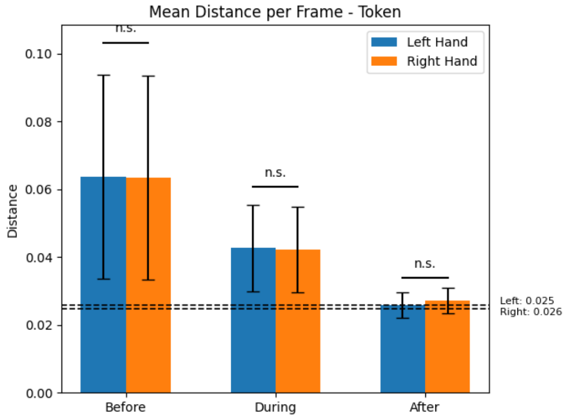


Figure 5: Average hand movement per frame for the left and right hand in the "Before", "During" and "After" phases of hesitation markers. The bars show the average distance per frame (±SEM), and the two overlapping dashed lines show the average distance of each hand during the "During" phase of the dialog timestamps.

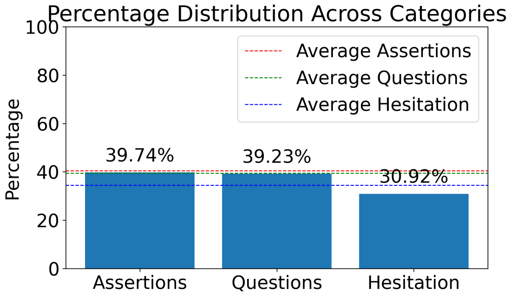


Figure 6: Average amount of time speaker looks at conversation partner during assertions, questions and hesitations. Bars represent weighted average, while horizontal lines indicate unweighted average.

4.3 Pink noise (T2)

Pink noise fluctuation has been observed in various data sources. Here we assess whether it also describes movement, where movement is hand movement tracked at wrist position. The most frequently used method for detecting $1/f$ scalings is *Detrended Fluctuation Analysis* (DFA; Peng et al., 1994). DFA returns a value for exponent α that characterizes pink noise when it ranges from 0.75 to 1.25 (Delignieres and Marmelat, 2012, p. 4). We carried out DFA by means of Python's *nolds* library (Schölzel, 2019).

However, due to gaps or insufficient length in the timestamp data, which are essential for DFA, the following analysis includes only 59 participants. In our pipeline, each individual timestamp is treated as a separate analysis unit. We make use of two different timestamps: (i) the time intervals defined by the utterances of a dialog (called *Dialog Timestamps* in the following), (ii) the time intervals taken only from hesitation marker tokens (*Hesitation Timestamps*). For a given timestamp, we first extract the segment of hand-tracking data corresponding to that timestamp. The three-dimensional positional coordinates (x,y,z) within that segment are converted into a one-dimensional series of frame-to-frame step lengths (the Euclidean distance between consecutive 3D points). DFA is then applied to this step-length series to compute the α-coefficient for that specific timestamp. By repeating this process for all timestamps associated with a given participant and a given hand (left or right), we obtain a list of α-values (one per timestamp) for each participant-hand combination—see fig. 8. We then compute the mean α across all valid timestamps for each participant–hand combination. Finally, to characterize group-level behavior, we aggregate these per-participant means across the entire sample: the mean (α_μ) and the median (α) are computed separately for left and right hands across all participants. In other words, each timestamp contributes one α-estimate per hand; these are averaged within participants to yield a single α per participant per hand; and those participant-level α's are in turn summarized by their overall mean and median, which are shown in table 5.

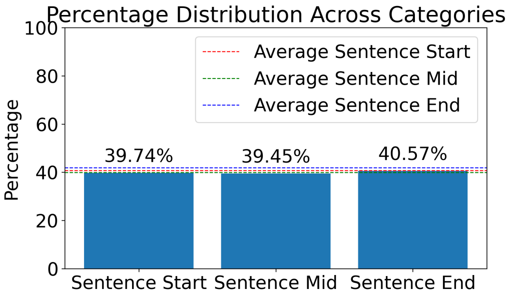
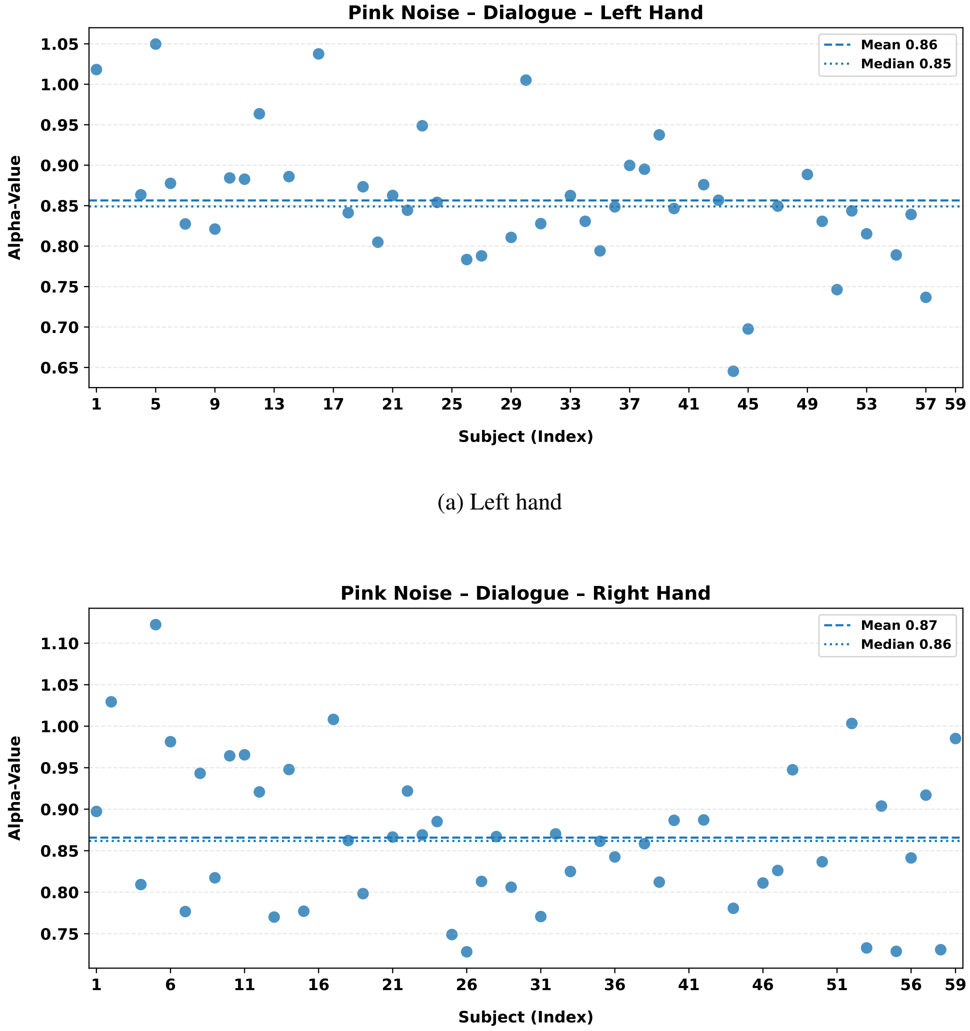


Figure 7: Average amount of time speaker looks at conversation partner during start, mid and end part of his sentence. Bars represent weighted average, while horizontal lines indicate unweighted average.

4.4 Head orientation as gaze approximation (T1)

As can be observed from fig. 6 and fig. 7, the head of one interlocutor is oriented towards the dialogue partner in only about one third of the time. Where do interlocutors look at the remaining time? To answer this question, we plotted the projections of



(a) Left hand

(b) Right hand

Figure 8: Pink noise hand movement, dialogue timestamps; x-axis labels show the index of the participants

Hand	Left	Right
Dialog Timestamps		
Mean	0.856	0.865
Median	0.848	0.859
Hesitation Timestamps		
Mean	0.938	0.945
Median	0.888	0.979

Table 5: Pink Noise: Averaged alpha coefficients for dialogue timestamps and hesitation timestamps (see main text for explanation)

head orientations of each individual interlocutor on a sphere surrounding the interlocutors. Two examples are shown in 9. These graphics are slightly less accurate than our calculations for the percentage values, since we need to assume a fixed head and body position for both persons during the entire conversation. Fixed positions are obtained by calculating the median value from all tracked position coordinate points for the respective part. Once fixed values are obtained, we calculate the view rays from the fixed head position and place the rectangle for body simulation onto the fixed body position point. Intersections with the rectangle are marked as red dots inside the rectangle while all view rays, regardless of intersecting with the rectangle, are intersected with the sphere. The sphere has the fixed head position as center and its radius is set to fully include the rectangle inside the sphere. The intersection points are again marked onto the sphere, containing a heatmap color scheme, showing the density of frequently viewed areas. Lastly an average view ray is calculated from all received head rotations and drawn originating from head position and with the length of the sphere's radius. The resulting projections show a high density area for interlocutors' view directions at approximately head level of their partner, with point distributions around that area. Very few points are measured where view directions go in completely different directions. This would speak for a higher percentage value of on-looking time, but we observe various plots with slight shifts of these cluster to the right left or above of the other interlocutor. Such variations would explain the the lower percentage values and could stem from only using the head rotation value to calculate view directions instead of integrating the eye tracking data. The further integration of eye-tracking data and the fine-tuning of the calculation method for gaze direction should lead to more detailed results with possibly longer on-looking times and is the subject of current work.

Brief summary:

T1 Participants look at each other about ⅓ of the time. There is no difference of partner-orientation for turn transitions, but there is for hesitations.

T2 Hand movements exhibit ¹⁄f fluctuations.

T3 Hesitation is associated with fewer partner-directed head movements, but has no effect on hand movements.

T4 An average turn overlap of $-214.90\,\text{ms}$ is observed.

5 Discussion

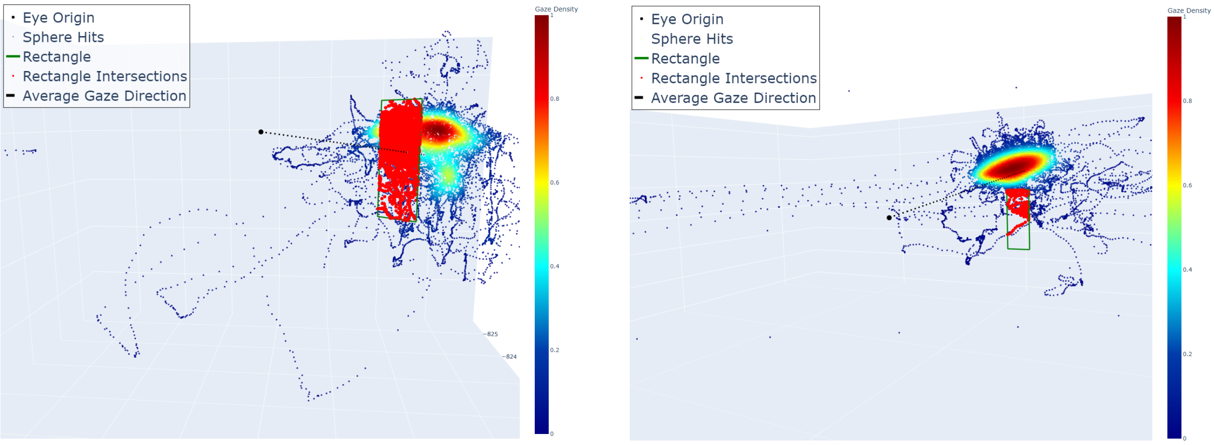
The analyses reported in section 4 show a couple of differences of VR-mediated FraGA dialogues and what we would expect from RW dialogues. Why is that?

Perhaps the most surprising result is that, on average, overlap in speech is the rule for turn transitions, not the exception. This observation is probably due to the asymmetry of the type of dialogue (direction givings). One participant, the Router, is the main speaker, because s/he gives the follower directions (cf. the proportions of speaking times and number of tokens in table 2). This means that the Follower in general has to interrupt the Router if s/he wants to take the turn, for instance, for requesting more information. Interruptions are naturally accompanied by an overlap of speeches (which is part of the nature of interrupting).

The finding that partner-directed gaze is avoided during hesitations can most likely be explained by findings on cognitive processing load and gaze aversion. Monitoring faces is an informationally rich, costly process (Doherty-Sneddon et al., 2001). This means that face monitoring through gaze ties up cognitive resources which are needed for word elicitation, especially when production difficulties occur. Therefore, averting the gaze is a strategy of avoiding cognitive overload (Glenberg et al., 1998).

Until now, gaze is approximated by head orientation, which is a simplification. However, the results so far seem to indicate that the gaze behavior between VR and RW is actually quite different. On the one hand, in VR no difference in head orientation across turns, turn transitions and conversational act types has been found. On the other hand, there are dialogue instances where nearly no head-approximated gaze is directed to the interlocutor. Both results deviate from the previous findings on RW (see section 2). Accordingly, there seems to be a rather different perception of the VR interaction situation compared to RW. We can only speculate about the reasons. It is conceivable that participants do not take avatar gaze as trustworthy as human gaze. As a consequence, partner monitoring is taken less seriously, in the extreme nearly fully avoided. It is also possible that the results are slightly biased by the precision of the tracking method and its visualization in avatars. That is, the eye direction displayed in the avatar might be slightly distorted. Given the very subtle system of gaze and mutual gaze, even small deviations can irritate VR users and affect head and eye movements.

Hand movement data are shown to follow ¹⁄f fluctuations. The fitting is nearly perfect when data are aggregated on the temporal windows defined by sentence tokens. It is still not well-understood what causes pink noise distributions (Delignieres and Marmelat, 2012; Van Orden et al., 2011). However, ¹⁄f scalings indicate a property of a non-random, complex systems. It characterizes, for instance, many kinds of physiological data (Sejdić and Lipsitz, 2013). Now hand-movements, including hand and arm gestures, are bodily actions, too. Therefore, some degree of "pinkiness" can be expected from this anatomical fact. In addition, hand movements that are part of gestures are subject to semantic constraints. For instance, a gesture occurrence usually "coheres" with an expression from speech, the affiliate (Schegloff, 1984), which mostly is a lexical item (Mehler and Lücking, 2012). A gesture and its affiliate normally exhibit a tight temporal



(a) "Good" example: interlocutor is main focus

(b) "Bad" example: main focus is slightly ahead of interlocutor

Figure 9: Three dimensional representation of tracked eye data during the entire conversation. Tracking points are projected onto a sphere around the average head position of the speaker, while the body of the conversation partner is represented as a two dimensional rectangle at his average body position.

relationship (McNeill, 1992). That is, affiliation imposes additional systematicity in movement data, which is probably captured by the token-based aggregation of the DFA and which could explain why the scaling coefficient is very close to 1.

The vast majority of the participants had little or no VR experience, cf. table 1. Accordingly, it cannot be ruled out that a certain degree of deviation from the RW results is simply due to the fact that the participants behaved hesitantly and had to adapt to the VR environment. We will be able to say more about this when the experiments are repeated with new participants, who are generally becoming more and more familiar with VR in their daily lives or as part of their studies.

On a more abstract level, FRAGA shows the usefulness of VR-based approches for multimodal computing. Here, we exemplified this by means of speech-to-text transcriptions (with token times, speaker diarization, and hesitation detection), hand movement data, and head orientations. Such analyses will improve with the tools and methods that are available for multimodality research (Henlein et al., 2024). However, FRAGA also shows that dialogs in the real world differ in important aspects from avatar-mediated VR dialogs. Therefore, the methodology of VR and the insights gained from it cannot simply be transferred to RW. Rather, FRAGA can be seen to address a field of communication research *sui generis*, namely avatar-based communication (e.g., Wei et al., 2024). Given the need for social immersive remote communication (roughly, from letters to phone and video calls to avatars), it is necessary to investigate and understand the characteristics of VR interactions and how they compare to RW interactions.

6 Outlook

We introduced FRAGA and first multimodal analyses using this resource. FRAGA is a dataset on human–human interaction in VR environments using VR glasses. A feature of the VR–based approach is the integrated tracking of different signals. Developing and using more automatized processing methods (e.g., gesture recognition, facial expression recognition, eye gaze tracking data and mutual gaze detection) is ongoing work. In this context, an anonymized version of FRAGA is in preparation and will be released to the public.

Acknowledgments

We thank two anonymous reviewers for their helpful feedback that improved the paper. AH, AL, and AM would like to thank the *German Research Foundation* (DFG, grant number 502018965) for its support.

References

Giuseppe Abrami, Alexander Mehler, Mevlüt Bagci, Patrick Schrottenbacher, Alexander Henlein, Christian Spiekermann, Juliane Engel, and Jakob Schreiber. 2023. Va.Si.Li-Lab as a collaborative multi-user annotation tool in virtual reality and its potential fields of application. In *Proceedings of the 34th ACM Conference on Hypertext and Social Media*, HT'23, pages 1–9, New York, NY, USA. Association for Computing Machinery.

Giuseppe Abrami, Dominik Alexander Wontke, Gurpreet Singh, and Alexander Mehler. 2024. Va.Si.Li-ES: VR-based dynamic event processing, environment change and user feedback in Va.Si.Li-Lab. In *Proceedings of the 35th ACM Conference on Hypertext and Social Media*, HT '24, pages 357—-368, New York, NY, USA. Association for Computing Machinery.

Mevlüt Bagci, Alexander Mehler, Giuseppe Abrami, Patrick Schrottenbacher, Christian Spiekermann, Maxim Konca, Jakob Schreiber, Kevin Saukel, Marc Quintino, and Juliane Engel. 2025. Simulation-based learning in virtual reality: Three use cases from social science and technological foundations in terms of Va.Si.Li-Lab. *Technology, Knowledge and Learning*.

Rick Dale, Riccardo Fusaroli, Nicholas D. Duran, and Daniel C. Richardson. 2013. The self-organization of human interaction. In Brian H. Ross, editor, *Psychology of Learning and Motivation*, number 59 in Psychology of Learning and Motivation, chapter 2, pages 43–95. Academic Press.

Ziedune Degutyte and Arlene Astell. 2021. The role of eye gaze in regulating turn taking in conversations: A systematized review of methods and findings. *Frontiers in Psychology*, 12.

Didier Delignieres and Vivien Marmelat. 2012. Fractal fluctuations and complexity: Current debates and future challenges. *Critical Reviews in Biomedical Engineering*, 40(6):485–500.

Michel Denis. 1997. The description of routes: A cognitive approach to the production of spatial discourse. *Current Psychology of Cognition*, 16:409–458.

Gwyneth Doherty-Sneddon, Lesley Bonner, and Vicki Bruce. 2001. Cognitive demands of face monitoring: Evidence for visuospatial overload. *Memory & Cognition*, 29(7):909–919.

Starkey Duncan. 1974. On the structure of speaker–auditor interaction during speaking turns. *Language in Society*, 3(2):161–180.

Chiara Gambi, Torsten Kai Jackmann, and Maria Staudte. 2015. The role of prosody and gaze in turn-end anticipation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, pages 764–769.

Jonathan Ginzburg, Chiara Mazzocconi, and Ye Tian. 2020. Laughter as language. *Glossa*, 5(1):104.

Arthur M. Glenberg, Jennifer L. Schroeder, and David A. Robertson. 1998. Averting the gaze disengages the environment and facilitates remembering. *Memory & Cognition*, 26(4):651–658.

Charles Goodwin. 1981. *Conversational Organization: Interaction between Speakers and Hearers*. Academic Press, New York.

Alina Gregori, Federica Amici, Ingmar Brilmayer, Aleksandra Ćwiek, Lennart Fritzsche, Susanne Fuchs, Alexander Henlein, Oliver Herbort, Frank Kügler, Jens Lemanski, Katja Liebal, Andy Lücking, Alexander Mehler, Kim Tien Nguyen, Wim Pouw, Pilar Prieto, Patrick Louis Rohrer, Paula G. Sánchez-Ramón, Martin Schulte-Rüther, Petra B. Schumacher, Stefan R. Schweinberger, Volker Struckmeier, Patrick C. Trettenbrein, and Celina I. von Eiff. 2023. A roadmap for technological innovation in multimodal communication research. In *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management.*, number 14029 in HCII 2023. Lecture Notes in Computer Science, pages 402–438, Cham. Springer.

Alexander Henlein, Anastasia Bauer, Reetu Bhattacharjee, Aleksandra Ćwiek, Alina Gregori, Frank Kügler, Jens Lemanski, Andy Lücking, Alexander Mehler, Pilar Prieto, Paula G. Sánchez-Ramón, Job Schepens, Martin Schulte-Rüther, Stefan R. Schweinberger, and Celina I. von Eiff. 2024. An outlook for AI innovation in multimodal communication research. In *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management.*, HCII 2024. Lecture Notes in Computer Science, pages 182–234, Cham. Springer Nature Switzerland.

Christopher T. Kello, Gregory G. Anderson, John G. Holden, and Guy C. Van Orden. 2008. The pervasiveness of 1/f scaling in speech reflects the metastable basis of cognition. *Cognitive Science*, 32(7):1217–1231.

Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26(1):22–63.

Kobin H. Kendrick, Judith Holler, and Stephen C. Levinson. 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.

Angelika Kothe, Volker Hohmann, and Giso Grimm. 2025. Effect of avatar head movement on communication behaviour, experience of presence and conversation success in triadic conversations. *Preprint*, arXiv:2504.20844.

Robert M. Krauss and Uri Hadar. 1999. The role of speech-related arm/hand gestures in word retrieval. In Ruth Campbell and Lynn S. Messing, editors, *Gesture, speech, and sign*, pages 93–116. Oxford University Press, Oxford.

Stephen C. Levinson. 2024. *The Dark Matter of Pragmatics: Known Unknowns*. Elements in Pragmatics. Cambridge University Press, Cambridge.

Stephen C. Levinson and Francisco Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6:731.

Andy Lücking, Kirsten Bergman, Florian Hahn, Stefan Kopp, and Hannes Rieser. 2013. Data-based analysis of speech and gesture: The Bielefeld Speech and Gesture Alignment Corpus (SaGA) and its applications. *Journal on Multimodal User Interfaces*, 7(1-2):5–18.

Andy Lücking, Kirsten Bergmann, Florian Hahn, Stefan Kopp, and Hannes Rieser. 2010. The Bielefeld speech and gesture alignment corpus (SaGA). In *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, LREC 2010, pages 92–98, Malta. 7th International Conference for Language Resources and Evaluation.

Oded Mayo and Ilanit Gordon. 2020. In and out of synchrony—behavioral and physiological dynamics of dyadic interpersonal coordination. *Psychophysiology*, 57:e13574.

David McNeill. 1992. *Hand and Mind*. Chicago University Press, Chicago.

Alexander Mehler, Mevlüt Bagci, Alexander Henlein, Giuseppe Abrami, Christian Spiekermann, Patrick Schrottenbacher, Maxim Konca, Andy Lücking, Juliane Engel, Marc Quintino, Jakob Schreiber, Kevin Saukel, and Olga Zlatkin-Troitschanskaia. 2023. A multimodal data model for simulation-based learning with Va.Si.Li-Lab. In *Proceedings of HCI International 2023*, Lecture Notes in Computer Science, pages 539–565. Springer. Accpeted.

Alexander Mehler and Andy Lücking. 2012. Pathways of alignment between gesture and speech: Assessing information transmission in multimodal ensembles. In *Proceedings of the International Workshop on Formal and Computational Approaches to Multimodal Communication under the auspices of ESSLLI 2012, Opole, Poland, 6-10 August*.

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

Norman Murray and Dave Roberts. 2006. Comparison of head gaze and head and eye gaze within an immersive environment. In *2006 Tenth IEEE International Symposium on Distributed Simulation and Real-Time Applications*, pages 70–76.

C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger. 1994. Mosaic organization of DNA nucleotides. *Phys. Rev. E*, 49(2):1685–1689.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Katharina J. Rohlfing, Giuseppe Leonardi, Iris Nomikou, Joanna Rączaszek-Leonardi, and Eyke Hüllermeier. 2020. Multimodal turn-taking: Motivations, methodological challenges, and novel approaches. *IEEE Transactions on Cognitive and Developmental Systems*, 12(2):260–271.

Emanuel A. Schegloff. 1984. On some gestures' relation to talk. In J. Maxwell Atkinson and John Heritage, editors, *Structures of Social Action. Studies in Conversational Analysis*, Studies in Emotion and Social Interaction, chapter 12, pages 266–296. Cambridge University Press, Cambridge, MA.

Christopher Schölzel. 2019. Nonlinear measures for dynamical systems.

Ervin Sejdić and Lewis A. Lipsitz. 2013. Necessity of noise in physiology and medicine. *Computer Methods and Programs in Biomedicine*, 111(2):459–470.

Michael J. Spivey. 2023. Cognitive science progresses toward interactive frameworks. *Topics in Cognitive Science*, 15(2):219–254.

Guy C. Van Orden, Heidi Kloos, and Sebastian Wallot. 2011. Living in the pink: Intentionality, wellbeing, and complexity. In Cliff Hooker, editor, *Handbook of the Philosophy of Science: Philosophy of Complex Systems*, 10. Elsevier.

Oleg Špakov, Howell Istance, Kari-Jouko Räihä, Tiia Viitanen, and Harri Siirtola. 2019. Eye gaze and head gaze in collaborative games. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, ETRA '19, New York, NY, USA. Association for Computing Machinery.

Xiaoying Wei, Xiaofu Jin, and Mingming Fan. 2024. Communication in immersive social virtual reality: A systematic review of 10 years' studies. In *Proceedings of the Tenth International Symposium of Chinese CHI*, Chinese CHI '22, page 27–37, New York, NY, USA. Association for Computing Machinery.

Mario Zusag, Laurin Wagner, and Bernhad Thallinger. 2024. Crisperwhisper: Accurate timestamps on verbatim speech transcriptions. In *Proc. Interspeech 2024*, pages 1265–1269.