Evaluating Study Design and Strategies for Mitigating the Impact of Hand Tracking Loss

Ylva Ferstl
Facebook Reality Labs
Trinity College Dublin
USA; Ireland

Rachel McDonnell
Trinity College Dublin
Ireland

Michael Neff
Facebook Reality Labs
University of California, Davis
USA

Figure 1: Sample frames from the experiment. (1) The male character in the FK condition, and (2) the same frame in the IK condition and (3) in the IK *full error* condition where hand tracking was lost for the character's right hand which subsequently remains hanging at shoulder height. (4) The female character in the IK *hybrid* condition: When tracking for the hand is lost around waist height (character's right hand), the hand will be moved to a rest position. (5) The character's right hand moved to a rest position. (6) Both hands moved to a rest position.

ABSTRACT

Social virtual reality uses motion tracking to place people in virtual environments as animated avatars. Often this tracking only measures the position and orientation of the head and hands, and from this estimates the body pose. Optical hand tracking is an important technology to enable such avatars, but can frequently fail and cause motion errors when the hands are visually obscured. This paper presents three amelioration strategies to handle these errors and demonstrates experimentally that all three are effective in reducing their impact. This setting is also used to explore general issues around study design for motion perception. Different strategies for presenting stimuli and soliciting input are compared. The presence of a simultaneous recall task is shown to reduce but not eliminate sensitivity to motion errors. Finally, it is shown that motion errors are interpreted, at least in part, as a shift in interlocutor personality.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in collaborative and social computing**.

This work is licensed under a Creative Commons Attribution International 4.0 License.

SAP '21, September 16–17, 2021, Virtual Event, France
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8663-0/21/09.
https://doi.org/10.1145/3474451.3476235

KEYWORDS

motion perception, hand tracking, VR avatars, IK errors, study design

ACM Reference Format:
Ylva Ferstl, Rachel McDonnell, and Michael Neff. 2021. Evaluating Study Design and Strategies for Mitigating the Impact of Hand Tracking Loss. In *ACM Symposium on Applied Perception 2021 (SAP '21), September 16–17, 2021, Virtual Event, France.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3474451.3476235

1 INTRODUCTION

With continuous improvements in optical hand tracking technology, first-person camera-based hand tracking is increasingly preferred over controller-based tracking in Virtual Reality. 1st person camera-based tracking relies on cameras mounted on the VR headset, and computer vision to detect the hands within the camera image. This controller-free tracking allows users more natural interaction with the virtual environment, increasing enjoyment and engagement. However, optical hand tracking can still fail frequently due to sub-optimal visual conditions. For example, the user's hand may leave the visibility space of the head cameras, or one hand may occlude the other, or the hand may be moving too fast, creating motion blur in the camera image. When hand tracking is lost in these scenarios, the default solution is to leave the hand of the user's avatar hanging where it was last tracked, and, when tracking resumes, the hand suddenly pops to the new position, creating a jump in the body motion. This prominent error in the rendered motion may decrease perceptions of realism and impact user experience.

In this work, we first investigate users' perception of these motion errors resulting from hand tracking loss for a conversational

partner, and propose and assess three different amelioration strategies for these scenarios. We propose a method for simulating hand tracking loss occurrences and generate samples for an perceptual experiment.

Second, we investigate aspects of experiment design. No common standards have been codified for assessing users' perception of motion in social interaction. We explore stimulus presentation, question design, the impact of providing viewers with a simultaneous task and whether errors are read as social signals.

These themes are explored through four experiments involving several hundred participants. The first experiment confirms people's sensitivity to tracking errors and shows sensitivity increases with improved motion quality. Experiment 2 assesses the error amelioration strategies, showing that all are beneficial, and compares different survey prompts and presentation modes. Experiment 3 shows that error sensitivity is reduced when participants are given an additional conversation-based task. Experiment 4 shows that motion errors are also interpreted, at least in part, as shifts in character personality, which will have repercussions for social VR.

2 BACKGROUND

2.1 Motion Errors

Numerous previous studies have explored people's sensitivity to motion errors. People were more sensitive to horizontal than vertical velocity errors when viewing jumping animations, and more sensitive to added accelerations than decelerations [Reitsma and Pollard 2003]. Hodgins et al. [2010] assessed sensitivity to fairly substantial anomalies in facial and body motion for a scenario involving an arguing couple. Facial anomalies were found to be more disturbing than body anomalies, and gaze tracking revealed that a majority of attention was focused on the face, which may be explanatory of these results. The important role of visual attention for error detection was also illustrated in Harrison et al. [2004]: For a simple, two link line drawing as a proxy for a human character, people were less able to detect changes in limb length when given a task of counting rotations or under presence of a low contrast distractor in the scene. Context, specifically the background shown behind a motion, can also impact perceptions of the emotional quality of a motion [Heimerdinger and LaViers 2019].

For a snooker game scenario, people were found to be more tolerant of errors when shown a realistic versus an abstract visual environment [Reitsma and O'Sullivan 2009]. This may have been caused by response bias: people were more likely to report errors in the abstract environment whether or not they were there.

Body tracking in VR significantly increases users' embodiment and social presence [Eubanks et al. 2020], and some work has investigated the impact of tracking errors on user experience. Embodiment was found to degrade for an athletic task as latency exceeds 125ms, and severly so over 300ms [Waltemate et al. 2016], and latency sensitivity is influenced by the speed of the motion performance [Hoyet et al. 2019]. In a social task, no decline in social presence was found for even severe lag and jitter, but certain errors impacted embodiment, enjoyment and perceived usability [Toothman and Neff 2019].

For avatar-object interaction, tracked motion will frequently not align perfectly with the virtual objects and a choice between preserving visual fidelity (not allowing the hand to penetrate objects) or preserving motion fidelity has to be made. Users preferred preservation of visual fidelity, though their task performance was better under preservation of motion fidelity [Canales et al. 2019].

For an error-prone task in VR using hand-tracking, displaying the embodied hand wrongly moving into the symmetrically opposite direction elicited a neural response consistent with semantic or conceptual violations, different from the neural response to self-generated errors [Padrao et al. 2016]. This difference was larger the more body ownership a participant felt for the virtual body.

Limited work has assessed the impact of errors on co-speech gesture. Perceptual work often focused on validating synthesis algorithms [Kucherenko et al. 2020] or measuring perceived personality (e.g. [Neff et al. 2011, 2010; Smith and Neff 2017]) and emotion (e.g. [Castillo and Neff 2019]). Adding discontinuities to gesture was found to decrease the impression of Emotional Stability [Smith and Neff 2017], so errors may have the impact of changing the perceived character personality, rather than being viewed as errors. Previous work found limited sensitivity for detecting mismatches between agent voice and gesture [Ennis et al. 2010], suggesting the potential to substitute motion to mitigate tracking errors. However, while some errors in gesture motion may go unnoticed, temporal misalignment can impact the perceived *speech*: Bosker and Peeters [2021] report a manual McGurk effect where the timing of a beat gesture (a non-meaningful, rhythmic gesture) can influence the vowel a listener perceives and potentially change the perceived meaning.

2.2 Study Design

In running a perceptual study, a large number of variables must be determined that can each impact the results. Do you put the motion in context or view it in isolation? Are there distracting activities or gaze draws or is the participant free to focus on the stimuli? Is it more important to measure awareness of errors or their impact on subjective criteria like presence or personality? What is the best way to present stimuli and is there a preferred question formulation? At this point, there are no established standards and many variations on the exemplars above have been used. Some of these challenges have also been discussed in Zell et al. [2020], as well as specifically for evaluation of co-speech gesture in Wolfert et al. [2021].

Two standard techniques developed for evaluating video quality were compared by Nehmé et al. [2019] for detecting errors in 3D models. The Double Stimulus Impairment Scale (DSIS) provides a reference, pristine video, followed by a degraded version. The user rates the level of degradation on a discrete 5-point scale: Imperceptible (5), Perceptible but not annoying (4), Slightly annoying (3), Annoying (2), Very annoying (1), with a recommended 10s presentation. The Absolute Category Ratings with Hidden Reference (ACR-HR) presents each impaired video individually and has it rated on a 5-point quality scale: bad, poor, fair, good and excellent. Other standard alternatives include Pairwise Comparison (PC), where users pick one of two options, and the Subjective Assessment Methodology for Video Quality (SAMVIQ), which uses a 101-point scale to rate a single video and allows users to change earlier ratings after viewing more videos. Nehmé et al. found DSIS to be more

stable and accurate for finding errors in geometry viewed in VR, meaning that fewer subjects were required to reach a particular error level. ACR rates looked to become comparable after a learning phase.

Comparing the use of user ratings, eye gaze and EEG for detecting errors in video, Tauscher et al. [2017] found each to be informative in different ways. For example, saccades gave a useful measure of attention, and EEG required large errors to produce a signal.

Using free text response versus ratings on a predefined scale for character personality judgements, some overlap in results was found, but different personality traits were more prominent in the different forms of rating [Liu et al. 2015].

Viewing perspective may impact people's perception of error. In an exercise where people had to adjust their avatar's weight to match their own, participants, particularly males, made more error in first person than third person perspective [Thaler et al. 2019].

3 METHOD

There were 4 main user studies, designed to investigate users' general sensitivity to motion errors (Exp. 1, Sec. 4), perception of amelioration strategies under varying study design (Exp. 2, Sec. 5), influence of task attention (Exp. 3, Sec. 6), and motion error induced shifts in personality perception (Exp. 4, Sec. 7).

3.1 Procedure

Due to pandemic restrictions, all experiments were video-based and were completed from users' homes. To ensure reasonable viewing conditions, a minimum screen size of 13" was required and all videos were automatically played in full screen mode, could only be played once, and had to be played to completion before participants were able to answer prompts. Fair reimbursement was used to ensure response validity [Jonell et al. 2020]. An Institutional Review Board approved the study prior to data collection.

All perceptual experiments were created with Qualtrics and distributed via Amazon Mechanical Turk. Participants received $15 per hour. Except where noted, there were 100 participants per sub-experiment. The term sub-experiment refers to a between-subject condition and is chosen to distinguish from the motion conditions (Sec. 3.3). There were main 4 experiments, each consisting of 1-5 sub-experiments. Experiments lasted approximately 1 hour, except for DSIS (see Table 1) which took approximately 1.5 hours.

Participants first self-reported their eligibility to participate (e.g. adequate English knowledge and no visual impairments) and gave informed consent. Next, they received instructions for the experiment and watched a video clip presenting examples of the range of stimuli to follow. After the example video, the experiment started. A trial always consisted of watching a video clip followed by a question prompt. In DSIS only, participants watched two video clips (the *clean* version of a clip followed by a degraded version) instead of one before reaching the question prompt. The current and total trial numbers were presented throughout the experiment to inform participants of their progress.

Since we are particularly interested in understanding sensitivity to error during social interaction, test clips were selected of individual people engaged in dialogue from the *Talking With Hands 16.2M* dataset [Lee et al. 2019]. This consisted of high quality, full body, motion capture data, along with speech audio. The input motion was the same for each experiment and consisted of 10 different clips from each of 10 motion captured subjects (7 male and 3 female), for a total of 100 utterances. Each clip was approximately 20 seconds long and selected to contain coherent speech without interruptions by the conversation partner.

The high-quality original motion was processed to mimic 1st person camera-based VR tracking using an inverse kinematics algorithm referred to as "three point IK" (3pt IK). For this, the six degrees of freedom of the head and hands were extracted from the motion capture and used to generate a new full body motion with the custom 3pt IK solver contained in the Oculus VR SDK. The solver takes as input the 3D world positions and rotations of the head and the two hands and estimates the upper body pose and root position. Tracking failures were simulated as described in the next section.

All motion data was rendered on a male or female model depending on the performer's gender. The used 3pt IK algorithm only reconstructs the upper body motion, therefore the lower body is faded out and a framing is selected that shows the character from mid-thigh to the top of his/her head, as shown in Fig. 1, mimicking what may be observed during conversational interaction with another person.

3.2 Simulating Tracking Failures

Five measures were implemented using the Unity3D game engine to simulate when tracking loss would likely occur during inside-out tracking. We first determined common error sources by consulting VR tracking developers and then tuned error source thresholds by empirically comparing to actual hand tracking losses with the Oculus Quest 1.

To simulate the percentage of each hand that was visible for the head cameras, one hand was colored in red, one in green, and a virtual camera was placed on the character's head, with a field of view of 160 degrees and tilted 15 degrees downwards. The virtual head camera's image was rendered to an image which was analyzed for the number of red and green pixels pertaining to the hands. The number of colored pixels found was scaled according to the distance from the camera image's center to account for the strong distortion resulting from the wide camera field of view.

Hand tracking may also fail if too few fingers are visible to the head cameras; e.g. even with a large portion of the back of the hand being visible, the hand may not be identified as such without visible fingers. To account for this, we virtually divide the hand into segments, including the base, mid, and tip of each finger. Rays are cast from the location of the head camera to the hand segments and visibility is assessed by the number of segments reached successfully. Tracking is considered dropped if there is less than 40% visibility.

We assess the hand's angle with respect to the line of projection of the head camera by calculating the dot product between the following vectors: (1) The vector connecting the camera to hand, and (2) the vector connecting the base of the hand to the base of the middle finger. The absolute value of dot product is 1 if the vectors are parallel and reaches 0 as they become perpendicular to each

Table 1: Prompts used to measure error in different sub-experiments.

Abbreviation	Full Name	Prompt	Responses
DSIS	Double Stimulus Impairment Scale	Please rate the motion error in the second clip compared to the first	Imperceptible(5), Perceptible but not annoying(4), Slightly annoying(3), Annoying(2), Very annoying(1)
DSIS_NP	Non-Paired DSIS	Please rate the motion error in this clip.	as above
ACR	Absolute Category Rating	Please rate the motion quality in this clip.	Excellent, Good, Fair, Poor, Bad
NAT_L	Likert Naturalness	The motion in this clip appears natural	Strongly agree - Strongly disagree (7 point, standard Likert labels on each value)
ERR_L	Likert Error	This motion contains errors	as above

other. We define an inadequate hand-camera angle as a dot product larger than 0.75.

To account for the constraints of a physical head camera losing focus for objects very close to the lens, we define the minimum distance of the hand from the head camera required for visibility as 0.12 m. To account for motion blur obscuring the image of a physical head camera, we define a maximum speed of 3 m/s of the hand above which visibility is lost.

3.3 Error Amelioration Strategies

We developed a number of error amelioration strategies for handling hand tracking failures. In embodied VR applications, people's motion must normally be reconstructed with only data on the location of the head and two hands. As described in Sec. 3.1, 3pt IK reconstruction is simulated by feeding the motion-captured head and hand data to the IK solver contained in the Oculus VR SDK. The output of this process produces the highest quality stimuli used in the experiments because the hand data is always present. It is referred to as *clean*, since no errors have been introduced.

1st person camera-based hand tracking uses cameras mounted on the VR headset to reconstruct the hand pose. This can fail when the optical quality is too low due to the hands being too far from the cameras, self-occlusions, the hand angle being parallel to the lines of projection and/or motion blur. To simulate the impact of these error sources, we calculate their likely occurrence during the motion sequence, as described in Sec. 3.2. When one of these errors occur for some number of frames, tracking is considered dropped and the hand location is frozen at the last valid frame. In the *full error* motion condition, no attempt is made to lessen the impact of this error. In practice, additional tracking failures may occur due to object occlusion or poor lighting, but those are not modeled here.

Three amelioration strategies were devised for reducing the impact of these unavoidable tracking errors. They provided the remaining motion conditions:

(1) *fade*: Upon loss of hand tracking, the joint angles of the respective hand and arm are frozen. When tracking resumes, the hand is re-aligned with the tracking position by interpolating the respective joint angles over time to create a smooth transition with similar speed to the other gesture motion. This strategy should work well for speakers with high gesture frequency, where tracking losses are usually short.
(2) *rest*: If hand tracking has been lost for longer than 0.4 seconds, the hand is moved to a rest position (hand hanging by the side of the body). If tracking resumes within 0.4 seconds, the hand is re-aligned with the tracking position by interpolating the respective joint angles over time to create a smooth transition. This strategy addresses severe tracking losses of long duration, avoiding prolonged unnatural poses such as the hand hanging in mid-air.
(3) *hybrid*: If hand tracking has been lost for longer than 2 seconds, or if hand tracking was lost around waist height for at least 0.08 seconds, move the hand to a rest position. If hand tracking is lost and the previous two conditions do not apply, use *fade*. Motivation for using the height of the hand upon occlusion comes from the frequent loss of hand tracking when subjects move their hands to a rest position by the side of the body, normally leading to the virtual hands remaining hanging at waist height where tracking was lost. This solution attempts to find a balance between leaving the hand hanging in an awkward position for too long and too quickly retracting a hand when tracking may resume from the same general location. This strategy seeks to combine the advantages of *fade* and *rest*, with potential to perform the best.

We also considered predictive models of error amelioration. Experiments were performed using motion momentum at occlusion time to predict they trajectory, with damping to slow the motion; however, this did not yield good results. In addition, we observed large differences in gesture style between speakers and therefore believe a good model for addressing the complexities of gesture motion prediction would need to be speaker-specific. For versatility and robustness, we therefore chose the above amelioration strategies.

3.4 Statistical Analysis

All data was fit with Cumulative Link Mixed Models (CLMMs), which compare differences in distributions of ordinal data [Christensen 2015, 2018]. CLMMs treat responses as categorical, ordered data, which is an appropriate approach for Likert data. Analysis was performed in R using the *ordinal* package. Post-hoc analysis

was performed with least square means and pairwise comparisons using Tukey correction, using the lsmeans package in R.

4 EXP. 1: MOTION ERROR SENSITIVITY

The first goal of Experiment 1 was to confirm that the errors in wrist position were noticeable. For this, two conditions were compared: *clean*, the highest quality motion with no artificial errors, and *full error*, which contained all errors with no amelioration strategy.

The second goal was to compare two different forms of motion reconstruction. The target VR application receives as input 6 DOF (degree of freedom) information for the head and hands and uses an IK algorithm to solve for a full body skeleton pose from these three points. This is a heavily underspecified problem and hence the resulting body pose may not always be natural. This increased error in the body pose may impact people's ability to notice errors in the wrist position. It may also be that future algorithms are able to better reconstruct the body pose from the three input points. For both these reasons, we want to compare people's sensitivity to error in the 3pt IK scenario and a higher quality motion baseline. For the latter, we directly use the input motion capture and generate errors in the motion by freezing the joint angles during the periods when tracking is expected to be lost. This creates higher quality body motion, with the same periods of tracking error. We dub this reconstruction method *FK*, because it relies on forward kinematics to determine the pose from joint angles, and refer to the 3pt IK reconstruction method as *IK*.

There were two sub-experiments, one for IK and one for FK stimuli. Each contained the *clean* and *full error* conditions. The test motion was generated from the 100 clips described in Sec. 3.3. For each sub-experiment, there were two sets of clips, with one half of the clips having errors introduced in one set and the other half having errors in the other set, and participants were equally distributed across the two clip sets. 40 participants partook in each sub-experiment. This experiment was run without speech audio.

The prompt used in this experiment was "This motion contains errors," which participants rated on a 7-point Likert scale, ranging from Strongly Agree to Strongly Disagree.

4.1 Results

For FK, the mean for the *clean* clips was 3.05 (SD 1.73) on a 7-point Likert scale and for *full error*, the mean was 5.16 (SD 1.98), plotted in Fig. 2. The distributions are significantly different, (z=29.33, p<2e-16), reflecting that people reliably detected more error in the clips that had jumps in the wrist position.

For IK, the mean for the *clean* clips was 4.38 (SD 1.87) on a 7-point Likert scale and for *full error*, the error condition, the mean was 5.05 (SD 1.71), plotted in Fig. 2. The distributions are significantly different, (z=11.73, p<2e-16), again showing that people reliably detected more error in the clips with jumps in the wrist position.

A secondary analysis compared performance on the two motion types, FK and IK, by forming a single cumulative link model on the data for both sub-experiments. This showed significant main effects for condition (z=32.75, p<2e-16) and motion type (z=21.78, p<2e-16), as well as a significant interaction between them (z=-17.97, p<2e-16). Post-hoc analysis with least square means using Tukey correction shows that for the *clean* condition, people observed significantly less error for the FK reconstruction (z=-21.78, p<.0001). In the *full error* condition, they observed significantly more error with the FK reconstruction (z=3.653, p=0.0015).

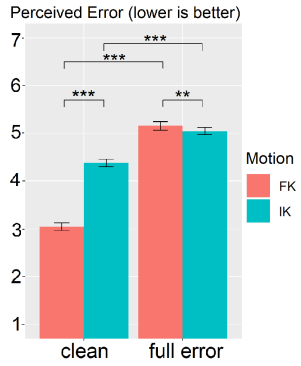

Figure 2: Perceived Error for FK and IK based motion reconstructions, including the clean (unmodified) and full error conditions.

4.2 Discussion

First, Experiment 1 confirmed that the errors generated by simulated drops in hand tracking were noticeable to observers with both motion reconstruction methods. Second, the best quality motion condition (*clean*) was seen as having significantly more error for the IK reconstruction than FK. This indicates that participants were sensitive to the errors introduced by the IK processing. Error ratings of the *full error* clips were much closer across motion conditions, but still significantly higher in the FK condition. While the difference was small, it seems to indicate that participants found tracking errors more noticeable when the base motion quality was higher.

5 EXP. 2: AMELIORATION STRATEGIES AND PROMPT DESIGN

There were two objectives for Experiment 2. First, we wanted to understand which amelioration strategy would be most effective for addressing the inevitable tracking errors. Second, we wanted to understand whether question type and response form impact the results obtained. There were five sub-experiments, each using a different prompt to obtain ratings (Table 1), and separate sets of participants. The prompts included ACR and DSIS, as two standard metrics, along with Likert scale questions that have been common in previous motion evaluation work. Some prompts focused on rating the amount of error in the motion (DSIS, DSIS_NP and ERR_L), while others focused on motion quality (ACR, NAT_L). DSIS is a paired test, where the best quality motion for a clip is shown first, followed by the degraded clip (or a repeat of the best quality motion). All other tests only show a single stimuli before asking for a rating. DSIS_NP was added as a single stimuli variant of DSIS to separate presentation mode from response form.

Each sub-experiment contains all 5 motion conditions and all 100 utterances (Sec. 3.3). Each utterance was rendered with each motion condition, and 5 clip sets were then formed that each contained 20 clips at each motion condition (2 per speaker). Participants were

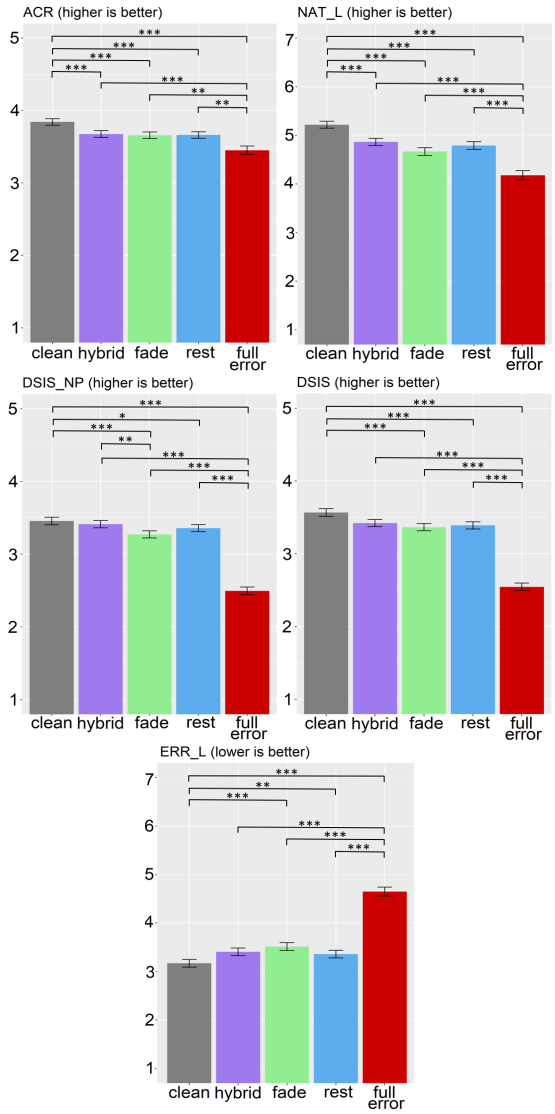

Figure 3: Ratings by condition for each of the prompts. Note that some prompts are 5-point and some are 7.

randomly assigned to view one of these 5 sets, with approximately 20 participants per set (100 participants per sub-experiment). This way, every motion condition of of every utterance was included while an individual participant only saw one condition for a given utterance.

5.1 Results

All question prompts showed significant differences between conditions (see Fig. 3). For ACR, NAT_L, DSIS and ERR_L, post-hoc analysis identified three groups: *clean* performed significantly better than all other conditions, *full error* performed significantly worse, and the 3 amelioration conditions were in the middle and not significantly different from each other. For DSIS and NAT_L, the p-values for the pairwise comparisons that led to these groups are all $p<0.0001$ and for ACR and ERR_L, $p<0.01$ in all relevant comparisons. The exception to the three level groupings is DSIS_NP. For this prompt, *clean* was not significantly different from *hybrid* ($z=-1.583$, $p=0.51$) and *hybrid* was significantly better than *fade* ($z=-3.823$, $p=0.0012$).

There was some variation between the speakers from the dataset. The relative ratings of the speakers is generally consistent across the various prompts, as is the ordering of the five motion conditions. There is variation across speakers in the average difference between *clean* and *full error* ratings, but this variation does not appear to be significant. (Illustrating figures in appendix.)

5.2 Discussion

Given that *full error* performed significantly worse in all sub-experiments, it is clear that all amelioration strategies led to improvement. There was no clear separation between the amelioration approaches as they were often statistically indistinguishable. With caution, we suggest that the *fade* strategy may be slightly worse. The main weaknesses of *fade* - potentially leaving an arm in awkward position for an extended period - will likely be more apparent in prolonged conversations; because our stimuli were short and consisted of segments with gesturing that lead to quicker recovery this weakness may not have become as apparent.

In terms of question choice, all conditions except DSIS_NP provided the same groupings of conditions ratings on this dataset, so would lead to the same result on this test. A closer inspections suggests that prompts based on error (DSIS, ERR_L and DSIS_NP) worked particularly well for differentiating the *full error* case from the rest, but did not offer as much separation for *clean* from the amelioration conditions. A possible explanation is that people viewed the error prompts as an identification task and could identify issues with *full error* particularly well, but they provided a more holistic judgment for the naturalness/quality questions which better separated the top four conditions. A participant may notice sudden jumps in the motion and still decide that the overall motion quality was good or natural. However, when specifically prompted to report if the motion contained errors, the participant may also agree. Finally, NAT_L shows more variation in the amelioration ratings than ACR. Further testing should be performed to confirm this hypothesis, but it appears that NAT_L may be a good candidate for a holistic judgment of motion style where differences are subtle.

6 EXP. 3: INFLUENCE OF TASK

In real world applications of character technology, users are (hopefully) engaged with the experience, interacting with characters, trying to complete game goals, etc. Their attention is not fully focused on trying to detect motion errors. Experiment 3 is designed to investigate if users are less sensitive to motion errors when given an interaction task that requires their attention, thus providing a more realistic test context. Participants were told they would be asked a question on the character's dialog after each clip. The questions were multiple choice and required close listening to the dialog. Examples include: "What happened to the driver? a) He was killed, b) He fled the scene, c) He was in an accident, or d) He won the race" and "The speaker talks about a... a) Lego movie, b) Disney movie, c) Theatre performance, or d) A horror movie."

Content questions were presented together with and the motion quality rating question. Two sub-experiments were run as part of this experiment. One used the Likert Error rating prompt ERR_L and the other the the Likert Naturalness prompt NAT_L (Table 1).

6.1 Results

Fig. 4 shows the results for both sub-experiments. For NAT_L, *clean* was significantly better than all other conditions and *full error* was worse. *Hybrid* was significantly better than *fade* (z=3.127, p=0.015) and there was a tendency for *rest* to also be significantly better than *fade* (z=2.717, p=0.052). There was no significant difference between *hybrid* and *rest* (z=0.416, p=0.994).

For ERR_L, *clean* was only significantly better than *full error* and *fade* (z=-3.653, p=0.0024), with no significant difference for *hybrid* (z=-1.784, p=0.38) and *rest* (z=-1.537, p=0.54). There was no significant difference between *hybrid*, *rest*, and *fade*. *Full error* was significantly worse than all conditions.

Fig. 5 shows a comparison of the results from Exp. 2 with those from Exp. 3 that added content questions. For NAT_L, comparing the base model and the model that included content questions, there is no significant difference for the *clean* condition (z=-2.858, p=0.12), but all the conditions with some error are rated significantly lower in the base model than for the model with additional content recall task (*hybrid* z=-4.163, p=0.0013, *fade* z=-3.973, p=0.0028, *rest* z=-4.820, p=0.0001, full error z=-5.562, p<.0001). For ERR_L, in all cases, the error ratings are significantly lower in the base model than in the model with dialog content questions. Notably, the spread between conditions is much larger in the base case.

For each participant, we calculated their average score and average accuracy answering questions for each of the error conditions. Regression lines for the extreme cases of *full error* and *clean* are plotted in Fig. 6. The lines for the other error conditions lie between these, but are omitted for visual clarity. Participants who were more accurate answering questions also tended to be more accurate in their motion error detection.

6.2 Discussion

Even with the added question answering task, all the amelioration techniques improved participants impression of motion quality,

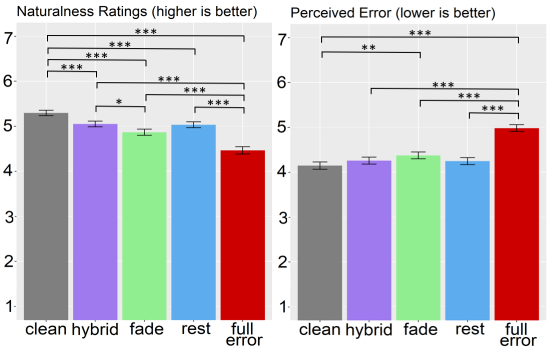

Figure 4: Ratings for Exp. 3 (Influence of Context) across condition for both the error detection and naturalness sub-experiment.

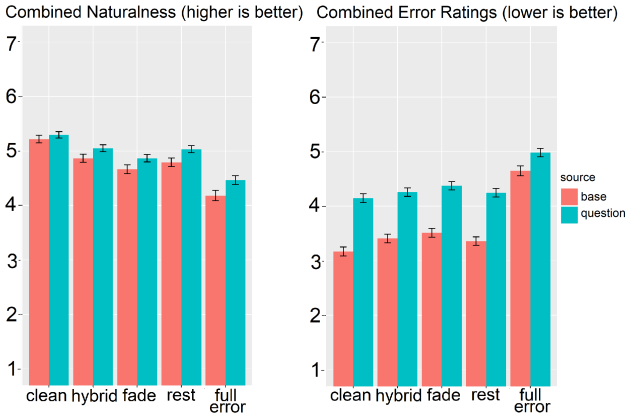

Figure 5: Combined ratings for the base version of the experiment and the version that included content questions.

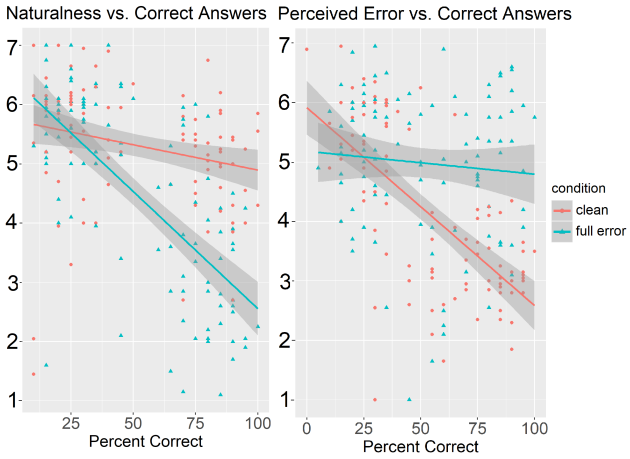

Figure 6: Regression lines fit to scatter plots of participants' accuracy answering questions and ratings of motion. For visual simplicity, only the extreme *full error* and *clean* conditions are shown. The amelioration conditions lie in between (full plot included in Appendix).

both in terms of naturalness and the amount of error. There is more evidence that *fade* is the weakest of the techniques, with significantly lower naturalness ratings than *hybrid* and a tendency for them to be lower than *rest*. It was also the only amelioration condition that participants rated with significantly more error than *clean*, although the difference between the amelioration conditions was not significant. The naturalness prompt is more effective at separating the amelioration conditions from the best case (*clean*). This may suggest that people are not consciously aware of things they would call errors in the amelioration case, but still find the motion quality lower. This provides more evidence that NAT_L may be a good candidate for assessing subtle motion variation.

The presence of a conversational task reduced sensitivity to motion quality differences, as expected. People noticed error less

in clips with error and thought clips with error were more natural than participants not engaged in a conversational task.

Since participants who were more accurate at answering the questions also tended to be more accurate in their motion observations, it does not appear that people were dividing limited attentional resources between the two tasks. Rather, it seems like some participants paid more careful attention to all aspects of the task than others. Alternatively, some participants might have had more difficulty with the conversational task, perhaps due to speaker accents, leading them to use more attentional resources for listening to the speech, resulting in less available resources (and hence lower performance) for the motion observation task, while also doing more poorly on the conversational task.

In both tests, it appears that there was less variance as a function of an additional conversational task for the 'easy' condition. For ERR_L, it is easiest to detect that error is present in *full error*. It is similarly easiest to detect that the *clean* clips are natural.

7 EXP. 4: PERSONALITY PERCEPTION

Noticeable motion errors may degrade users' experiences. They also, or alternatively, may shift their impression of the person they are interacting with. Previous work has established that small changes in the performance of gesture can reliably influence the perceived personality of a character (e.g. [Smith and Neff 2017]). We sought to understand if the motion errors produced by tracking loss can also impact perceived personality, establishing that such errors lead to a shift in the impression of an interlocutor.

An experiment was conducted using the IK stimuli from Experiment 1, conisting of the *clean* and *full error* conditions. The stimuli presentation was structured the same, but in this experiment, instead of rating error, participants were asked to rate the character's personality by providing Likert responses to the prompts of the Ten-Item Personality Inventory for each clip [Gosling et al. 2003]. This is a compact instrument for measuring the Five Factor personality model (Extroversion, Openness to Experience, Emotional Stability, Agreeableness and Conscientiousness)[Norman 1963], a widely used model of personality in social psychology. 45 participants partook in this experiment, 22 saw the first set of motions and 23 the second.

7.1 Results

Results are summarized in Fig. 7 and show small, but consistent variation across the personality traits. Since there is no reason to expect a relationship across personality traits, a separate cumulative link model was used to analyze each personality trait individually. If we use conservative Bonferroni correction to adjust our alpha to $p \leq 0.01$ to account for the five tests, the differences for Extroversion (z=-3.084, p=0.0020) and Openness to Experience (z=-2.68, p=0.0074) are statistically significant. The differences for Agreeableness (z=-2.31, p=0.021), Conscientiousness (z=1.34, p=0.18), and Emotional Stability (z=-2.045, p=0.041) were not. For both Extroversion (*clean* mean=4.11 SD=1.26, *full error* mean=3.99, SD = 1.26) and Openness to Experience (*clean* mean=4.199 SD=1.088, *full error* mean=4.126, SD =1.092), the presence of error shifts the perceived perception towards the negative end of the scale.

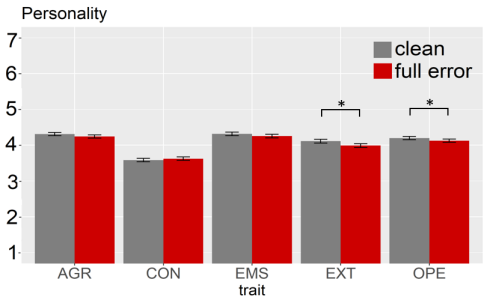

Figure 7: Ratings for Exp. 4 comparing the perceived personality of clips with and without error.

7.2 Discussion

The significant differences in Extroversion and Openness to Experience, and nearly significant changes in Agreeableness and Emotional Stability, suggest that the presence of error is shifting people's impression of the avatar, instead merely being viewed as a technical flaw. For all traits where the change was significant or nearly so, the error shifted the trait towards the more negative end of the spectrum. This could suggest that there was a general halo effect, where the error led to people making more negative interpretations of the people overall. In all cases, however, the differences are very small so caution should be used in determining the importance of this effect.

8 DISCUSSION AND CONCLUSION

In this work, we study hand tracking loss for conversational interaction. First, we proposed a method for simulating headset-based hand tracking loss occurrences on error-free motion-captured data and we assess participants' sensitivity to these errors, including effects on perceived personality. Next, we propose three error amelioration strategies and show that each improves perceived motion quality. Finally, we investigate study design questions.

While all three error amelioration strategies improved perceived motion quality, there was a tendency for *fade* to be worse than the others, *rest* and *hybrid*. Sensitivity to motion quality was lower when engaged in a conversational recall task, but amelioration still made a significant improvement. Sensitivity to motion errors was also higher for higher quality base motion.

Most question prompts were able to separate the stimuli into three levels: *clean*, ameliorated error (*fade*, *rest*, *hybrid*) and *full error*. Prompts based on gauging error produced particularly distinct ratings for the high error case, but performed less well in terms of separating the amelioration and *clean* stimuli. Indeed, ERR_L in Exp. 3 and DSIS_NP in Exp. 2 did not always produce significantly different ratings for the *clean* baseline and amelioration conditions. Overall, a naturalness Likert scale better separated the *clean* and amelioration conditions and was a good choice in this study when holistic judgment of motion style was desired. We found no advantage of prior showing of the *clean* reference motion (*DSIS*) for motion error sensitivity.

The presence of error significantly lowered the perception of the personality traits Extroversion and Openness to Experience, indicating the people in part interpret error as a "feature" of their interlocutor. The magnitude of this change was small, however.

There is room for improvement in the amelioration strategy since none was consistently indistinguishable from the *clean* motion. Machine learning methods could be used for determining optimal response when tracking fails. Another improvement could come from applying amelioration strategies directly to the IK input: By feeding a hypothesized pose into the IK algorithm, the rest of the body pose is adjusted accordingly and would likely lead to an overall more natural appearance of the motion. It would also be interesting to test these strategies in VR where they could be applied to a person's own avatar. Different strategies may be preferable in a first person versus third person view; while for a third person view, the most visually appealing solution may be preferred, in a first person view, the solution spatially closest to the user's hands may best preserve user engagement and presence. Users may also alter their gesture behavior due to awareness of the limited tracking performance, and therefore in-VR gesture behavior may differ from the captured motion used in this study.

REFERENCES

Hans Rutger Bosker and David Peeters. 2021. Beat gestures influence which speech sounds you hear. *Proceedings of the Royal Society B* 288, 1943 (2021), 20202419.

Ryan Canales, Aline Normoyle, Yu Sun, Yuting Ye, Massimiliano Di Luca, and Sophie Jörg. 2019. Virtual grasping feedback and virtual hand ownership. In *ACM Symposium on Applied Perception 2019*. 1–9.

Gabriel Castillo and Michael Neff. 2019. What do we express without knowing? Emotion in Gesture. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 702–710.

Rune Haubo B Christensen. 2015. Analysis of ordinal data with cumulative link models—estimation with the R-package ordinal. *R-package version* 28 (2015).

Rune Haubo B Christensen. 2018. Cumulative link models for ordinal regression with the R package ordinal. *Submitted in J. Stat. Software* (2018).

Cathy Ennis, Rachel McDonnell, and Carol O'Sullivan. 2010. Seeing is Believing: Body Motion Dominates in Multisensory Conversations. *ACM Transactions on Graphics (SIGGRAPH)* 29, 4, Article 91 (July 2010), 9 pages. https://doi.org/10.1145/1778765.1778828

James Coleman Eubanks, Alec G Moore, Paul A Fishwick, and Ryan P McMahan. 2020. The Effects of Body Tracking Fidelity on Embodiment of an Inverse-Kinematic Avatar for Male Participants. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 54–63.

Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in personality* 37, 6 (2003), 504–528.

Jason Harrison, Ronald A Rensink, and Michiel Van De Panne. 2004. Obscuring length changes during animated motion. *ACM Transactions on Graphics (TOG)* 23, 3 (2004), 569–573.

Madison Heimerdinger and Amy LaViers. 2019. Modeling the interactions of context and style on affect in motion perception: stylized gaits across multiple environmental contexts. *International Journal of Social Robotics* 11, 3 (2019), 495–513.

Jessica Hodgins, Sophie Jörg, Carol O'Sullivan, Sang Il Park, and Moshe Mahler. 2010. The saliency of anomalies in animated human characters. *ACM Transactions on Applied Perception (TAP)* 7, 4 (2010), 1–14.

Ludovic Hoyet, Clément Spies, Pierre Plantard, Anthony Sorel, Richard Kulpa, and Franck Multon. 2019. Influence of Motion Speed on the Perception of Latency in Avatar Control. In *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE, 286–2863.

Patrik Jonell, Taras Kucherenko, Ilaria Torre, and Jonas Beskow. 2020. Can we trust online crowdworkers? *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents* (Oct 2020). https://doi.org/10.1145/3383652.3423860

Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. 2020. The GENEA Challenge 2020: Benchmarking gesture-generation systems on common data. (2020).

Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha S Srinivasa, and Yaser Sheikh. 2019. Talking With Hands 16.2 M: A Large-Scale Dataset of Synchronized Body-Finger Motion and Audio for Conversational Motion Analysis and Synthesis.. In *ICCV*. 763–772.

Kris Liu, Jackson Tolins, Jean E Fox Tree, Michael Neff, and Marilyn A Walker. 2015. Two techniques for assessing virtual agent personality. *IEEE Transactions on Affective Computing* 7, 1 (2015), 94–105.

Michael Neff, Nicholas Toothman, Robeson Bowmani, Jean E Fox Tree, and Marilyn A Walker. 2011. Don't scratch! Self-adaptors reflect emotional stability. In *International Workshop on Intelligent Virtual Agents*. Springer, 398–411.

Michael Neff, Yingying Wang, Rob Abbott, and Marilyn Walker. 2010. Evaluating the effect of gesture and language on personality perception in conversational agents. In *International Conference on Intelligent Virtual Agents*. Springer, 222–235.

Yana Nehmé, Jean-Philippe Farrugia, Florent Dupont, Patrick LeCallet, and Guillaume Lavoué. 2019. Comparison of subjective methods, with and without explicit reference, for quality assessment of 3D graphics. In *ACM Symposium on Applied Perception 2019*. 1–9.

Warren T Norman. 1963. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology* 66, 6 (1963), 574.

Gonçalo Padrao, Mar Gonzalez-Franco, Maria V Sanchez-Vives, Mel Slater, and Antoni Rodriguez-Fornells. 2016. Violating body movement semantics: Neural signatures of self-generated and external-generated errors. *Neuroimage* 124 (2016), 147–156.

Paul SA Reitsma and Carol O'Sullivan. 2009. Effect of scenario on perceptual sensitivity to errors in animation. *ACM Transactions on Applied Perception (TAP)* 6, 3 (2009), 1–16.

Paul SA Reitsma and Nancy S Pollard. 2003. Perceptual metrics for character animation: sensitivity to errors in ballistic motion. In *ACM SIGGRAPH 2003 Papers*. 537–542.

Harrison Jesse Smith and Michael Neff. 2017. Understanding the impact of animated gesture performance on personality perceptions. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 49.

Jan-Philipp Tauscher, Maryam Mustafa, and Marcus Magnor. 2017. Comparative analysis of three different modalities for perception of artifacts in videos. *ACM Transactions on Applied Perception (TAP)* 14, 4 (2017), 1–12.

Anne Thaler, Sergi Pujades, Jeanine K Stefanucci, Sarah H Creem-Regehr, Joachim Tesch, Michael J Black, and Betty J Mohler. 2019. The influence of visual perspective on body size estimation in immersive virtual Reality. In *ACM Symposium on Applied Perception 2019*. 1–12.

Nicholas Toothman and Michael Neff. 2019. The Impact of Avatar Tracking Errors on User Experience in VR. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 756–766.

Thomas Waltemate, Irene Senna, Felix Hülsmann, Marieke Rohde, Stefan Kopp, Marc Ernst, and Mario Botsch. 2016. The impact of latency on perceptual judgments and motor performance in closed-loop interaction in virtual reality. In *Proceedings of the 22nd ACM conference on virtual reality software and technology*. 27–35.

Pieter Wolfert, Nicole Robinson, and Tony Belpaeme. 2021. A Review of Evaluation Practices of Gesture Generation in Embodied Conversational Agents. *arXiv preprint arXiv:2101.03769* (2021).

Eduard Zell, Katja Zibrek, Xueni Pan, Marco Gillies, and Rachel McDonnell. 2020. From Perception to Interaction with Virtual Characters. (2020).

APPENDIX

Fig. 8 extends Fig. 6 to include all motion conditions: It shows regression lines that are fit to the scatter plots of participants' accuracy answering questions and ratings of motion. Fig. 9 shows the ratings from Experiment 2 per speaker, averaged over all conditions. They are broken down by condition in Fig. 11.

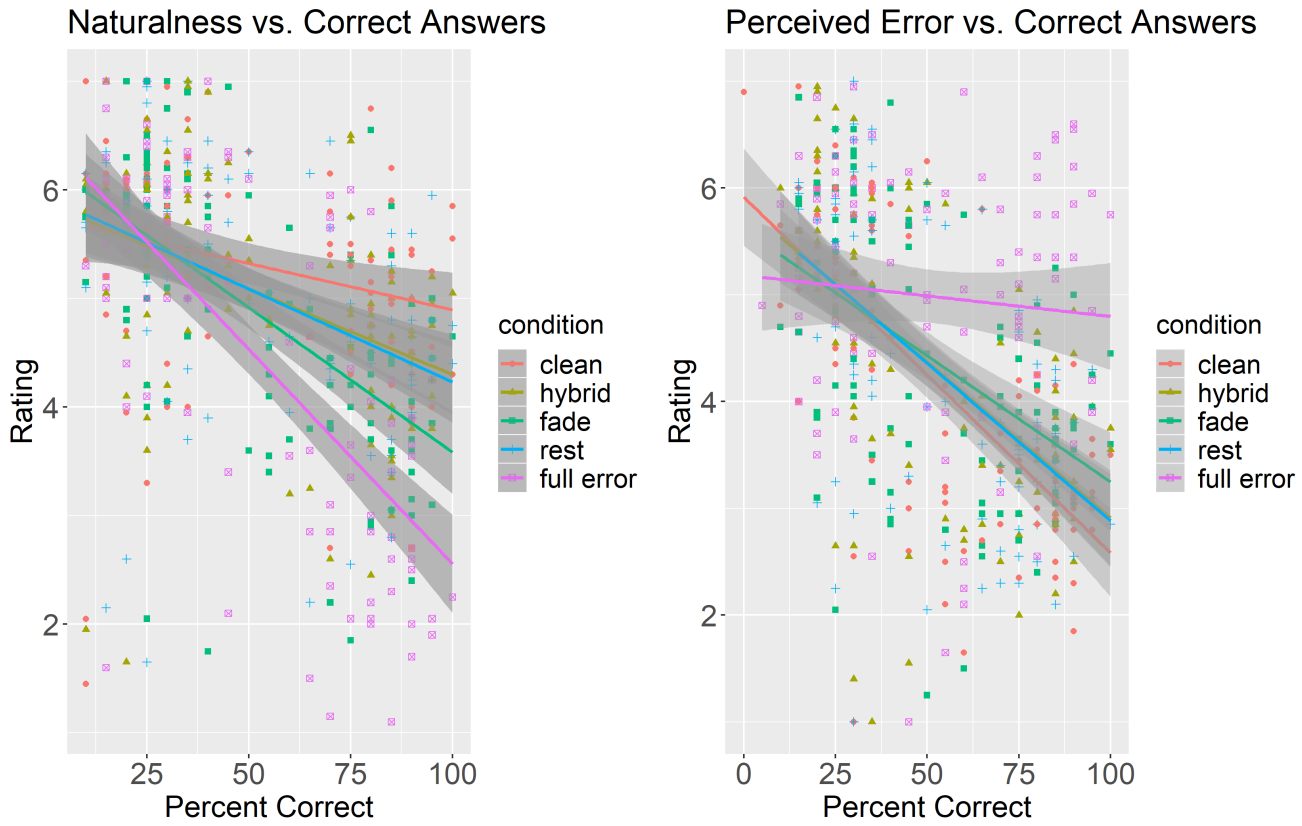

Figure 8: Regression lines fit to scatter plots of participants' accuracy answering questions and ratings of motion, for all motion conditions.

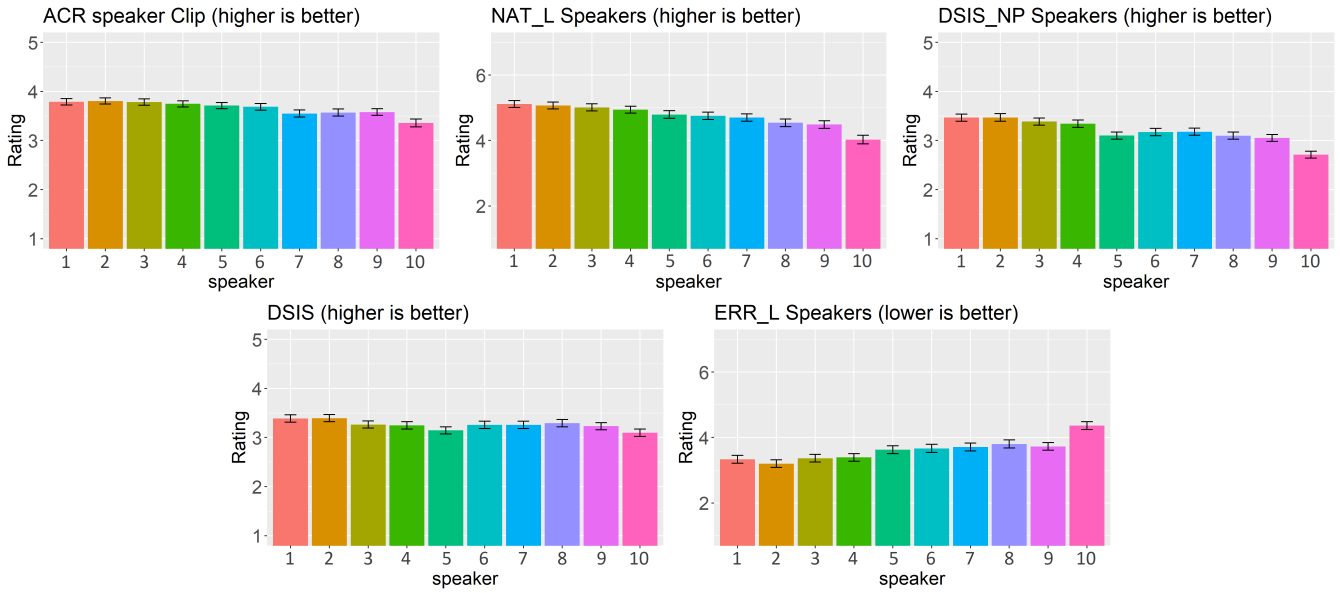

Figure 9: Ratings by speaker from Exp. 2 with the various prompts, averaged over all conditions. In all cases, the speakers are sorted by their order in the NAT_L experiment.

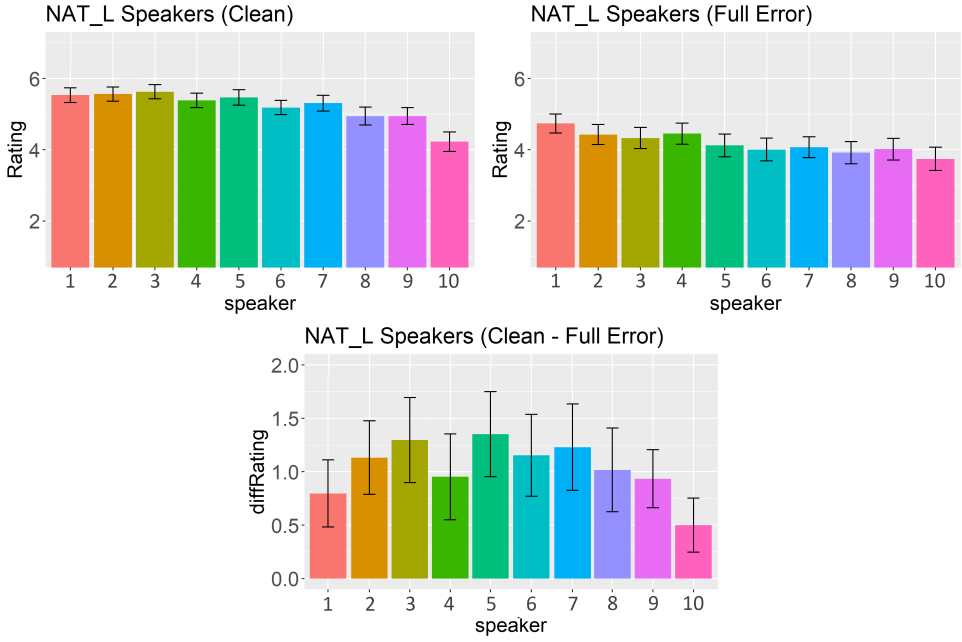

Figure 10: Ratings by speaker from Exp. 2 for the naturalness Likert scale. From left, the clean condition, the full error condition and the difference.

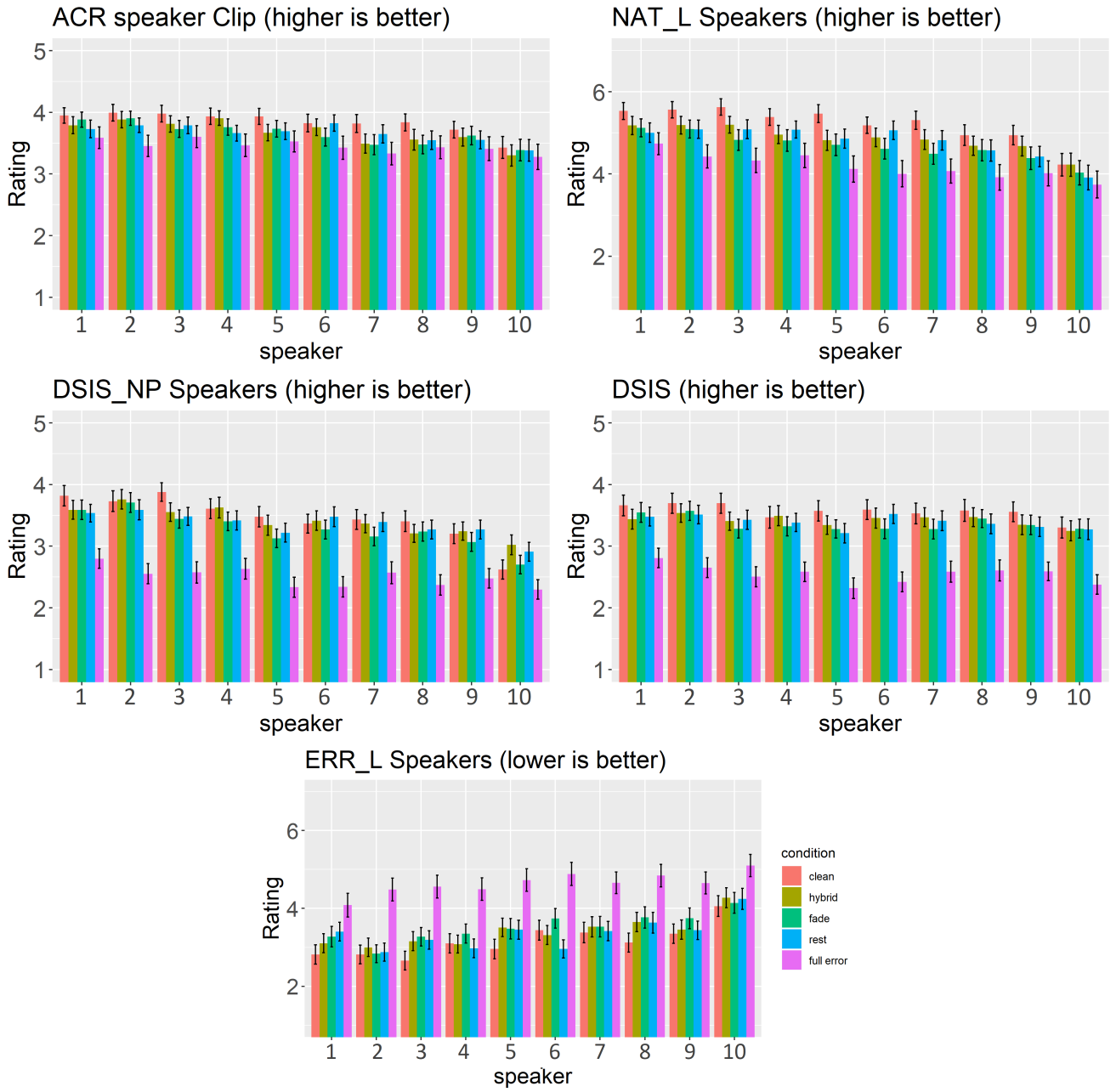

Figure 11: Ratings by speaker from Exp. 2 with the various prompts. In all cases, the speakers are sorted by their order in the NAT_L experiment.