Gaze and Head Rotation Analysis in a Triadic VR Job Interview Simulation

Saygin Artiran* Poorva S. Bedmutha† Aaron Li‡ Pamela Cosman§

University of California, San Diego

ABSTRACT

Virtual reality (VR) systems have shown potential in analyzing human behavior across various domains. We present the design and development of a VR-based job interview simulation tailored for analyzing gaze and head rotation behaviors in a context with two virtual interviewers. Our system allows users to encounter common interview questions and quantifies how they share their attention (gaze and head rotations) to engage with multiple interviewers based on their conversational role (speaking or listening). We detect voice activity to identify the start of user speech and guide the backchannels (head nods or verbal cues such as "uh-huh") given by the virtual interviewers. We track the user's gaze and use geometric yaw rotation adjustment given the yaw and position readings of the VR headset to find the head orientation of the user relative to the interviewers in the VR environment. The system enables the exploration of whether backchannels trigger an attention shift, or joint attention, among other gaze and head orientation analyses. The VR application can enable people to practice answering common interview questions while also practicing social skills of eye contact and sharing attention among conversational partners.

Keywords: Social Behavior, Social Modulation, Job interview Practice.

Index Terms: Human-centered computing—Human computer interaction (HCI)—Virtual reality; User studies

1 INTRODUCTION

The immersive and interactive nature of virtual reality (VR) allow users to engage in realistic simulated environments, providing a unique opportunity for experiential learning. Users can learn new skills, repeatedly practice complex tasks, and immerse themselves in realistic simulations that resemble real-world environments.

In social sciences, VR tools that can analyze social behavior like gaze, head rotation and gesturing can assist researchers in understanding communication patterns, and social dynamics. VR allows for the practice of social skills in a controlled environment. Users can engage in hands-on learning, making mistakes without fear of repercussions in the real world. This lets people experience specific interactions that might otherwise be difficult to access. In this study, we present a novel VR-based simulation for triadic (three-way) job interviews. Our design enables a user to interact with two virtual interviewers in a three-way conversation and familiarize with common interview questions. The system analyzes their gaze and head rotation behaviors, as well as how they react to the conversational backchannels (e.g., head nods) given by the interviewers.

Individuals move their gaze in a type of active sensing or to signal another person in social settings. Gaze characteristics may differ based on conversational role, which is called the social modulation of gaze. When listening, individuals tend to make more eye contact than when they are speaking [16, 36]. Head rotation and gaze direction tend to align heavily with each other ($\sim$70% of the time) [31]. Given the strong connection between gaze and head rotation, a similar social modulation of head orientation can be expected. Like gaze and head rotation, backchannels, either verbal (e.g., "yes", "uh-huh") or nonverbal (e.g., head nods, shakes, and tilts), are crucial to conversations as they bridge the conversationalists.

People's gaze behavior and head orientation tendencies and the way they react to backchannels can provide insight into whether they tend to perform effective joint attention and whether or not they socially exclude conversationalists. Joint attention is defined as the intentional coordination of an individual's focus of attention with that of another person, resulting in paying attention to the same item for social reasons. This skill is especially important during the early stages of development which reinforces later stages of development as well [35]. An individual's attention direction can be estimated based on their gaze if their face is visible, and otherwise, their head or body orientation can provide some information [28]. In the case of multiple listeners, speakers can avoid excluding the listeners by gazing at them or orienting towards them for a few seconds [30]. This includes acknowledging their backchannels.

Effective usage of gaze and head orientation can increase the chances of obtaining and retaining employment. Applicants who looked straight ahead rather than down during job interviews were perceived to be more confident and dependable and were more likely to be hired [2]. In other early work, researchers connected higher perceived self-confidence and competence levels among interviewees to the effective usage of nonverbal cues such as eye contact and head orientation [12, 27]. In more recent studies, higher interview scores were reported when the participants gazed more at the interviewer, with minimal looking around [34], and employment was more likely in case of effective head orientations towards conversational partners [14].

This study makes two main contributions: (1) Although gaze behavior is widely studied, conversational engagement through head rotation is not. We are the first to create a social modulation analysis of these behaviors in a realistic mock job interview using a portable VR headset. Since the behavior analysis is fully automated, it allows for inexpensive solo practice. (2) While job interview practice tools involving one virtual interviewer exist, the system presented in this paper constitutes the first virtual job interview with two interviewers, which allows for novel behavioral analyses such as mirroring behavior and attention sharing.

2 RELATED WORK

2.1 Gaze Behavior Analysis

Eye tracking is widely integrated with VR head-mounted displays (HMDs). Wang *et al.* [37] investigated visual attention allocation in VR driving simulators. The authors of [11] proposed a machine learning (ML) approach to predict visual attention in VR environments. Their work demonstrated the potential of leveraging large-scale gaze datasets and ML algorithms for gaze analysis in VR.

Current eye-tracking technologies, especially those integrated into HMDs, face challenges related to calibration and data quality [22,25]. Wearing glasses, contact lenses, or mascara, as well as physiological

*e-mail: sartiran@ucsd.edu
†e-mail: pbedmutha@ucsd.edu
‡e-mail: a1li@ucsd.edu
§e-mail: pcosman@ucsd.edu

factors such as eye color may hinder gaze tracking, potentially necessitating repeated calibration or impacting the reliability of data. Researchers have developed algorithms to mitigate the problems of calibration drift, headset slippage, and blinking [5, 20, 38].

2.2 Head Orientation Analysis

Researchers have used VR headsets to investigate head movements and their roles in perceiving and interacting with virtual scenes [8, 23]. Xiao *et al.* [40] used signal processing to identify which head motion patterns are performed by interlocutors to enact acceptance or blame in dyadic (two-person) interactions. In [4], researchers devised an ML-based model to estimate conversational engagement levels based on head gestures that were detected using the measurements of an augmented reality (AR) headset.

VR technology for head orientation analysis has proven valuable in various domains, including social interactions, and training simulations [17, 39]. Researchers have examined how individuals naturally use head movements during social interactions in immersive virtual environments [26]. These studies showed that when used effectively, head rotations can make interlocutors feel included.

2.3 Social Behavior Analysis in Triadic Conversations

Dyadic interactions are unable to capture important behaviors such as social exclusion. To address this, researchers have turned to triadic conversations. Most such studies did not use VR [3, 18, 41].

In [15], head movements on their own were documented to slightly undershoot a speaker's position in a VR-based triadic conversation setting as listeners slightly offset their horizontal head rotation angle instead of directly facing the speaker. Using immersive VR and measuring head/hand movements, Miller *et al.* [21] explored synchrony in three-person teams. Synchrony was affected by the context of the virtual environment and gaze behaviors. In another effort investigating synchrony within triads, Tarr *et al.* [33] built a setup where participants, represented by virtual agents, partook in a joint movement activity with two other participants.

2.4 Behavioral Mirroring and Joint Attention

Unintentional behavioral mirroring refers to involuntarily matching another interlocutor's behaviors and mannerisms in a social setting. In an early study that did not use VR [10], the authors explored how unconscious mirroring can facilitate smooth interactions and increase liking between individuals. Results pointed to a direct proportion between the level of mirroring and the self-perceived rapport. Novotny *et al.* [24] investigated the influence of mirroring in interviews. After the initial interviews, the participants who paired up with interviewers who mirrored their behaviors were more willing to discuss further information compared to the control group (where the interviewers deliberately avoided mirroring).

Aburumman *et al.* [1] examined how head gestures performed by virtual interviewers affected the trust and liking towards them. Head gestures that seemed to be driven by naturalistic behavior rules led to better synchronization and a stronger sense of mutual understanding. Hence, having virtual interviewers that can give realistic backchannels might result in more immersive and realistic VR-based interview experiences. VR technologies have also enabled groups with social communication differences such as autistic school-aged children to practice joint attention, engage in more normative eye contact patterns and initiate a greater number of interactions [29].

2.5 Job Interview Practice

Tools have been developed to support the practice of social and professional skills in job interviews. Strickland *et al.* [32] designed a job interview practice suite that comprised multiple approaches, including VR simulations. In other studies, participants performed better after VR-based job interview practices [6, 9]. Additionally, lower anxiety and higher self-confidence levels were reported after VR-based interview practices [6].

In summary, previous research has demonstrated the potential of VR for studying human behavior and social interactions. VR simulations of job interviews can let individuals tailor their social communication skills to increase their employment chances while reducing the potential costs of having that same practice delivered by a human coach [7].

3 Methods

In this section, we describe our VR-based mock job interview application (which was based on our design in [5]), which allows users to practice for job interviews with multiple interviewers and analyzes users' gaze and head rotation tendencies. We describe the simple voice activity detection (VAD) algorithm we used and how it connects to the timing of backchannels for the virtual interviewers. We also present our geometric yaw rotation adjustment approach that calculates a user's head rotation angle around a predefined fixed vertical axis based on the location of the VR headset and its rotation.

3.1 VR Job Interview

Users wore an HTC Vive Pro Eye VR headset to visualize the virtual office space, comprising two virtual interviewers along with common office objects (Fig. 1a). The 3D coordinates and angular velocities (pitch, yaw, roll) of the headset in each time step are tracked using Unity. The eye tracking uses the built-in tracker of the VR headset, accessible through Unity using the Vive SRanipal SDK. During the interview, at each time step, the system collects gaze origin and direction and finds the virtual object that intersects with the gaze ray. The system records the object label, intersection location, and time spent in each time step which is not fixed across time steps. The interviewers' faces were divided into forehead, eyes, and mouth (Fig. 1b). After each session, the total time and percentage spent looking at each face region are obtained. Gaze data is processed using the algorithm from [5] which improves eye tracking accuracy.

The job interview was for a video game company as video gaming is a common interest of young adults, and because neurodivergent individuals are strongly represented in the video game industry. More generally, the simulation is for people with a STEM background who might seek employment in tech. The interview has 43 questions, including both general questions ("Can you tell me about your strengths?"), and more specific ones ("Can you think of some ways that games can promote positive play?"). The virtual interviewers ask questions one by one, not referring back to previous questions. The users are unscripted, responding as they wish, including pauses for thinking. To signal that they fully answered a question, the users use the controller's trigger button. If needed, a question could be repeated by pressing the controller's trackpad.

Each interviewer asks roughly half of the questions; the assignments, fixed across subjects, were determined to make the conversations realistic. For example, one interviewer might ask multiple

(a) (b)

Figure 1: (a) Virtual office space, (b) Interviewer face regions.

questions in a row during which the other interviewer is mostly silent, except for backchannels. Using the Live Link Face app, the voices and facial animations of two individuals were recorded while they read predefined job interview lines to a camera. The recorded data was processed in Blender and added on head models as key points. Photos of the individuals were used to create facial textures. The head models included head nodding and shaking animations, and small movements such as head tilts, displayed when the users talk, so the interviewers look interested. In addition to the physical animations, the interviewers provide verbal backchannels such as "uh-huh" and "hmm". An interviewer turns their head towards the other when the other asks a question. They can also turn their head when the other interviewer gives a backchannel while the user is speaking. These behaviors allow the virtual interviewers to react to each other's behaviors in a realistic way. Other than backchannels and head turns, the interviewers do not engage in non-verbal communication such as arm movements.

3.2 Voice Activity and Backchannel Timing

Our system tracks voice activity to detect the starting point of an interviewee's response since we do not want, for example, an interviewer to say "uh-huh" before the user has even started their answer. At each time point, we compute the root mean squared (RMS) value of 48,000 audio samples to quantify the average speech signal magnitude. We apply a threshold to a Gaussian weighted window of RMS values to make a decision on voice activity. We have 3 parameters: threshold t_{RMS}, window size w, and standard deviation (s.d.) of the Gaussian distribution σ. The algorithm detects voice activity when the weighted average RMS in a window is greater than t_{RMS}.

To validate and tune our VAD algorithm, we had 2 individuals (1m/1f) wear the headset and read 3 scripts displayed in VR with and without additional background noise. To generate background noise, we played office sounds on YouTube. Each recording took around 3 minutes and we annotated the time intervals where the reader was silent for longer than 0.5s. The optimal parameter set ($\hat{t_{RMS}}$, $\hat{w}$, $\hat{\sigma}$) = (0.07, 29, 18) maximized the sum of true positive rate, true negative rate, and intersection over union value (94.1%, 91.3%, and 91.4%). As our application runs at about 45 time points per second, a window of $\hat{w}$=29 corresponds to 0.64s, using 0.32s of data from the past and from the future. Hence, when a user starts speaking, the starting point of their speech is recognized 0.32s later.

For our application, we had to determine when the virtual interviewers ought to provide a backchannel. We analyzed recordings of 6 in-person 10-minute professional conversations involving an interviewee and two individuals who acted as interviewers to learn the amount of time an interviewer takes to give a verbal backchannel (VB) or a physical backchannel (PB) after the interviewee starts responding. The setup in these recordings was identical to our virtual setup as the actors sat across from the interviewee around an oval table. Average durations between two consecutive VBs and PBs were also recorded. Our findings are in Table 1 where standard deviations are given in parentheses. The times the interviewers waited until giving either backchannel type were similar. Henceforth, we refer to the interviewer who asked the most recent question as *IntQ*, and the other interviewer as *IntO*. On average, *IntQ* gives the first backchannels at t+8.5s (s.d.=2.8s), and *IntO* at t+9.5s (s.d.=3.5s) where t marks the start of the interviewee's response. The average time between two VBs or two PBs is roughly 9s (s.d.=2.4s) for *IntQ*, and 13s (s.d.=5.5s) for *IntO*. The shortest interviewee response that received a backchannel in the recordings was about 2.5s.

In our application, the virtual interviewers provide backchannels randomly based on these timings. Based on the question, after 2.5s of voice activity is detected that started at time t, *IntQ* gives an initial backchannel at a time drawn uniformly randomly from the interval $t + 8.5 \pm 2.8$s, while *IntO* has an interval $t + 9.5 \pm 3.5$s, unless the user pushes the controller button to indicate the end of their answer

Table 1: Time to first backchannel, and backchannel spacing

Type	Avg. Initial Wait Time		Avg. Time In Between	
	IntQ	*IntO*	*IntQ*	*IntO*
Verbal	9.3s (4.6s)	10.3s (3.4s)	9.7s (3.5s)	13.7s (8.6s)
Physical	8.1s (0.9s)	9s (3.6s)	8.3s (1.3s)	12.1s (2.3s)

prior to the backchannel occurrence. For *IntQ*, the duration between consecutive backchannels is drawn from 9 ± 2.4s, and from 13 ± 5.5s for *IntO* (Table 1).

With this design, we can explore the effects of interviewer backchannels. We consider four cases: *QbOl*: *IntQ* may give backchannels, *IntO* solely listens, *QlOb*: *IntQ* solely listens, *IntO* may give backchannels, *QbOb*: both interviewers may give backchannels, and *QtOb*: *IntQ* turns his head towards *IntO* when *IntO* gives a backchannel, but does not give backchannels to the user. For 5 questions that were known from [5] to have consistently brief subject responses (less than the average initial backchannel times in Table 1), the virtual interviewers did not give backchannels. Ignoring these questions, the remaining 38 questions were assigned with 10 each to the first two cases, and 9 each to the other two, where questions with answers of different lengths were distributed evenly across the cases. These assignments determined the virtual interviewer behavior and were fixed across participants. We did not consider the opposite of *QtOb* (*QbOt*) as we anticipated the users to mostly gaze at and face *IntQ*, irrespective of their conversational role, which reduces the chances of a reaction to *IntO*'s head turns.

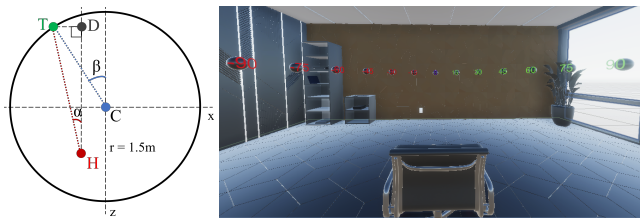
3.3 Geometric Yaw Rotation Adjustment

We collect information about head movements by tracking the headset object's position. The headset object is defined as a child object of the camera rig object which determines the starting location and orientation of the user within the application. When a child object rotates, the rotations are computed with respect to its parent object's axes, given that the parent object does not rotate. In our application, the camera rig object never moves or rotates. When we track the headset's position/rotation, we track position changes/rotations with respect to the camera rig, and its axes. Then, irrespective of headset location, a yaw angle of 0° corresponds to facing straight ahead. However, a target that is placed at 45° with respect to the center of the camera rig will not be at 45° for a shifted headset location. Since we aim to explore the social inclusion tendencies of individuals and analyze their head turns towards the interviewers, the targets in our application are the interviewers.

The virtual interviewer positions in the virtual setting are determined by their 3D-center location. The distance from the center of the camera rig to the centers of the interviewers is 1.5m. In Fig. 1a, the right-hand side of the scene represents positive yaw angles, from the perspective of someone sitting in the empty chair. The interviewer on the left side of Fig. 1a (*Int*-1) is centered at -22.5° (spanning $-22.5° \pm 10°$), and the other one (*Int*-2) at 22.5° (spanning $22.5° \pm 9°$) with respect to the z (forward) axis of the camera rig.

We aim to study orienting towards the virtual interviewers. Ray casting is commonly used for 3D target acquisition and it could have been useful if the interviewers were blocked by objects. However, since there is a clear path between the user and the interviewers, knowing the rotation and location of the headset with respect to the fixed origin suffices to geometrically find engagement with (orientation towards) our targets.

Pitch, yaw, and roll angles describe the rotation of an object around the horizontal, vertical, and forward axis. Orienting towards a virtual interviewer is explained by the yaw values. Our algorithm uses the (H_x, H_z) location of the headset with respect to the camera rig object and the measured yaw angle α to compute the yaw rotation β that corresponds to facing the same location from the center of the camera rig. The relation between these quantities for one of the



(a) (b)

Figure 2: (a) C, H, and T represent the center of the camera rig, headset, and target location that the user is facing. D is the projection of T on the forward z axis of the headset. α is the yaw angle measured by the headset. β is the output yaw angle of the algorithm. x-axis is the horizontal motion axis in Unity, (b) Head rotation data collection setup.

six cases given in Fig. 2a is:

$$\begin{aligned} &tan\alpha = |DT|/|DH| = (r \cdot sin|\beta| - |H_x|)/(r \cdot cos|\beta| + |H_z|) \\ &= r \cdot sin(|\beta| - |\alpha|) = sin|\alpha| \cdot |H_z| + cos|\alpha| \cdot |H_x| \\ &=> \beta = \alpha - sin^{-1}((sin|\alpha| \cdot |H_z| + cos|\alpha| \cdot |H_x|)/r) \end{aligned} \quad (1)$$

The remaining cases are similar. To validate this approach, 9 subjects (8m/1f) oriented their heads towards 13 virtual spheres placed horizontally 1.5m away from the center of the camera rig. The spheres were placed at angles between -90° and 90°, in increments of 15° (Fig. 2b). The subjects oriented towards each sphere in a random order as instructed by an experimenter who tracked the timestamp and sphere corresponding to each command. Head rotation data was collected from 9 different positions: left, center, right, back-left, back, back-right, front-left, front, and front-right. The subjects were seated at the center of the camera rig for the center position. The other positions correspond to being seated half a meter away from the center position in the given direction. All of the spheres were targeted from each position.

For the center position, no yaw adjustments were necessary and the mean absolute difference (MAD) between the recorded yaw angles (α) and the ground truth values (β^{GT}) can be regarded as a baseline for our algorithm. Here, β^{GT} is the horizontal angle with respect to the center of the camera rig at which the sphere was located (whichever sphere the subject was instructed to face). For the center position, MAD = $(\sum_{i=1}^{S} |\beta_i^{GT} - \alpha_i|)/S$, where S is the number of spheres that were faced from a position. All 13 spheres were oriented towards, but sometimes $S > 13$ because of potential instruction repetitions. The center MAD averaged over all subjects was 3°. This small error is due to the fact that subjects cannot orient exactly towards the spheres in this VR task.

For positions other than the center one, MAD = $(\sum_{i=1}^{S} |\beta_i^{GT} - \beta_i|)/S$. Table 2 presents the MAD values before and after the algorithm adjusts the recorded yaw values. Based on Table 2, the MADs between the processed yaw angles and the true values show significant improvement for all positions. Additionally, the average MAD achieved by our algorithm over all positions is 3°, which is equal to the baseline, suggesting the algorithm does not introduce any additional error beyond the error of orienting towards a sphere.

Table 2: MAD Values Before/After Processing the Yaw Values

Case:	Left	Right	Back-Left	Back	Back-Right	Front-Left	Front	Front-Right
Diff. Before Algorithm (°):	12.2	13.2	15.4	10.7	16.4	20.9	12.6	19.7
Diff. After Algorithm (°):	2.6	2.7	3.2	3.1	3.4	3	3.2	2.7

4 Results

To examine our VR job interview simulation's capabilities as a gaze and head rotation analysis tool for triadic conversations, five individuals (4m/1f graduate students; mean age=26.6 (s.d.=2.6)) did the simulation. On average, a session took 22 minutes. We aimed to see how subjects divided their attention while listening and responding to a question, using gaze and head rotation, and to investigate the impact of backchannels given by the virtual interviewers.

4.1 Gaze and Head Rotation Results

Fig. 3 shows the social modulation of the subjects' gaze. On average, subjects gazed at the interviewers' faces more when they were listening than speaking (98.3% versus 86%). This comports with the findings in [5, 16, 19, 36]. Regardless of the conversational role, the subjects mainly targeted the interviewers' eyes, specifically *IntQ*'s eyes. Although less than the eyes for both cases, subjects looked at the interviewers' mouths, mostly *IntQ*'s mouth, more when listening compared to speaking. While speaking, although the main focus was *IntQ*'s eyes, the amount of gaze directed towards *IntO*'s eyes increased. The amount of gaze at the foreheads was negligible.

We investigated the effect of conversational role on the region the subjects faced (Fig. 4). We defined 5 regions for the yaw angles: *IntQ/IntO*: -22.5°±10° or 22.5°±9° based on which interviewer asked the most recent question, *Interior-Q/Interior-O*: [-12.5°, 0°] or [0°, 13.5°] based on who *IntQ* is, and *Exterior:* [-180°, -32.5°] and [31.5°, 180°]. The subjects did not turn to the latter at all, so Fig. 4 does not include this region and we drop it from further discussion. As listeners and speakers, the subjects faced the interviewers more than the other regions. They primarily dwelt around *IntQ* when listening. As speakers, they turned to *IntO* more which was also more than the amount of time they spent in *Interior-Q*. The *Interior-O* region was consistently targeted the least.

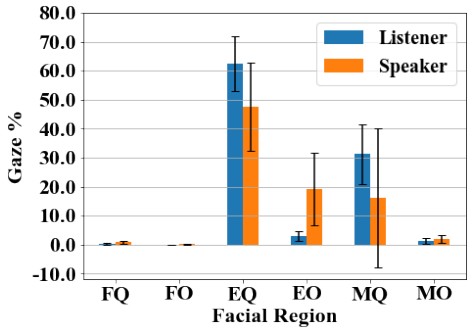


Figure 3: Social modulation of gaze behavior based on interviewer roles. F, E, M, Q, O denote forehead, eyes, mouth, *IntQ*, and *IntO*.

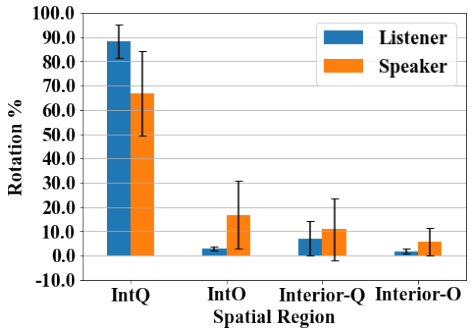


Figure 4: Social modulation of head rotation based on interviewer roles.

We also checked if the subjects acted differently towards the interviewers. On average, while they were responding to *Int*-1's questions, the yaw angles were negative 80% of the time. The yaw values were positive 72% of the time for *Int*-2's questions. Since this difference is not large, we conclude that appearances and behaviors of the interviewers did not affect subject behavior.

4.2 Effects of Backchannels

As introduced in Section 3.1, this VR application can also be used to analyze the effects of backchannels produced by the virtual interviewers. The four cases we investigated were: *QbOl*, *QlOb*, *QbOb*, and *QtOb*, in which *Q* and *O* denote *IntQ* and *IntO*, and b, l, t indicate whether or not that interviewer can give backchannels, or turns to the other interviewer when the other gives a backchannel. The metric we used for this analysis was the maximum absolute yaw angle change triggered by each interviewer backchannel/head turn, averaged over the total number of backchannels/head turns for that case, across all 5 subjects. Overall, 69, 67, and 37 backchannels, and 50 head turns were executed for the four cases, respectively. Backchannels and head turns that the interviewers perform are animations with known durations. The average lengths of the animations for VBs, PBs, and head turns during the five interview sessions were 3s, 6s, and 5s, respectively. These lengths include the actual backchannel duration plus a buffer to give users time to react.

We are interested in the maximum absolute difference between β recorded at the beginning of a backchannel/head turn and at a time point while that backchannel/head turn is being performed. For the 5 questions where the interviewers did not give any backchannels, the average maximum unsigned yaw change was 6.2°. The maximum unsigned yaw shifts on average for the four cases given above were 11.5°, 12.7°, 13°, and 23.4° (Fig. 5). *QtOb* triggered the biggest reactions, meaning that when *IntQ* turns towards *IntO* following a backchannel by *IntO*, the subjects mirrored *IntQ*'s head turn to some extent, to perform joint attention. The fact that the maximum unsigned yaw shift values were larger in the presence of interviewer backchannels compared to the short answer cases which had no backchannels suggests that subjects adjusted their behaviors to interact with both interviewers when they received attention from them in the form of conversational backchannels.

The VR-based job interview simulation successfully helped us to identify an outlier subject behavior. For this subject, the maximum yaw shifts for the four cases given above were 2.4°, 2°, 2.7°, and 5.4° (all significantly smaller than the average; Fig. 5). Additionally, this subject gazed at *IntO*'s face for 4.7% of the time when speaking (25.4% on average among the other subjects). Excluding the outlier subject further amplified some of our findings. Compared to the previous 68.5%, the subjects gazed at the interviewers' eyes for 76.7% of the time when speaking, widening the gap between the eye contact percentages for the two conversational roles. For the speaking role, the percentage of gaze pointed at *IntQ*'s mouth dropped from 16.2% to 4.2% (s.d. from 24% to 2.1%), shifting to both interviewers' eyes. Furthermore, the first three values shown in Fig. 5 increased by around 2.5%, whereas the last one by 4.5%, making the margin between the reaction to the last case and the rest even more apparent. These observations show that the outlier subject tended to hyper-focus on *IntQ* with little to no reaction to *IntO*'s backchannels or *IntQ*'s head turns. They also gazed at *IntQ*'s mouth considerably more than the rest of the subjects.

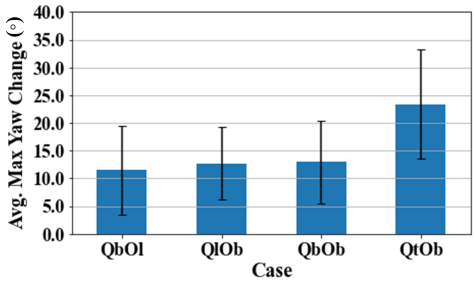


Figure 5: Maximum unsigned yaw shifts for each interviewer backchannel behavior case, averaged across all subjects.

4.3 User Experience Survey Results

After the simulations, the subjects were asked to rate the repeatability, complexity, user-friendliness, and realism of the application using a 5-point Likert scale. The average scores (with s.d.s in parentheses) for the four aspects were 4.2 (1.2), 2.2 (0.9), 4.6 (0.6), and 4.3 (0.8), respectively. High repeatability, realism, and user-friendliness scores, and low complexity scores point to the suitability of our VR job interview simulation as a self-deliverable practice tool that can also let them know whether they tend to hyper-focus on a conversationalist. The open-ended questions asked about what the subjects liked the most, and what could be improved about the design of the system and user experience. The subjects liked the realism of the interviewers, and their behaviors not being robotic. They thought the interviewers should be able to talk to each other or show more emotion, and wished the headset was more comfortable, and they could take breaks.

5 Discussion and Conclusion

Turning towards people and making eye contact with them can make people feel included in a conversation. In professional settings such as job interviews, candidates who use gaze and head movements to optimally address all interviewers can increase their chances of employment [14, 34]. In this study, we presented a VR-based triadic job interview simulation that enables users to get familiar with common interview questions while also informing themselves about how they spread their attention through gaze and head rotations.

Some of the behaviors have been examined in prior studies on screen-based or in-person settings [13, 16, 19, 36] or in immersive environments [5]. However, most of the behaviors studied here have not previously been examined in the context of VR, whether screen-based or immersive. Specifically, we are studying attention sharing (gaze and head rotations) in triadic conversations, in which a subject has to divide their attention between two conversational partners, which is not a commonly studied topic using VR.

Whether they were listeners or speakers, our participants primarily faced whichever interviewer asked the most recent question, and gazed at their eyes. These findings on social modulation were in agreement with earlier work which was almost exclusively not in VR [16, 19, 36] showing that our system can extend these analyses to the context of immersive VR and can provide them in a fully automated way. Our novel analysis of backchannel effects showed that when the interviewer who asked the most recent question turned towards the other interviewer (when that interviewer gave a backchannel), participants also turned their head towards the other interviewer, a kind of mirroring or joint attention behavior. The participants reacted more mildly when the interviewers performed backchannels together or separately, without any head turns.

The participants favored the repeatability, user-friendliness, and realism of our design. One of the long-term goals of the system is to allow users to practice answering common job interview questions, while also practicing and receiving feedback on communication skills. It can be difficult to share attention among multiple listeners, and make them all feel addressed. This VR application would be able to help people practice such attention-sharing skills.

6 Limitations and Future Work

This study is limited by the relatively small number of participants, majority of whom was male. In our future iterations, we will expand the participant set to prevent potential behavioral biases. We also intend to create different question sets, so that users can redo the interviews without encountering identical questions. This would also allow for observing how gaze and head orientation changes across interview sessions. We will also explore user reactions to a situation where neither interviewer gives any backchannels even when the user speaks for an extended duration. Lastly, we intend to create a mechanism to provide simple feedback to the users on their social inclusion behavior, so that users who exclude an interviewer can be made more aware and may choose to practice social inclusion.

Acknowledgments

We wish to thank Onur Tepencelik for the recordings we analyzed for interviewer backchannel timing, and all of our colleagues for helpful discussions. We would also like to thank our participants. This work was supported by the National Science Foundation under grant DUE-1928604.

References

[1] N. Aburumman, M. Gillies, et al. Nonverbal communication in virtual reality: Nodding as a social signal in virtual interactions. *Intl. J. of Human-Computer Studies*, 164:102819, 2022.

[2] J. G. Amalfitano and N. C. Kalt. Effects of eye contact on the evaluation of job applicants. *J. of Employment Counseling*, 14(1):46–48, 1977.

[3] A. D. Arndt et al. Who do I look at? Mutual gaze in triadic sales encounters. *J. of Business Research*, 111:91–101, 2020.

[4] S. Artiran, L. Chukoskie, et al. HMM-based detection of head nods to evaluate conversational engagement from head motion data. In *29th European Signal Processing Conference*, pp. 1301–1305. IEEE, 2021.

[5] S. Artiran et al. Measuring social modulation of gaze in autism spectrum condition with virtual reality interviews. *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, 30:2373–2384, 2022.

[6] R. M. Aysina, Z. A. Maksimenko, and M. V. Nikiforov. Feasibility and efficacy of job interview simulation training for long-term unemployed individuals. *PsychNology Journal*, 14(1):41–60, 2016.

[7] J. Bersin. The corporate learning factbook 2014: Benchmarks, trends, and analysis of the US training market, 2014.

[8] J. Brookes, M. Warburton, M. Alghadier, et al. Studying human behavior with virtual reality: The Unity experiment framework. *Behavior Research Methods*, 52:455–463, 2020.

[9] B. Chang, J.-T. Lee, Y.-Y. Chen, and F.-Y. Yu. Applying role reversal strategy to conduct the virtual job interview: A practice in second life immersive environment. In *4th Intl. Conference on Digital Game And Intelligent Toy Enhanced Learning*, pp. 177–181. IEEE, 2012.

[10] T. L. Chartrand and J. A. Bargh. The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology*, 76(6):893, 1999.

[11] F. Fathy, Y. Mansour, et al. Virtual reality and machine learning for predicting visual attention in a daylit exhibition space: A proof of concept. *Ain Shams Engineering J.*, 14(6):102098, 2023.

[12] R. J. Forbes and P. R. Jackson. Non-verbal behaviour and the outcome of selection interviews. *Journal of Occupational Psychology*, 53(1):65–72, 1980.

[13] M. Freeth and P. Bugembe. Social partner gaze direction and conversational phase; dactors affecting social attention during face-to-face conversations in autistic adults? *Autism*, 23(2):503–513, 2 2019.

[14] M. C. Hatch. Candidates' head and eye orientations in job interviews: Effects on impression formation. 2022.

[15] L. Hladek and B. U. Seeber. Behavior in triadic conversations in conditions with varying positions of noise distractors. In *Fortschritte der Akustik–DAGA'23*, pp. 916–918, 2023.

[16] A. Kendon. Some functions of gaze direction in social interaction. *Acta Psychologica*, 26:22–63, 1967.

[17] S. Kloiber, V. Settgast, C. Schinko, et al. Immersive analysis of user motion in VR applications. *The Visual Computer*, 36:1937–1949, 2020.

[18] H. Lu, M. F. McKinney, T. Zhang, and A. J. Oxenham. Investigating age, hearing loss, and background noise effects on speaker-targeted head and eye movements in three-way conversations. *The Journal of the Acoustical Society of America*, 149(3):1889–1900, 2021.

[19] H. Mansour and G. Kuhn. Studying "natural" eye movements in an "unnatural" social environment: The influence of social activity, framing, and sub-clinical traits on gaze aversion. *Quarterly Journal of Experimental Psychology*, 72(8):1913–1925, 2019.

[20] R. Mantiuk, B. Bazyluk, and R. K. Mantiuk. Gaze-driven object tracking for real time rendering. In *Computer Graphics Forum*, vol. 32, pp. 163–173. Wiley Online Library, 2013.

[21] M. R. Miller et al. Synchrony within triads using virtual reality. *Proc. of the ACM on Human-Computer Interaction*, 5(CSCW2):1–27, 2021.

[22] J. D. Morgante et al. A critical test of temporal and spatial accuracy of the Tobii T60XL eye tracker. *Infancy*, 17(1):9–32, 1 2012.

[23] J. R. J. Neo, A. S. Won, and M. M. Shepley. Designing immersive virtual environments for human behavior research. *Frontiers in Virtual Reality*, 2:603750, 2021.

[24] E. Novotny, M. G. Frank, and M. Grizzard. A laboratory study comparing the effectiveness of verbal and nonverbal rapport-building techniques in interviews. *Communication Studies*, 72(5):819–833, 2021.

[25] M. Nystrom, R. Andersson, K. Holmqvist, and J. Van De Weijer. The influence of calibration method and eye physiology on eyetracking data quality. *Behavior Research Methods*, 45(1):272–288, 2013.

[26] X. Pan and A. F. d. C. Hamilton. Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape. *British J. of Psychology*, 109(3):395–417, 2018.

[27] C. K. Parsons and R. C. Liden. Interviewer perceptions of applicant qualifications: A multivariate field study of demographic characteristics and nonverbal cues. *Journal of Applied Psychology*, 69(4):557, 1984.

[28] D. I. Perrett et al. Organization and functions of cells responsive to faces in the temporal cortex. *Philosophical Trans. of the Royal Society of London. Series B: Biological sciences*, 335(1273):23–30, 1992.

[29] V. Ravindran, M. Osgood, et al. Virtual reality support for joint attention using the floreo joint attention module: Usability and feasibility pilot study. *JMIR Pediatrics and Parenting*, 2(2):e14429, 2019.

[30] W. R. Steele. *Presentation skills 201: how to take it to the next level as a confident, engaging presenter*. Outskirts Press, 2009.

[31] R. Stiefelhagen and J. Zhu. Head orientation and gaze direction in meetings. In *CHI'02 Extended Abstracts on Human Factors in Computing Systems*, pp. 858–859, 2002.

[32] D. C. Strickland, C. D. Coles, and L. B. Southern. JobTIPS: A transition to employment program for individuals with autism spectrum disorders. *J. of Autism and Developmental Disorders*, 43(10):2472–2483, 2013.

[33] B. Tarr, M. Slater, and E. Cohen. Synchrony and social connection in immersive virtual reality. *Scientific Reports*, 8(1):3693, 2018.

[34] F. Tian, S. Okada, and K. Nitta. Analyzing eye movements in interview communication with virtual reality agents. In *Proceedings of the 7th International Conference on Human-Agent Interaction*, pp. 3–10, 2019.

[35] P. Venuprasad, T. Dobhal, A. Paul, T. N. Nguyen, A. Gilman, et al. Characterizing joint attention behavior during real world interactions using automated object and gaze detection. In *Proc. of the 11th ACM Symposium on Eye Tracking Research & Applications*, pp. 1–8, 2019.

[36] R. Vertegaal et al. Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, pp. 301–308, 2001.

[37] Z. Wang et al. Relationship between gaze behavior and steering performance for driver–automation shared control: A driving simulator study. *IEEE Trans. on Intelligent Vehicles*, 4(1):154–166, 2018.

[38] M. Weier, T. Roth, et al. Predicting the gaze depth in head-mounted displays using multiple feature regression. In *Proc. of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pp. 1–9, 2018.

[39] A. S. Won, J. N. Bailenson, S. C. Stathatos, and W. Dai. Automatically detected nonverbal behavior predicts creativity in collaborating dyads. *Journal of Nonverbal Behavior*, 38:389–408, 2014.

[40] B. Xiao, P. Georgiou, B. Baucom, and S. S. Narayanan. Head motion modeling for human behavior analysis in dyadic interaction. *IEEE Transactions on Multimedia*, 17(7):1107–1119, 2015.

[41] E. Zima, C. Weiß, and G. Brône. Gaze and overlap resolution in triadic interactions. *Journal of Pragmatics*, 140:49–69, 2019.