An Evaluation of Other-Avatar Facial Animation Methods for Social VR

Peter Kullmann
peter.kullmann@uni-wuerzburg.de
Julius-Maximilians-Universität
Würzburg, Germany

Timo Menzel
timo.menzel@tu-dortmund.de
TU Dortmund University
Dortmund, Germany

Mario Botsch
mario.botsch@tu-dortmund.de
TU Dortmund University
Dortmund, Germany

Marc Erich Latoschik
marc.latoschik@uni-wuerzburg.de
Julius-Maximilians-Universität
Würzburg, Germany

Figure 1: *Virtual character addresses observer.* Screenshots show, left to right, static face, synthesized expression, and tracked expression. In our study, animation was either in sync with audio or delayed by the animation system's inherent latency.

ABSTRACT

We report a mixed-design study on the effect of facial animation method (static, synthesized, or tracked expressions) and its synchronization to speaker audio (in sync or delayed by the method's inherent latency) on an avatar's perceived naturalness and plausibility. We created a virtual human for an actor and recorded his spontaneous half-minute responses to conversation prompts. As a simulated immersive interaction, 44 participants unfamiliar with the actor observed and rated performances rendered with the avatar, each with the different facial animation methods. Half of them observed performances in sync and the others with the animation method's latency. Results show audio synchronization did not influence ratings and static faces were rated less natural and less plausible than animated faces. Notably, synthesized expressions were rated as more natural and more plausible than tracked expressions. Moreover, ratings of verbal behavior naturalness differed in the same way. We discuss implications of these results for avatar-mediated communication.

CHI EA '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3585617

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **Virtual reality**.

KEYWORDS

Facial Animation, Facial Expression Tracking, Facial Expression Synthesis, Naturalness, Plausibility, Observation Study, Virtual Reality, Performance Capture

ACM Reference Format:
Peter Kullmann, Timo Menzel, Mario Botsch, and Marc Erich Latoschik. 2023. An Evaluation of Other-Avatar Facial Animation Methods for Social VR. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3544549.3585617

1 INTRODUCTION

We use language with bodily motion to express our thoughts, ideas, and internal states. This gesture-speech unity plays a crucial role in conveying information, both with and without intention [40]. Facial expressions play a prominent role among other nonverbal behaviors, such as posture, proxemics, and paralinguistics. They also contribute to verbal behavior and are visible in most situations. Human face movements have been investigated in depth, often using the Facial Action Coding System (FACS) to deconstruct expressions into their underlying components based on anatomical movements [6]. For example, lip movement fosters speech comprehension [20],

and expressions on the level of micro-facial movements indicate deception [19]. Human eye morphology even facilitates perceiving gaze direction[14]. Gaze direction is used as input to detect information about our environment and as output to signal attention [9] and regulate interaction [13].

The importance of facial expressions carries over to how (virtual) humans are perceived in virtual, augmented, and mixed reality scenarios (VR, AR, MR - XR for short). Here, virtual humans are usually classified depending on agency as either avatars (controlled by human input) or embodied agents (controlled algorithmically)[3]. In immersive interpersonal communication, like in physical reality, faces are at the center of attention and crucial to conversation results [24, 30]. Behavioral and visual fidelity of virtual humans has shown to be important to XR experiences' realism, plausibility, and presence and incorporated into several models thereof [15, 34, 35].

Recent progress in 3D reconstruction methods allows affordable creation of virtual humans [4], even allowing to attune their facial expressiveness to personal idiosyncrasies [21]. While this high visual fidelity can be achieved with consumer hardware, behavioral fidelity in XR setups is most commonly restricted to microphone input and tracking three devices in 3D space (typically two hand controllers and a headset). Newer headsets provide facial expression tracking, and existing ones can be modified with external hardware to do so. As an alternative, facial expressions can also be synthesized from audio and head movements. To decide which approach to follow, it is important to compare the effect of tracking expressions to synthesizing them on naturalness and plausibility. More natural and more plausible nonverbal behavior could improve the experience of interpersonal communication in XR. It is also relevant to compare facial animation methods with and without their inherent latency.

To investigate this, we compare the perception of facial expressions captured with ARKit, Apple's face tracking solution, to expressions created with the Oculus Avatar SDK, Meta's widely used system for facial expression synthesis based on speaker audio, head movement, and tagged gaze targets, to a baseline of a static face. We let participants observe an avatar's short performances in VR and rate them in terms of naturalness and plausibility. Performances are either shown with their original latency or in-sync with audio.

2 RELATED WORK

Several approaches have been used to make virtual faces come to life that we roughly divide into either expression tracking or expression synthesis. Exemplary works for synthesized expressions include animated eyes being preferred over static eyes when viewing oneself in a virtual mirror [5] and veridical gaze preferred over synthesized gaze in a dyadic avatar-mediated interaction [31]. Gonzalez-Franco and colleagues showed increased self-identification with self-avatars' pre-baked animations [10]. Murcia-López and colleagues let participants select animation parameters for stylistic characters that are included in the Oculus Avatar SDK [22]. It activates blendshapes with FACS-like semantics as follows: lip-sync is created by retrieving phonemes in a temporal convolutional network, eye behavior is based on dynamic saliency-based gaze targeting with blinks about every six seconds or more often during speech and gaze, and ambient micro-expressions are linked to lip sync and eye gaze events. It has also widespread use in commercial applications, e.g. in PokerStars VR, Epic Roller Coasters, or Tribe XR. Other, more data-driven approaches created realistic facial animation from audio, but are not readily integrated to game engines common in XR research [7, 28, 38].

Tracking facial expressions in XR setups is challenging since head-mounted displays obstruct large face areas. Before headsets with built-in facial expression tracking became available, previous works have deployed custom-built hardware or composed existing sensors, commonly fusing lower face tracking sensors with eye-tracking headsets [17, 25, 29, 39].

Apart from different expressions outputs, facial animation methods can differ in processing duration, introducing audiovisual misalignment. Since light travels faster than sound we are not used to quicker audio and are sensitive to even small synchronization errors: perceptual thresholds have been reported from around 80ms [37] to around 180ms [41]. Hence, we can synchronize animations to recorded audio to isolate the effect of expression output irrespective of processing duration.

We decided to compare facial expressions synthesized with the Oculus Avatar SDK to facial expressions tracked with ARKit. Both use FACS-like action unit semantics and provide plugins to forward expression data to game engines.

3 METHOD

Our study followed a 2x3 design with between-groups factor *latency adjustment* (animations in sync with audio vs. delayed by latency inherent to the animation system) and within-groups factor *facial animation method* (face tracking vs. expression synthesis vs. static face). Written approval for the study was obtained from the ethics committee of the Institute for Human-Computer Media (MCM) of the Julius-Maximilians-Universität Würzburg[1].

We hypothesize two effects:

1. *Main effect of facial animation method*: static faces are perceived as less natural and less plausible than animated (synthesized and tracked) faces.

2. *Interaction effect of latency adjustment x facial animation method*: In original latency, the quicker method is perceived as more natural and plausible than the slower method. When latency is adjusted for, both synthesized and tracked facial expressions are rated equally natural and plausible.

3.1 Character Creation

We based our virtual human on a full-body scan using the approach from Achenbach and colleagues [1]: In a custom-built rig of 94 DSLR cameras, multi-view images of our scanning subject were captured to generate a dense point cloud. A template model's pose and shape parameters were then optimized to fit this point cloud.

In a second step, we personalized blendshapes with the automated pipeline of Menzel and colleagues [21].

Finally, minor scanning artifacts in texture and mesh were corrected manually, resulting in a skinned mesh resembling our actor with high visual detail and ready for real-time animation.

[1] https://www.mcm.uni-wuerzburg.de/forschung/ethikkommission/

3.2 Performance Capture

Figure 2: *Performance capture setup.* Actor facing a mannequin as target while equipped with head tracker, microphone, and hand controllers, with face tracker on table.

To initiate unscripted natural behavior, we examined a set of questions intended to be used in a classroom for English lessons [2], similar to Lee and colleagues' conversation prompts [16]. We then picked questions we thought to inspire spontaneous, casual answers about non-intimate topics that neither require too much background knowledge nor reveal personal information, such as "What makes someone a good driver?" or "Are holidays really relaxing? What stressful things are involved in taking a holiday?".

We instructed a trained actor to freely respond to our selected questions as if asked by a recently met acquaintance. Their imagined conversation partner was embodied as a mannequin sat opposite of him at a table (cf. Figure 2). We set a target of about 30 seconds per response, because longer observation times of expressive behavior have not shown to predict interpersonal outcomes better [2].

To record the performer's behavior we equipped him with Valve Index Knuckles tracking hand motions and a HTC Vive tracker to record head movements. We recorded a video and facial expressions with an iPhone 12 Pro that ran a custom-built app. It used official Apple SDKs to record the phone's screen with camera feed and microphone input and the detected facial expressions as 52 FACS-like coefficients together with 3D poses for head, both eyes, and the point currently looked at.

For our virtual reconstruction of the scene, we let the actor raise his arms to a T-pose and probed several landmark positions in the room with an additional HTC Vive Controller: the corners of the table, iPhone camera, the mannequin's chin, forehead, and eyes, and the actor's tip of the nose, nasion, corners of the mouth, and wrists.

3.3 Animation

3.3.1 Coordinate System Alignment. To find the offset between head tracker origin and the skinned mesh's skull joint, we registered virtual face mesh landmarks with corresponding face landmarks probed in the performance capture setup. This was further refined by manually matching the tracker placement on the virtual character in a juxtaposition of face sensor screen capture and the virtual character rendered in a matching perspective.

To match coordinate frames from ARKit and SteamVR, we firstly registered matching vertices from ARKit's face mask with vertices from the full-body mesh with corresponding blendshape weights applied. Secondly, we applied the previous offset from head tracker to virtual skull joint. Finally, we further adjusted the offset so that the gaze rays hit the iPhone camera at the time of recording when the actor truly looked at it.

Similarly, we offset head poses forwarded to the Oculus Avatar SDK to have the SDK template avatar match the actor avatar's eye level and direction.

3.3.2 Body pose. To infer the actor's body pose, we used the VRIK solver from RootMotion's Final IK package[3]: hand and head end effector targets followed respective tracker trajectories with a fixed pelvis target at the seat and elbow bend goals at the armrests. We retrieved offsets between trackers and end effectors from the recorded T-pose.

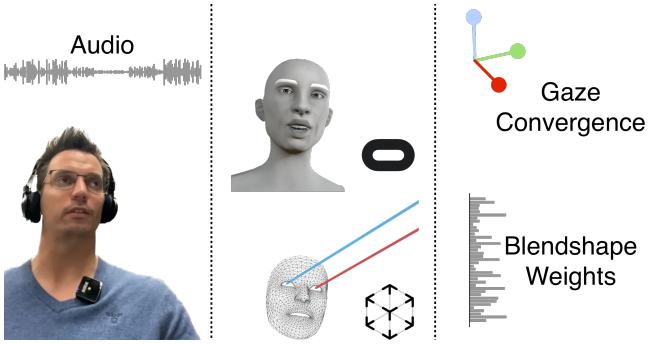

Figure 3: *Facial Animation Pipeline.* Performance (left) is turned into corresponding animation parameters (right) from either audio and tagged gaze targets (Meta Oculus Avatar SDK, top center), or RGBD tracking (Apple ARKit, bottom center). Resulting avatar renders shown in Figure 1.

3.3.3 Facial Expression. Data for both tracked and synthesized facial animation was brought into a shared animation parameter space to drive the virtual actor mesh (cf. Figure 3). We transferred ARKit expressions directly since the mesh had exactly matching blendshapes. Mapping expressions from the Oculus Avatar SDK directly to their semantically matching ARKit blendshapes resulted in different expressiveness. Therefore, for each Oculus Avatar SDK expression, we computed its closest resemblance built from a combination of ARKit blendshapes and baked it as new target shape.

To transfer gaze direction, we rotated the actor avatar's eyes so that their visual axes point towards the look-at point. For tracked expressions we used the reported look-at point, for synthesized expressions we derived the look-at point as the middle of where the Oculus Avatar SDK's avatar gaze rays are closest.

[2]http://teflpedia.com/Teflpedia:Conversation_questions

[3]https://assetstore.unity.com/packages/tools/animation/final-ik-14290

3.3.4 Latency Adjustment. For synthesized expressions we precalculated phoneme weights for the audio track to synchronize them. The Oculus Avatar SDK used them with an offset of 40ms. To get them in their original latency, we live-fed the actor audio into the SDK.

To synchronize tracked expressions with the audio track, we offset the animation track by our measured motion-to-photon latency. We recorded a person repeatedly opening their mouth with a bilabial plosive and manually counted the number of frames until their mirrored ARKit mesh opened its mouth. This yielded an offset of 152.5ms. For original latency, we used the tracked expressions without delay.

The body pose was synchronized to audio by slowly swinging a controller in the face tracker's camera frustum and delaying the audio track so that movement peaks co-occur. Stauffert and colleagues report a motion-to-photon latency for a HTC Vive tracker [36] of 56.14ms, which we used as delay for body pose tracking data.

3.4 Immersive Observation

Observation in VR occurred seated from the perspective of the mannequin facing the actor during performance capture. Since we focus on effects of facial animation, we occluded the character's lower body and forearms by placing a table between observer and actor in the otherwise empty virtual environment. We used a Reverb G2 Omnicept as headset (resolution of 2160x2160 pixels per eye at 90Hz refresh rate).

3.5 Procedure

Our study procedure, depicted in Figure 4, took about 60 minutes. We welcomed participants and let them read the study briefing. After we answered questions about the procedure, participants gave informed written consent to their participation and use of their data. Then they filled out digital questionnaires about previous XR and gaming experience, demographics, and symptoms related to simulator-sickness using the Simulator Sickness Questionnaire (SSQ) [12] on a dedicated workstation. The experimenter performed quasi-random group assignment using covariate-adaptive randomization [11]. Accordingly, participants were evenly assigned to the latency adjustment conditions (all animations either in sync with audio or delayed by the animation system's latency) across biological sex and previous XR and gaming experience. Each participant then observed and rated the virtual character in four different performance blocks.

In each block, the performance was shown in its three facial animation variations, first rated with a static face, then with synthesized and tracked facial expression in randomized order. Thus, every participant rated twelve observations. To conclude, participants reported SSQ scores, familiarity with the shown actor before the experiment, and gave optional study feedback in an open text field.

Before the first trial, the experimenter instructed them on how to don the VR headset including adjusting lens spacing, strap fit, and a sound test. When ready, participants were instructed to sit relaxed and the virtual camera was calibrated. Once calibrated, the previously set up black screen overlay was removed to reveal the virtual scene. In it, a text panel informed participants about the virtual character about to be shown and instructed them to observe him attentively. Then, the panel showed the prompt he would respond to. Performances ended by hiding the virtual character and displaying a prompt to remove the headset to continue with the questionnaire on the computer. Here, we investigated *behavior naturalness* by asking for agreement to three statements on a Likert-scale (7 points from "completely disagree" to "completely agree"): 1. The verbal behavior of the virtual character seemed natural, 2. The nonverbal behavior of the virtual character seemed natural, 3. The verbal and nonverbal behavior of the virtual character fit together. To assess *appearance and behavior plausibility*, we asked for agreements to six statements from the same-named dimension of the Virtual Human Plausibility Questionnaire (VHPQ [18]) on a Likert-scale (7 points from "completely disagree" to "completely agree"): 1. The behavior of the virtual character seemed plausible, 2. The appearance of the virtual character seemed plausible, 3. The virtual character's behavior matched its appearance, 4. The behavior and appearance of the virtual character were coherent, 5. The virtual character behaved as I would expect it to behave, 6. I could predict how the virtual character would behave by its appearance.

3.6 Participants

We recruited 48 participants via our university's participation management system. They were free to pick a time slot provided on weekdays during normal working hours and compensated with €10. We excluded four from analysis - two due to technical issues with the setup, one because of language comprehension issues, and one because they reported to be familiar with the actor.

The 44 participants we included for analysis (29 of them female) had a mean age of 26.1 years (SD=6.1) and mostly had a higher education entrance qualification (23) or completed studies (19).

4 RESULTS

We used R v4.2.2[27] for analysis, aggregated Likert-scale ratings as interval-level data [23], and evaluated effects with linear mixed modeling using nlme[26]. We used random intercepts for all variables and modeled factors and interactions as fixed effects, as proposed by Field et al. [8]. All effects are reported as significant at $p<.05$ using log-likelihood ratio. Descriptive values are shown in Table 1. To test our hypotheses, we used two planned orthogonal contrasts throughout our analysis: firstly comparing observation of static faces to observation of either animated faces (tracked or synthesized), secondly comparing synthesized to tracked expressions. Since post-exposure SSQ scores were lower than pre-exposure in both conditions, we did not analyze simulator sickness further.

4.1 Behavior Naturalness

Rating of *nonverbal behavior naturalness* was significantly affected by facial animation method ($\chi^2(2)=93.24$, $p<.001$), but not by latency adjustment ($\chi^2(1)=.25$, $p=.62$). There was no significant interaction between latency adjustment and facial animation method ($\chi^2(2)=.50$, $p=.78$). Contrasting ratings of static faces with ratings

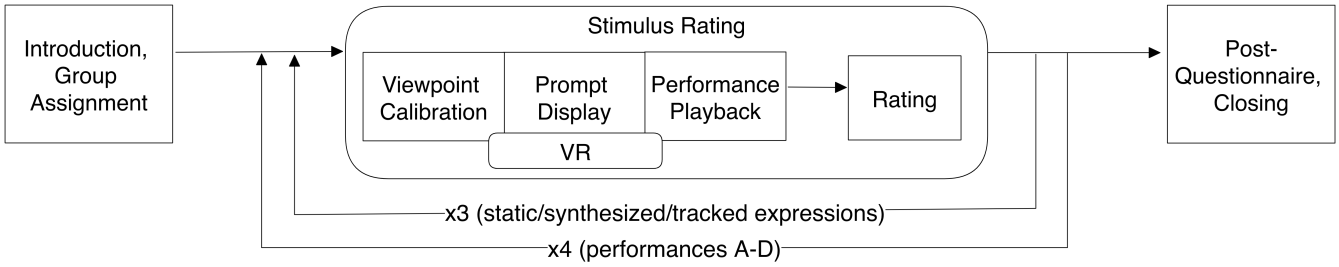

Figure 4: *Experiment Procedure*

of synthesized and tracked expressions revealed a significant difference (b=.68, t(86)=11.8, p=<.001). Likewise, contrasting synthesized with tracked animations proved significant (b=-.5, t(86)=-5, p=<.001).

Verbal behavior naturalness rating was also significantly affected by facial animation method ($\chi^2(2)$=68.36, p<.001), but not by the method's latency ($\chi^2(1)$=.37, p=.54). There was no significant interaction between latency adjustment and facial animation method ($\chi^2(2)$=2.62, p=.27). Contrasting the static face with the two dynamic methods (synthesized and tracked) revealed a significant difference (b=.65, t(86)=9.74, p=.006). Likewise, the contrast comparing synthesized to tracked animations proved significant (b=-.5, t(86)=-5, p=<.001).

Match between verbal and nonverbal behavior was significantly affected by facial animation method ($\chi^2(2)$=167.53, p<.001), but not by latency adjustment ($\chi^2(1)$=1.19, p=.28). Ratings show no significant interaction between latency adjustment and facial animation method ($\chi^2(2)$=2.24, p=.33). Contrasting animated with non-animated faces revealed significant differences (b=1.01, t(86)=18.9, p<.001), as did contrasting both animated face versions (b=-.56), t(86)=-6.08, p<.001).

4.2 Appearance and Behavior Plausibility

Rating of *appearance and behavior plausibility* was significantly affected by the facial animation method ($\chi^2(2)$=104.69, p<.001), but not by the method's latency ($\chi^2(1)$=.53, p=.47). There was no significant interaction between latency adjustment and facial animation method ($\chi^2(2)$=.34, p=.84).

Contrasting ratings of static faces with the two dynamic methods (synthesized and tracked) revealed a significant difference (b=.55, t(86)=13.0, p=<.001). Likewise, the contrast comparing synthesized to tracked animations proved significant (b=-.36, t(86)=-5.45, p=<.001).

4.3 Qualitative Feedback

Several participants mentioned they felt directly addressed by the avatar. A few participants speculated about what changed between observations within a performance block. Some guessed that faces continuously moved more realistically from trial to trial within a block, although we randomized the condition order. Multiple participants mentioned they had difficulty in differentiating the two animated conditions.

5 DISCUSSION

Static faces were, on average, rated as less natural and less plausible than animated faces. Notably, this effect was also shown for verbal naturalness. While we had formulated our hypothesis H1 towards the overall difference in naturalness, we highlight this difference because verbal behavior was never manipulated and always equal per performance block.

Against our hypothesis H2, the performance ratings did not reveal an interaction effect between facial animation method and its synchronization to speaker audio. This might be due the relatively small audiovisual skew we used between groups. Also, rating differences in verbal behavior naturalness hint at participants not necessarily focusing on the two channels (verbal/ nonverbal) distinctly, while not paying close attention to their temporal alignment.

While our planned contrasts did show significant differences, we expected the comparison between the two animated performances two show inverted differences: All plausibility and naturalness ratings were, on average, higher for synthesized expressions than for tracked ones. In other words, participants found non-personalized, "generic" expressions fit the actor's avatar better than ones from the actor himself. This might stem from artifacts in the tracked facial expressions or the synthesized expressions including more prosocial cues.

As a first suspicion, we retroactively re-watched the tracked performances with a focus on tracking artifacts. While common in live tracking data, e.g. in the form of jittery lips, sudden jumps after tracking loss, and eyes and/or mouth not closing completely, these issues usually do not occur in facial expression synthesis. Synthesized expressions are smoother by design. However, we did not find such artifacts in the captured facial performance.

In a further exploratory analysis we compared the two animated face variations. In social settings, gaze behavior includes directing gaze at another one's face (face-gaze) or eyes (eye-gaze), simultaneously looking at each other's face (mutual gaze) or eyes (eye contact), and intentionally not looking at another person (gaze avoidance). Since we did not record observer eye gaze, we looked at how often the virtual human's eyes were directed toward the observer. We calculated how long in total the virtual human's gaze rays hit the mannequin's head. More specifically, we checked whether a sphere moving from eye origin along the gaze direction hit a capsule collider placed to fit the mannequin head. The tracked expressions included more gaze towards the mannequin, but with shorter dwell times.

Latency	Face	Appearance + Behavior Plausibility	Behavior Naturalness		
			Verbal	Nonverbal	Match
Original	Static	3.56 (1.26)	4.39 (2.36)	3.26 (1.91)	2.28 (1.45)
	Synthesized	5.67 (0.98)	6.36 (0.85)	5.85 (1.20)	6.14 (0.90)
	Tracked	4.83 (1.28)	5.68 (1.50)	4.72 (1.55)	4.89 (1.60)
Adjusted	Static	3.80 (1.38)	3.76 (2.38)	3.36 (0.50)	2.77 (1.51)
	Synthesized	5.75 (1.00)	6.35 (0.83)	5.88 (1.21)	6.08 (1.20)
	Tracked	5.02 (1.20)	5.73 (1.28)	5.01 (1.47)	5.09 (1.51)

Table 1: *Descriptive statistics.* Means of independent variables with standard deviations in brackets.

We suggest this rating difference in favor of synthesized expressions should be utilized for conversational agents or playing back monologues. Therefore, more natural and more plausible animations might be achieved without the need for facial expression tracking. For truly interactive settings with dynamic turn-taking our findings might not generalize. Since facial expression tracking also contains personal facial expression dynamics and veridical gaze points, showing synthesized expressions instead might contribute to misunderstandings. Still, achieving natural and plausible facial animations with tracking input from consumer devices is a valuable insight for XR researchers and developers.

6 LIMITATIONS AND FUTURE WORK

We presented one character across throughout all observations. This might have rendered the experimental setting less realistic, since people usually do not repeat themselves word by word. We opted for this approach because it allowed direct comparisons between the conditions, but suggest to also compare characters that differ in factors like gender, ethnicity, and voice. Similarly, we suggest exploring effects of facial animation depending on character familiarity.

Also, our approach could be extended to truly interactive setups. However, latency might show difficult as independent variable because audio and nonverbal behavior can then only be synchronized by delaying audio signals by the facial animation's processing duration. This amplifies the difference between conversation partners' "non-mutual realities" [32] and results in misunderstandings, e.g. in the form of overlapping talk [33].

Furthermore, future work should also address effects of nonverbal behavior on the perception of verbal behavior in more detail.

7 CONCLUSION

We explored how two facial animation methods (tracking or synthesis of facial expression) compared to each other and a baseline of a static face when used on a personalized, photorealistic virtual human. In a mixed-design observation study, 44 participants observed four performances, each in its three facial animation variations, and subsequently rated their appearance and behavior plausibility, and behavior naturalness. The between-groups factor of latency adjustment (animation in-sync with audio or delayed by recording latency) showed not to influence ratings significantly. However, the within-factor facial animation method showed to significantly affect ratings: Overall, performances were rated more plausible and more natural when shown with animated faces (synthesized/tracked expressions), even more so for synthesized ones than for tracked ones. This implies that natural, plausible facial animations for avatars do not require facial expression tracking when showing another avatar's monologue. We suggest further work to address these implications for character variations and truly interactive settings.

ACKNOWLEDGMENTS

Our research has been funded by the German Federal Ministry of Education and Research (BMBF) in the project ViLeArn More (project ID 16DHB2214) and by the Bavarian State Ministry For Digital Affairs in the project XR Hub (project ID A5-3822-2-16).

We thank Thomas Schröter for his patience during performance capture sessions, Sebastian Oberdörfer for for insightful manuscript discussions and are grateful to Andrea Bartl, Florian Kern, and David Mal for helping us refine method details. We are also thankful to the anonymous reviewers for their helpful comments.

REFERENCES

[1] Jascha Achenbach, Thomas Waltemate, Marc Erich Latoschik, and Mario Botsch. 2017. Fast Generation of Realistic Virtual Humans. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology* (Gothenburg, Sweden) *(VRST '17)*. Association for Computing Machinery, New York, NY, USA, Article 12, 10 pages. https://doi.org/10.1145/3139131.3139154

[2] Nalini Ambady and Robert Rosenthal. 1992. Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences: A Meta-Analysis. *Psychological bulletin* 111, 2 (1992), 256. https://doi.org/10.1037/0033-2909.111.2.256

[3] Jeremy N. Bailenson and Jim Blascovich. 2004. Avatars. In *Encyclopedia of Human-Computer Interaction*. Vol. 1. Berkshire Publishing Group, 64–6.

[4] Andrea Bartl, Stephan Wenninger, Erik Wolf, Mario Botsch, and Marc Erich Latoschik. 2021. Affordable But Not Cheap: A Case Study of the Effects of Two 3D-Reconstruction Methods of Virtual Humans. *Frontiers in Virtual Reality* 2 (Sept. 2021), 694617. https://doi.org/10.3389/frvir.2021.694617

[5] David Borland, Tabitha Peck, and Mel Slater. 2013. An Evaluation of Self-Avatar Eye Movement for Virtual Embodiment. *IEEE Transactions on Visualization and Computer Graphics* 19, 4 (April 2013), 591–596. https://doi.org/10.1109/TVCG.2013.24

[6] Jeffrey F. Cohn and Paul Ekman. 2005. Measuring Facial Action. In *The New Handbook of Methods in Nonverbal Behavior Research*. Oxford University Press, New York, NY, US, 9–64.

[7] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael Black. 2019. Capture, Learning, and Synthesis of 3D Speaking Styles. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 10101–10111. http://voca.is.tue.mpg.de/

[8] Andy P. Field, Jeremy Miles, and Zoë Field. 2012. *Discovering Statistics Using R*. Sage, London ; Thousand Oaks, Calif.

[9] Alexandra Frischen, Andrew P. Bayliss, and Steven P. Tipper. 2007. Gaze Cueing of Attention: Visual Attention, Social Cognition, and Individual Differences. *Psychological Bulletin* 133, 4 (2007), 694–724. https://doi.org/10.1037/0033-2909.133.4.694

[10] Mar Gonzalez-Franco, Anthony Steed, Steve Hoogendyk, and Eyal Ofek. 2020. Using Facial Animation to Increase the Enfacement Illusion and Avatar Self-Identification. *IEEE Transactions on Visualization and Computer Graphics* 26, 5 (May 2020), 2023–2029. https://doi.org/10.1109/TVCG.2020.2973075

[11] Man Jin, Adam Polis, and Jonathan Hartzel. 2021. Algorithms for Minimization Randomization and the Implementation with an R Package. *Communications in Statistics - Simulation and Computation* 50, 10 (Oct. 2021), 3077–3087. https://doi.org/10.1080/03610918.2019.1619765

[12] Robert S Kennedy, Norman E Lane, Kevin S Berbaum, and Michael G Lilienthal. 1993. Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness. *The international journal of aviation psychology* 3, 3 (1993), 203–220. https://doi.org/10.1207/s15327108ijap0303_3

[13] Chris L. Kleinke. 1986. Gaze and Eye Contact: A Research Review. *Psychological Bulletin* 100, 1 (1986), 78–100. https://doi.org/10.1037/0033-2909.100.1.78

[14] Hiromi Kobayashi and Shiro Kohshima. 2001. Unique Morphology of the Human Eye and Its Adaptive Meaning: Comparative Studies on External Morphology of the Primate Eye. *Journal of Human Evolution* 40, 5 (May 2001), 419–435. https://doi.org/10.1006/jhev.2001.0468

[15] Marc Erich Latoschik and Carolin Wienrich. 2022. Congruence and Plausibility, Not Presence: Pivotal Conditions for XR Experiences and Effects, a Novel Approach. *Frontiers in Virtual Reality* 3 (June 2022), 694433. https://doi.org/10.3389/frvir.2022.694433

[16] Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha S. Srinivasa, and Yaser Sheikh. 2019. Talking With Hands 16.2M: A Large-Scale Dataset of Synchronized Body-Finger Motion and Audio for Conversational Motion Analysis and Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 763–772. https://doi.org/10.1109/ICCV.2019.00085

[17] Hao Li, Laura Trutoiu, Kyle Olszewski, Lingyu Wei, Tristan Trutna, Pei-Lun Hsieh, Aaron Nicholls, and Chongyang Ma. 2015. Facial Performance Sensing Head-Mounted Display. *ACM Transactions on Graphics* 34, 4 (July 2015), 1–9. https://doi.org/10.1145/2766939

[18] David Mal, Erik Wolf, Nina Dollinger, Mario Botsch, Carolin Wienrich, and Marc Erich Latoschik. 2022. Virtual Human Coherence and Plausibility – Towards a Validated Scale. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE. https://doi.org/10.1109/vrw55335.2022.00245

[19] David Matsumoto and Hyisung C. Hwang. 2018. Microexpressions Differentiate Truths From Lies About Future Malicious Intent. *Frontiers in Psychology* 9 (Dec. 2018), 2545. https://doi.org/10.3389/fpsyg.2018.02545

[20] Harry McGurk and John MacDonald. 1976. Hearing Lips and Seeing Voices. *Nature* 264, 5588 (1976), 746–748. https://doi.org/10.1038/264746a0

[21] Timo Menzel, Mario Botsch, and Marc Erich Latoschik. 2022. Automated Blendshape Personalization for Faithful Face Animations Using Commodity Smartphones. In *Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology* (Tsukuba, Japan) *(VRST '22)*. Association for Computing Machinery, New York, NY, USA, Article 22, 9 pages. https://doi.org/10.1145/3562939.3565622

[22] Maria Murcia-Lopez, Tara Collingwoode-Williams, William Steptoe, Raz Schwartz, Timothy J. Loving, and Mel Slater. 2020. Evaluating Virtual Reality Experiences Through Participant Choices. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, Atlanta, GA, USA, 747–755. https://doi.org/10.1109/VR46266.2020.00098

[23] Geoff Norman. 2010. Likert Scales, Levels of Measurement and the "Laws" of Statistics. *Advances in Health Sciences Education* 15, 5 (Dec. 2010), 625–632. https://doi.org/10.1007/s10459-010-9222-y

[24] Catherine Oh Kruzic, David Kruzic, Fernanda Herrera, and Jeremy Bailenson. 2020. Facial Expressions Contribute More than Body Movements to Conversational Outcomes in Avatar-Mediated Virtual Environments. *Scientific Reports* 10, 1 (Dec. 2020), 20626. https://doi.org/10.1038/s41598-020-76672-4

[25] Kyle Olszewski, Joseph J. Lim, Shunsuke Saito, and Hao Li. 2016. High-Fidelity Facial and Speech Animation for VR HMDs. *ACM Transactions on Graphics* 35, 6 (Nov. 2016), 1–14. https://doi.org/10.1145/2980179.2980252

[26] José Pinheiro, Douglas Bates, and R Core Team. 2022. *nlme: Linear and Nonlinear Mixed Effects Models*. https://CRAN.R-project.org/package=nlme R package version 3.1-161.

[27] R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[28] Alexander Richard, Colin Lea, Shugao Ma, Jurgen Gall, Fernando de la Torre, and Yaser Sheikh. 2021. Audio- and Gaze-Driven Facial Animation of Codec Avatars. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 41–50. https://doi.org/10.1109/WACV48630.2021.00009

[29] Daniel Roth, Gary Bente, Peter Kullmann, David Mal, Chris Felix Purps, Kai Vogeley, and Marc Erich Latoschik. 2019. Technologies for Social Augmentations in User-Embodied Virtual Reality. In *25th ACM Symposium on Virtual Reality Software and Technology*. ACM, Parramatta NSW Australia, 1–12. https://doi.org/10.1145/3359996.3364269

[30] Daniel Roth, Carola Bloch, Anne-Kathrin Wilbers, Marc Erich Latoschik, Kai Kaspar, and Gary Bente. 2016. What You See Is What You Get: Channel Dominance in the Decoding of Affective Nonverbal Behavior Displayed by Avatars. In *66th Annual Conference of the International Communication Association*. Fukuoka, Japan.

[31] Daniel Roth, Peter Kullmann, Gary Bente, Dominik Gall, and Marc Erich Latoschik. 2018. Effects of Hybrid and Synthetic Social Gaze in Avatar-Mediated Interactions. In *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, Munich, Germany, 103–108. https://doi.org/10.1109/ISMAR-Adjunct.2018.00044

[32] Karen Ruhleder and Brigitte Jordan. 2001. Co-constructing non-mutual realities: Delay-generated trouble in distributed interaction. *Computer Supported Cooperative Work (CSCW)* 10, 1 (2001), 113–138. https://doi.org/10.1023/A:1011243905593

[33] Lucas M. Seuren, Joseph Wherton, Trisha Greenhalgh, and Sara E. Shaw. 2021. Whose Turn Is It Anyway? Latency and the Organization of Turn-Taking in Video-Mediated Interaction. *Journal of Pragmatics* 172 (Jan. 2021), 63–78. https://doi.org/10.1016/j.pragma.2020.11.005

[34] Richard Skarbez, Solene Neyret, Frederick P. Brooks, Mel Slater, and Mary C. Whitton. 2017. A Psychophysical Experiment Regarding Components of the Plausibility Illusion. *IEEE Transactions on Visualization and Computer Graphics* 23, 4 (April 2017), 1369–1378. https://doi.org/10.1109/TVCG.2017.2657158

[35] Mel Slater, Domna Banakou, Alejandro Beacco, Jaime Gallego, Francisco Macia-Varela, and Ramon Oliva. 2022. A Separate Reality: An Update on Place Illusion and Plausibility in Virtual Reality. *Frontiers in Virtual Reality* 3 (June 2022), 914392. https://doi.org/10.3389/frvir.2022.914392

[36] Jan-Philipp Stauffert, Florian Niebling, and Marc Erich Latoschik. 2020. Simultaneous Run-Time Measurement of Motion-to-Photon Latency and Latency Jitter. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, Atlanta, GA, USA, 636–644. https://doi.org/10.1109/VR46266.2020.00086

[37] R. Steinmetz. Jan./1996. Human Perception of Jitter and Media Synchronization. *IEEE Journal on Selected Areas in Communications* 14, 1 (Jan./1996), 61–72. https://doi.org/10.1109/49.481694

[38] Monica Villanueva Aylagas, Hector Anadon Leon, Mattias Teye, and Konrad Tollmar. 2022. Voice2Face: Audio-driven Facial and Tongue Rig Animations with cVAEs. *Computer Graphics Forum* (2022). https://doi.org/10.1111/cgf.14640

[39] Matias Volonte, Eyal Ofek, Ken Jakubzak, Shawn Bruner, and Mar Gonzalez-Franco. 2022. HeadBox: A Facial Blendshape Animation Toolkit for the Microsoft Rocketbox Library. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, Christchurch, New Zealand, 39–42. https://doi.org/10.1109/VRW55335.2022.00015

[40] Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and Speech in Interaction: An Overview. *Speech Communication* 57 (Feb. 2014), 209–232. https://doi.org/10.1016/j.specom.2013.09.008

[41] Audrey C Younkin and Philip J Corriveau. 2008. Determining the amount of audio-video synchronization errors perceptible to the average end-user. *IEEE Transactions on Broadcasting* 54, 3 (2008), 623–627. https://doi.org/10.1109/TBC.2008.2002102

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009