Real-Time Conversational Gaze Synthesis for Avatars

Ryan Canales
Clemson University
Clemson, SC, USA
rcanale@clemson.edu

Eakta Jain
University of Florida
Gainesville, Florida, USA
ejain@cise.ufl.edu

Sophie Jörg
University of Bamberg
Bamberg, Germany
Clemson University
Clemson, SC, USA
sophie.joerg@uni-bamberg.de

ABSTRACT

Eye movement plays an important role in face-to-face communication. In this work, we present a deep learning approach for synthesizing the eye movements of avatars for two-party conversations and evaluate viewer perception of different types of eye motions. We aim to synthesize believable gaze behavior based on head motions and audio features as they would typically be available in virtual reality applications. To this end, we captured the head motion, eye motion, and audio of several two-party conversations and trained an RNN-based model to predict where an avatar looks in a two-person conversational scenario. We evaluated our approach with a user study on the perceived quality of the eye animation and compared our method with other eye animation methods. While our model was not rated highest, our model and our user study lead to a series of insights on model features, viewer perception, and study design that we present.

CCS CONCEPTS

• **Computing methodologies** → **Virtual reality**; **Perception**; **Animation**.

KEYWORDS

gaze animation, avatars, motion perception, virtual reality

ACM Reference Format:
Ryan Canales, Eakta Jain, and Sophie Jörg. 2023. Real-Time Conversational Gaze Synthesis for Avatars. In *ACM SIGGRAPH Conference on Motion, Interaction and Games (MIG '23), November 15–17, 2023, Rennes, France.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3623264.3624446

1 INTRODUCTION

Virtual reality (VR) devices and social VR applications are increasingly becoming mainstream. It will soon become commonplace to use self avatars to interact with others in immersive virtual environments. In this work, we consider the generation of eye movements for avatars in a two-party conversation context. We focused on synthesizing conversational eye animation because our eyes are essential for expressing non-verbal conversational cues during face-to-face communication. They help mediate conversational turn taking, facilitate emotional expression, and signal engagement [Lee et al. 2002; Ruhland et al. 2014]. Since our eye movement is so important for communicating, it follows that enabling synthesized eye motion for real-time scenarios is a crucial milestone toward effective avatar-mediated communication. In fact, research has repeatedly shown that our overall impression of virtual humans improves when their eyes behave realistically [Lee et al. 2002; Roth et al. 2018]. Unfortunately, current methods for synthesizing eye motion for conversational virtual humans in real-time remain limited, and most VR devices in use still do not capture eye motions since eye tracking remains largely a premium feature.

We present a novel data-driven method for synthesizing the eye motion of virtual humans for use in VR and other real-time applications. Our method predicts the rotation of the eyes of a virtual human based on a user's speech features and the head movement of both the speaker and listener. Our algorithm is designed for avatar-mediated, two-party conversational scenarios, but would also be applicable to conversational agents where the head movements and audio are generated before the eye movements.

2 RELATED WORK

2.1 Eye Motions and Virtual Humans

Psychological research indicates that common patterns in our gaze behavior emerge during interpersonal interactions [Abele 1986; Ho et al. 2015]. For example, people attend to the speaker when they are listening. But the speaker tends to look away from the listener, because they are thinking, for example. Gaze also provides non-verbal cues during conversation, for example, when the speaker is finished, they may look at the listener to signal that they are ready for a response. In addition, gaze behavior also communicates personality traits.

Normoyle et al. [2013] demonstrated that small changes to the frequency of eye contact affects the trustworthiness of the virtual character. Jörg and colleagues [2018] found that even very small changes to eye motion on virtual characters can affect the perceived naturalness of the motion . Personality traits such as openness, conscientiousness, extraversion, agreeableness, and neuroticism can be discerned based on the eye motion of virtual humans [Ruhland et al. 2015]. Human sensitivity to an avatar's eye movements has motivated work in simulation of realistic eye motion for virtual characters and avatars in social situations.

Gaze behavior models designed to mimic natural eye motion during conversation have been shown to positively impact avatar-mediated communication. Simulating social gaze has been shown to improve the realism of face-to-face interactions during two party conversations in VR [Garau et al. 2003]. Lee et al. [2002] compared a procedural model to static and random gaze generators, and found

MIG '23, November 15–17, 2023, Rennes, France
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0393-5/23/11.
https://doi.org/10.1145/3623264.3624446

that responses for friendliness, engagement, and liveliness were higher when the procedural model was used. Oyekoya et al. [2009] found that a saliency-based gaze behavior model in a virtual conversational scenario was comparable to tracked gaze and significantly better than static and random gaze models in terms of naturalness and realism. Seele et al. [2017] compared three gaze-behavior models in a dyadic social VR setting: an off-the-shelf parameterized, scene agnostic eye motion generator, a scene-base gaze synthesis algorithm, and gaze from an eye tracker. They reported no significant impact of model type on quality of social interaction but trends to suggest that tracked gaze can improve presence and avatar realism.

2.2 Gaze Synthesis

Numerous techniques have been developed for synthesizing the eye motions of virtual humans [Ruhland et al. 2014]. Early conversational gaze models used statistics to infer gaze timings, such as eye contact and the frequencies of fixations and saccades [Lee et al. 2002; Vinayagamoorthy et al. 2004].

Gu and Badler [2006] incorporated attention in their gaze model for embodied conversational agents in multi-party conversational scenarios, allowing agents to look toward distracting objects during conversation. Iwao et al. [2012; 2013] synthesized the more subtle eye movements that occur during fixations using probability models derived from captured gaze data from two party conversations.

Subsequent to the early work on procedural models, data-driven models were investigated. They trained on captured motions to animate the eyes without explicit rules or assumptions. Ma et al. [2009] generated eye movement based on the head movement while Le et al. [2012] created a model that generated both head and eye motion based only on speech input for two party conversations. Jin et al. [2019] utilized a deep learning approach, training two recurrent neural networks (RNNs), one for the speaker and the other for listeners, to synthesize head and eye motions of virtual avatars based on the audio from participants in a three-party conversation.

In addition to communication specific gaze behavior models, there are several methods that synthesize or retarget gaze based on the location of gaze targets or the visual saliency of the virtual scene. Peters et al. [2010] developed a gaze shift model that animates the head, eyes, and blinks of a character based on the gaze target location and a parameter specifying the tendency the character moves their head. Taking into account physiological constraints, Andrist et al. [2012] designed a parametric gaze model that animates both the avatar's eyes and head towards a target in the scene. Pejsa et al. [2016] developed a tool for animators to modify the torso, head, and eye animation of a character to look towards a new target.

More recently, deep learning has been used for driving a character's gaze. For example, Klein et al. [2019] used an RNN to animate a character's upper-body as it follows a moving gaze target in real-time, and Goudé et al. [2023] trained a model to predict areas of interest in real-time based on the visual saliency of the scene. Although there has been success in synthesizing gaze for virtual characters, research into data-driven conversational gaze behavior models for use in immersive virtual environments is limited.

In our work, we focus on developing and evaluating a data-driven approach for animating the eyes of a virtual avatar based on the head motion and audio during a dyadic face-to-face conversation.

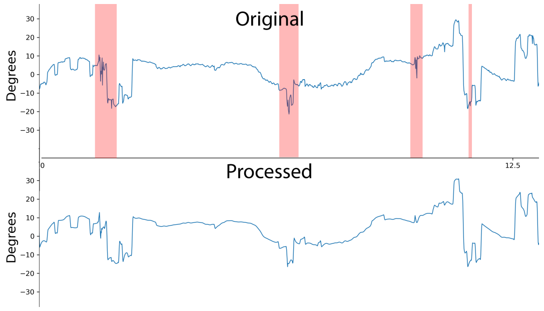


Figure 1: A sample from our data showing the horizontal gaze angle θ before processing (top) and after (bottom). Errors from tracking loss are highlighted in the original graph.

3 METHOD

To model gaze direction based on motion and speech inputs, we trained a recurrent neural network (RNN), which can capture temporal relations, on a dataset consisting of two-party conversations.

3.1 Data Collection and Preprocessing

3.1.1 Hardware. We recorded detailed gaze and pupil information from the left eye of one of the conversational partners (called the primary performer) at 120Hz using a Pupil Labs Core wearable eye tracker. A motion capture system consisting of 15 Optitrack motion capture cameras recorded the head movements (position and orientation) of both performers at 120fps. Finally, an Audio-Technica AT2020 microphone recorded audio in mono, sampled at 44.1kHz. We opted for an external eye tracker in combination with a motion capture system, as opposed to a head mounted display with an integrated eye tracker, so that the conversational partners could see each other leading to a more natural eye motions. Furthermore, based on our experience, this procedure leads to higher quality motion data.

3.1.2 Capture Protocol. In each capture session, two performers stood facing each other about one meter apart. Both performers wore a hat with reflective motion capture markers. The primary performer wore the eye tracker and the microphone was directed towards them. The eye tracker was calibrated to an accuracy within 3 degrees within the field of view. After the motion capture system and eye tracker are adjusted and calibrated, a clapperboard with motion capture markers was used to synchronize sound and head motions and to signal the start of the conversation. Performers were asked to converse naturally while standing in one spot. They could select a conversational topic from a list of topics designed for casual conversations in English classes [Teflpedia [n. d.]] or choose their own topic, for example, holiday plans, food preferences, and favorite media. Performers were instructed to avoid any identifying or personal information. Conversation were limited to roughly three minutes. The clapperboard was used again at the end of the conversation. There were 4 primary performers and 5 conversational partners. A total of 2179.5 seconds (about 36 minutes) of data from 11 conversations was recorded. The ratio that the primary performer's gaze was directed towards the other was 67% while speaking, 86% while listening, and 76% overall.

3.1.3 *Data Preprocessing.*

The captured data of each session is synchronized manually based on the cues from the clapperboard. The 3D gaze vector in space is computed combining the eye tracker data and the performer’s head position or rotation. The audio data is manually labeled to indicate when each performer is speaking.

Since the captured gaze data contains noise when the subject blinks, squints, or turns their eyes far from center, a multi-step gaze preprocessing pipeline was implemented to reduce the number of artifacts. The 3D gaze vector provided by the eye tracker is first converted to spherical coordinates, resulting in horizontal (θ) and vertical (ϕ) gaze direction angles. These angles are then smoothed using a Savitzky-Golay filter with a cubic curve and a window size of 11, to filter high frequency noise before further processing [Duchowski et al. 2016]. We then use the confidence for each sample (a value between 0 and 1, provided by Pupil [Kassner et al. 2014]), to linearly interpolate the gaze angles between high confidence (c > 0.9) samples within each conversation. Next, we implement an automatic saccade detection algorithm based on Jörg et al. [2018] which is a variant of Engbert and Kliegl’s [2003] saccade detection algorithm. The algorithm has two saccade detection sensitivity parameters: λ_{low} for small saccades and λ_{high} for large saccades. These parameters were tuned for each conversation in our dataset. Saccades were detected in two passes, with the first pass searching for large amplitude saccades, and the second pass for low amplitude saccades, and then merged such that there were no overlapping saccade instances. To smooth the data between the detected saccades, we apply a 2^{nd} degree Butterworth filter with a cutoff frequency of 3Hz at a 120Hz sampling rate. We enforce a minimum fixation duration of 5 frames or 50ms to limit erroneous detection of saccades. Finally, the resulting smoothed data is downsampled from 120Hz to 60Hz. A comparison of the original data and processed data is shown in Figure 1.

3.2 Network Architecture and Training

3.2.1 *Features.*

The following motion features were computed and used for training the network: 1) The 3D gaze direction (two angles in spherical coordinates, azimuth and elevation), 2) horizontal and vertical angles of the primary performers facing direction vector, 3) horizontal and vertical angles of the vector between the head positions of the two performers, 4) velocity of the primary performer’s head rotation computed as the sum of the squared difference in the performer’s head pitch and yaw from frame to frame.

We trained versions of our model with audio features either as relative pitch and intensity or as a binary feature indicating whether or not the primary performer was speaking. The audio pitch and intensity are computed over windows of 738 audio samples (16.7ms), corresponding to the duration of a single frame at 60fps. The pitch at a particular frame is the pitch with the maximum amplitude within the window of audio during the frame, and the intensity is the root mean squared value of the amplitudes for each sample. Similar to Jin et al. [2019], we use the difference in pitch and intensity features between two neighboring frames. Concretely, the relative pitch (p) and intensity (I) features for frame i are: $(p_i - p_{i-1}, I_i - I_{i-1})$.

3.2.2 *Neural Network Architecture.*

Our neural network (NN) architecture is inspired by the network used for gaze event classification by Kothari et al. [Kothari et al. 2020] and is shown in Figure 2. The input sequence is transformed by four linear layers then fed to a stacked 3 layer Gated Recurrent Unit (GRU). The output of the GRU is then transformed by two linear layers and a final linear layer makes a prediction. A GRU is used in this work since it is faster to train and is generally found to achieve results on par with a Long-Short Term Memory (LSTM). The model predicts the gaze direction as two angles, θ and ϕ, corresponding to the horizontal and vertical components of the unit gaze direction vector, respectively. The depth component of the gaze vector is not used due to the inaccurate depth estimate provided by the monocular eye tracker setup.

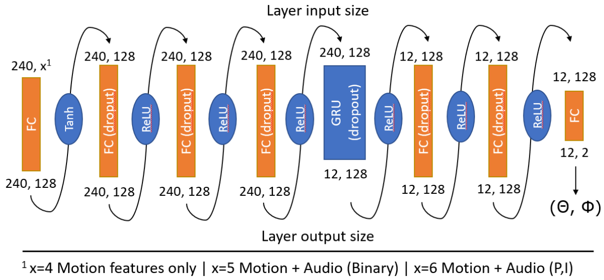


Figure 2: The neural network architecture with layer input and output sizes shown above and below each layer. The input is a sequence of feature vectors ($N = 240$) and the output is a 12 frame gaze direction prediction (θ, ϕ).

In our method, a sequence of the previous N frames (with $N = 240$) are used to compute the features and to provide context for the model to predict the eye motion over the next 12 frames.

3.2.3 *Training.*

We divide our data into training, validation, and test sets by taking random, non-overlapping subsequences from each conversation. We use subsequences that are 60% of each conversation for training, 20% for validation, and 20% for testing. Each of the features across all conversations in our training set are standardized to have zero mean and unit variance.

The model is trained on batches of 4 second (240 frames) feature sequences from the training set. The audio within the 4 second window is pre-emphasized [2016] to improve the signal to noise ratio before computing the relative pitch and intensity features over 735 samples, corresponding to one frame of motion. We also augmented the training set by mirroring the motion features horizontally, resulting in 12644 windows of data in the training set (twice the original 6322), 1965 in the validation set, and 1965 in the test set. The RNN is stateless, meaning each feature sequence is treated as independent from other feature sequences. The loss is the mean squared error between the predicted gaze angles and the corresponding captured gaze angles. We also prevented the occasional bad gaze sample (confidence below 0.6) from influencing the training of the model: the loss per sample is $C * \frac{1}{2} * [(\bar{\theta} - \theta)^2 + (\bar{\phi} - \phi)^2]$, where C is set to 0 if the sample confidence is below 0.6, and 1 otherwise. To prevent overfitting the training set, we implement neuron dropout with a probability of 0.5 during training for each of the fully connected layers between the first and final layers (labeled in Figure 2) and the GRU. We also use L2 regularization and select the model that performs best on the validation set.

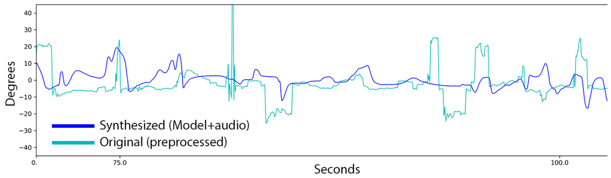


Figure 3: Example model output with motion and binary audio features in blue and the original preprocessed motion in cyan. The synthesized motion is noticeably smoother.

4 EVALUATION

4.1 Model Performance

Figure 3 shows an example of synthesized and recorded gaze angles from the test set for comparison. Synthesized example motions can be seen in the accompanying video. The model learned to compensate for avatar’s head movement and the movement of the other avatar. The model also avoids reproducing the artifacts found in the input motion, such as noise due to blinking or loss of tracking.

4.1.1 Computation speed. Our model is lightweight enough for use in real-time applications. A PC with an Nvidia RTX 3050Ti GPU and a 12th generation Intel i7 CPU can maintain the 60 frames per second required to run the model. We also implemented a VR setup where the model runs on a dedicated PC and is connected to a Meta Quest 2 via TCP. In this setup, the features are streamed from the headset to the PC, input into the model, then the model output is transmitted to the headset.

4.1.2 Ablation Study. We trained six different models to examine the effect of different features on the training, validation, and test set losses. The features tested were:

- Audio (binary): Only a binary audio feature was used for training the model. If the primary performer was speaking, this value was 1, otherwise it was 0.
- Audio (P, I): Only the relative pitch and intensity of the audio were used to train the model.
- Motion: Motion features only, as described in section 3.1.4
- Motion Augmented: Motion features augmented, in which the motion features in the training set were mirrored and added to the training set, called MA in the user study
- Motion Augmented + Audio (binary): Motion features and audio as a binary, called MAAB in the user study
- Motion Augmented + Audio (P,I): Motion features and audio as relative pitch and intensity

The losses for each model can be seen in Table 1.

4.1.3 Influence of Audio. Adding binary audio features to our model resulted in lower test set losses (see Table 1) in comparison to the augmented motion only. In a typical conversation, people look at their partner more often when listening than when talking. So, to evaluate if audio features influenced the output of the model, we compared the percentage of time during which the gaze is directed towards the other avatar when speaking and when listening. We found that with audio features, the gaze was directed at the other avatar about 93% of the time while speaking, and about 99% of the time while listening. Without audio features, the ratio while speaking is slightly higher at 96%, but the ratio is similar while listening at about 98%. Including binary audio features has created a higher difference between these ratios. However, these values are still much higher than in the original test set data, which has a ratio of about 72% while speaking and about 84% while listening.

Table 1: Models trained with different features and their associated training, validation, and test set losses. The highlighted boxes indicate the lowest loss for each column.

	Training	Validation	Test
Audio (binary)	0.02397	0.02303	0.02239
Audio (P, I)	0.02377	0.02267	0.02249
Motion	0.01779	0.02116	0.02099
Motion Aug	0.01884	0.02102	0.02098
Motion Aug + Audio (binary)	0.01843	0.02180	0.02082
Motion Aug + Audio (P,I)	0.01991	0.02131	0.02120

Further analysis revealed that audio may be a somewhat redundant feature already included in the head motion. We computed the Euclidean distance between the yaw (θ) and pitch (ϕ) of the primary performers head rotation between each frame i as follows: $\sqrt{(\theta_i - \theta_{i-1})^2 + (\phi_i - \phi_{i-1})^2}$. We also labeled each frame with either 0 for not speaking or 1 for speaking. We then fit a logistic regression model to this data, with the Euclidean distance between angles as the predictor for whether they are speaking, and found that is was a significant predictor ($\chi^2(1) \ll 0.001$).

4.2 User Study

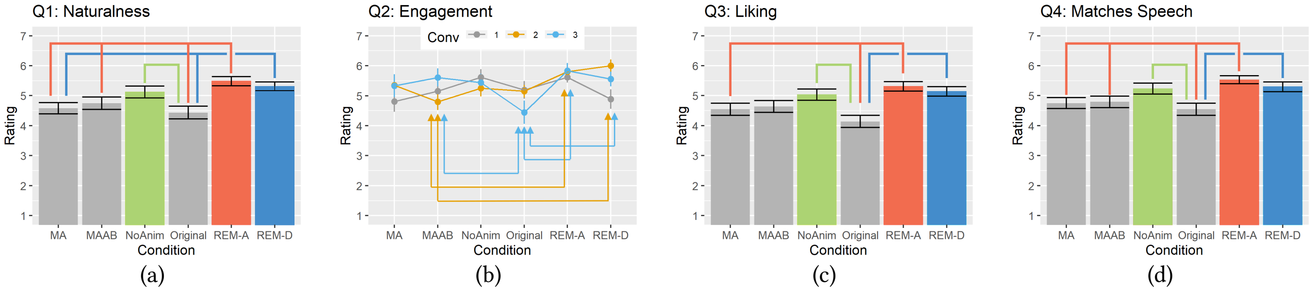
To evaluate the perceived quality of the eye animation, we conducted a user study in which participants watched and rated several videos of an avatar conversing with another person. The study was IRB approved.

4.2.1 Stimuli Creation and Procedure. We included six eye animation conditions in our study:

- preprocessed eye tracker data (Original)
- no eye animation (NoAnim)
- the Realistic Eye Movements (REM) Unity package[1] with adjusted parameters (REM-A)
- REM with default parameters (REM-D)
- our Motion Augmented model (MA) and
- our Motion Augmented + Audio (binary) model (MAAB)

We decided to not include the unprocessed original motion as it had obvious artifacts and as some (even if less detailed) processing would be possible in real time when using an HMD with eye tracking. So the processed captured motion was a higher bar. We included MA and MAAB in our evaluation because they had the lowest validation and test set losses (see Table 1) and to evaluate if adding audio as a feature makes a perceptually noticeable difference in the synthesized eye motion. We chose the "Realistic Eye Movements" package as a comparison as a highest quality procedural approach we could find. It generates microsaccades and saccades based on "Eyes Alive" by Lee et al. [2002] and has various user adjustable parameters, including saccade speed, magnitude, and

[1] Tore Knabe, Realistic Eye Movements, Unity Asset Store, 2021



(a) (b) (c) (d)

Figure 4: Our user study results. Color coded lines by animation condition indicate pairs with significantly different ratings in graphs (a), (c) and (d). Graph (b) shows significant differences between animation conditions color coded by conversation.

frequency. For REM-A, we set the ratio that the character looks at the other character to match that of our test set, at about 76%. The avatars we used in our evaluation were created using the avatar creation tool Ready Player Me[2]. The upper body and head movement as well as the blinking motion of all clips were animated using the captured data from the test set. So the only difference between conditions was the eyeball motion. The eyelids during blinks were animated using the piecewise function described by Duchowski et al. [2015]. The blink duration was taken from the eye tracker data. A minimum blink duration of 400 ms was set since it resulted in better looking eyelid animation for our stylized avatars. Eyelid saccades were also included in our clips, and were animated based on the vertical rotation of the eyes. The Oculus LipSync SDK[3] was used to animate the mouth based on the audio. Due to the large size of the eyes on the avatars used in the study, we also limit the amount the eyes can rotate vertically and horizontally.

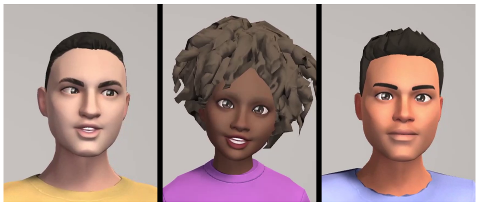


Figure 5: Avatars used in our evaluation. The animations were shown from the point of view of the interlocutor.

We selected three clips from our test set, each from a different conversation with a different performer (see Figure 5) and recorded them from the point of view of the other person. Clips were between 37 and 39 seconds long. We generated six versions of each clip, one for each eye animation condition, resulting in a total of 18 videos. We conducted the survey within subjects for the eye animation condition and between subjects for the conversation, meaning each participant watched six versions of the same conversation, corresponding to the six eye animation conditions. Participants were instructed to pay close attention to the eye animation. They first watched two videos for training. Then they watched the six versions (and two validation videos) in random order, viewing each video once. After each clip, participants rated their agreement with the following statements on a 7-point Likert scale:

Q1: The behavior of the avatar's eyes appeared natural.
Q2: The avatar seemed engaged in the conversation.
Q3: I liked the avatar's eye movements.
Q4: The avatar's eye movements matched its speech.

At the end of the study, there was an open response question asking participants to describe their thought process and any criteria considered while rating the animations.

4.2.2 Participants. Participants were recruited through Amazon Mechanical Turk. They received two US dollars after study completion. Data was collected from 120 participants, 40 per conversation. Completion time ranged from 7.8-61.4 minutes, with a median time of 13.5 minutes. We included two verification clips: a visual verification in which there were obvious errors in the eye animation and an audio verification where the voice told participants a specific answer they should select. We filtered out participants who responded neutrally or positively to whether the eye movement appeared natural, or if they did not select the answer as instructed by the audio verification clip. After filtering, there were 26 responses remaining for conversation 1, 20 for conversation 2, and 18 for conversation 3, for a total of 64 responses. Among the filtered responses, there were 27 females, 36 males, and 1 who preferred not to state their gender. The age range of participants was between 23 and 68 ($\mu = 36$).

4.2.3 Results. We compared the ratings from our study for each of the six eye animation conditions. The two independent variables in our analysis were the eye animation condition and the conversation. Each question (Q1-Q4) was a dependent variable. The mean ratings along with vertical standard error bars are graphed in Figure 4. Survey responses were analyzed using a multilevel linear mixed effects model for each question separately. Only significant results ($p < 0.05$) are presented.

There were significant main effects of animation condition for the responses to naturalness (Q1) ($\chi^2(5) = 31.97, p < 0.0001$), engagement (Q2) ($\chi^2(5) = 19.42, p < 0.005$), liking (Q3) ($\chi^2(5) = 35.02, p < 0.0001$), and speech matching (Q4) ($\chi^2(5) = 27.58, p < 0.0001$). We found no significant main effect of conversation for any of the questions. However, we did find an interaction between animation condition and conversation for engagement ($\chi^2(10) = 26.85, p < 0.05$). A post-hoc least squared means test with Tukey corrected p-values was done for each significant main effect found.

[2] readyplayer.me
[3] Oculus LipSync for Unity, 2021

For Q1, Q3, and Q4 the post-hoc tests revealed similar effects. For all three questions, the NoAnim condition, REM-A, and REM-D were rated significantly higher than Original and REM-A was rated significantly higher than MA. REM-D was only rated significantly higher than MA for Q1. For Q1 and Q4, REM-A was rated significantly higher than MAAB.

For engagement (Q2), there was a significant main effect of animation condition (REM-A was rated higher than Original, MA, and MAAB) as well as a significant interaction effect. Analyzing the ratings by conversation, we found significant differences between animation conditions for conversations 2 and 3. Both REM-A and REM-D were rated significantly higher than MAAB for conversation 2. REM-A, REM-D, and MAAB were rated significantly higher than Original for conversation 3. Figure 4 (b) shows significant differences between ratings for animation condition by conversation.

5 DISCUSSION AND LIMITATIONS

While our models were not able to reach the quality of the procedural animation and even received slightly lower average ratings than no animation, our results still lead to a series of insights.

An unexpected result from our user study is that the condition with **no eye animation was rated relatively high overall**. We speculate that the combination of head motions and blinking obscured the lack of eye motion and may even lead to the illusion of subtle eye movement. Perhaps in longer animation clips or in a face-to-face immersive scenario, static eyes would be more obvious or even detrimental to the experience.

Furthermore, we found that the **original, preprocessed motion was rated the lowest** for each measure in our user study. This result is not as surprising as it might first seem. The overall quality of the captured dataset, in terms of pupil detection confidence, is high at 0.92 on a scale from 0 (no pupil detected) to 1 (full confidence). However, while we processed the captured motion, the result still has slight remaining artifacts, for example from blinking, and is overly smooth at times. We estimate that our data quality was higher than typical captures since performers remained standing in one spot and mostly looked at each other. We compared our captured data to captures from a Meta Quest Pro and observed less noise and fewer artifacts in our data than the Quest Pro data.

The **procedural gaze models**, REM-A with an adjusted look-at parameter and REM-D with the default value, **were rated highest**, with REM-A rated slightly higher than REM-D overall. These models have the advantage of having correct saccadic profiles. We think that these animations were also rated highly because there is more perceived eye contact than with the other methods. Indeed, several comments from participants specifically mention eye contact as part of their consideration for their ratings.

Finally, **animations synthesized with our model were not rated highly** compared to the animations created with REM. We think a larger and better quality dataset might improve the results given this architecture. Our dataset with only 36 minutes may have trained the model to "play it safe" and generate eye motion that compensates for head movement without straying too much.

However, despite any issues mentioned so far, the original preprocessed motion should still perform best when it comes to **matching the speech** and we were curious to see if our model would learn to match the speech. Surprisingly, the conditions NoAnim, REM-A, and REM-D, which all do not take the speech into account at all were rated higher than the original motion for Q4 (speech matching). This observation as well as the similarity in ratings to Q1 (naturalness) and Q3 (liking) indicate that respondents might have considered naturalness or liking first in their rating for Q4. In other words, if they did not perceive the motion as being natural, then they would not perceive it to match the speech.

6 CONCLUSION AND FUTURE WORK

In this paper, we presented an RNN-based method to synthesize eye motions for conversations based on head motion and audio. We evaluated our results by comparing it to several conditions, such as realistic procedurally generated gaze. While our model was not rated as highly as procedural gaze, our model and user study provide a series of insights and conclusions.

We found that speaking may be correlated with a user's head motion, making it potentially possible to synthesize gaze without using audio features, which might enhance user privacy.

In our user study, we found that people are sensitive to perceived artifacts in eye motion, which is consistent with results from previous literature. Motions with artifacts were rated even worse than not having any animation, indicating that directly using eye-tracked motion in a virtual environment should be avoided.

In general, our results indicate that a procedural motion synthesis model may be sufficient for short interactions. However, future work should study longer dialogues where people make more nuanced judgements about the avatar's personality.

Our results suggest that engagement is correlated with eye contact. So perceived avatar engagement could potentially be controlled by tuning eye contact frequency, at least for short conversations.

Our results on speech matching indicate that a different type of experiment that clearly separates naturalness from speech matching would be necessary for future evaluations. A solution might be a recently presented study design such as the one used by Yoon et al. [2022], developed specifically to separate human-likeness (similar to naturalness) from appropriateness to speech regarding gestures. Interestingly, they compared a whole series of RNN based approaches and one procedural approach for synthesizing gestures and found that the procedural approach was considered most human-like, which might point towards more general limitations in our current use of RNNs for creating human motions.

Finally, there is potential in improving deep learning approaches as they should be able to synthesize eye motion that is more nuanced than a procedural model alone can. Future work using generative models, such as Generative Adversarial Networks (GANs), may lead to more plausible results than procedural methods, since they learn the distribution of both the input and output data and can generate new outputs indistinguishable from real examples. Furthermore, to learn plausible gaze patterns for complex or varied situations, e.g., a heated discussion, a large and diverse database would be necessary.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under grant numbers IIS-1652210 and IIS-1423189. The authors want to thank Nils Thürey for valuable comments and feedback.

REFERENCES

Andrea Abele. 1986. Functions of gaze in social interaction: Communication and monitoring. *Journal of Nonverbal Behavior* 10, 2 (1986), 83–101. https://doi.org/10.1007/BF01000006

Sean Andrist, Tomislav Pejsa, Bilge Mutlu, and Michael Gleicher. 2012. Designing Effective Gaze Mechanisms for Virtual Agents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. 705–714. https://doi.org/10.1145/2207676.2207777

Andrew Duchowski, Sophie Jörg, Aubrey Lawson, Takumi Bolte, Lech Świrski, and Krzysztof Krejtz. 2015. Eye Movement Synthesis with 1/f Pink Noise. In *Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games* (Paris, France) *(MIG '15)*. 47–56. https://doi.org/10.1145/2822013.2822014

Andrew T. Duchowski, Sophie Jörg, Tyler N. Allen, Ioannis Giannopoulos, and Krzysztof Krejtz. 2016. Eye Movement Synthesis. In *Proceedings of the 9th Biennial ACM Symposium on Eye Tracking Research & Applications* (Charleston, South Carolina) *(ETRA '16)*. 147–154. https://doi.org/10.1145/2857491.2857528

Ralf Engbert and Reinhold Kliegl. 2003. Microsaccades uncover the orientation of covert attention. *Vision research* 43, 9 (2003), 1035–1045.

Haytham M. Fayek. 2016. Speech Processing for Machine Learning: Filter banks, Mel-Frequency Cepstral Coefficients (MFCCs) and What's In-Between. https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html

Maia Garau, Mel Slater, Vinoba Vinayagamoorthy, Andrea Brogni, Anthony Steed, and M. Angela Sasse. 2003. The Impact of Avatar Realism and Eye Gaze Control on Perceived Quality of Communication in a Shared Immersive Virtual Environment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. 529–536. https://doi.org/10.1145/642611.642703

Ific Goudé, Alexandre Bruckert, Anne-Hélène Olivier, Julien Pettré, Rémi Cozot, Kadi Bouatouch, Marc Christie, and Ludovic Hoyet. 2023. Real-time Multi-map Saliency-driven Gaze Behavior for Non-conversational Characters. *IEEE Transactions on Visualization and Computer Graphics* 14, 1 (2023), 1–13. https://doi.org/10.1109/TVCG.2023.3244679

Erdan Gu and Norman I Badler. 2006. Visual attention and eye gaze during multiparty conversations with distractions. In *Intelligent Virtual Agents: 6th International Conference, IVA 2006, Marina Del Rey, CA, USA, August 21-23, 2006. Proceedings 6*. Springer, 193–204.

Simon Ho, Tom Foulsham, and Alan Kingstone. 2015. Speaking and Listening with the Eyes: Gaze Signaling during Dyadic Interactions. *PLOS ONE* 10, 8 (08 2015), 1–18. https://doi.org/10.1371/journal.pone.0136905

Tomoyori Iwao, Daisuke Mima, Hiroyuki Kubo, Akinobu Maejima, and Shigeo Morishima. 2012. Analysis and Synthesis of Realistic Eye Movement in Face-to-Face Communication. In *ACM SIGGRAPH 2012 Posters* (Los Angeles, California) *(SIGGRAPH '12)*. Article 87, 1 pages. https://doi.org/10.1145/2342896.2342999

Tomoyori Iwao, Daisuke Mima, Hiroyuki Kubo, Akinobu Maejima, and Shigeo Morishima. 2013. Generating Eye Movement during Conversations Using Markov Process. In *ACM SIGGRAPH 2013 Posters* (Anaheim, California) *(SIGGRAPH '13)*. Article 6, 1 pages. https://doi.org/10.1145/2503385.2503392

Aobo Jin, Qixin Deng, Yuting Zhang, and Zhigang Deng. 2019. A Deep Learning-Based Model for Head and Eye Motion Generation in Three-Party Conversations. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 2, 2, Article 9 (jul 2019), 19 pages. https://doi.org/10.1145/3340250

Sophie Jörg, Andrew Duchowski, Krzysztof Krejtz, and Anna Niedzielska. 2018. Perceptual Adjustment of Eyeball Rotation and Pupil Size Jitter for Virtual Characters. 15, 4, Article 24 (Oct. 2018), 13 pages. https://doi.org/10.1145/3238302

Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-Based Interaction *(UbiComp '14 Adjunct)*. 1151–1160. https://doi.org/10.1145/2638728.2641695

Alex Klein, Zerrin Yumak, Arjen Beij, and A. Frank van der Stappen. Oct 28, 2019. Data-driven Gaze Animation using Recurrent Neural Networks. In *Motion, Interaction and Games* (Newcastle upon Tyne, United Kingdom) *(MIG '19)*. ACM, 1–11. http://dl.acm.org/citation.cfm?id=3360054

Rakshit Kothari, Zhizhuo Yang, Christopher Kanan, Reynold Bailey, Jeff B. Pelz, and Gabriel J. Diaz. 2020. Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities. *Scientific reports* 10, 1 (2020), 1–18. https://doi.org/10.1038/s41598-020-59251-5

Binh H. Le, Xiaohan Ma, and Zhigang Deng. 2012. Live Speech Driven Head-and-Eye Motion Generators. *IEEE Transactions on Visualization and Computer Graphics* 18, 11 (2012), 1902–1914. https://doi.org/10.1109/TVCG.2012.74

Sooha Park Lee, Jeremy B. Badler, and Norman I. Badler. 2002. Eyes Alive. *ACM Trans. Graph.* 21, 3 (July 2002), 637–644. https://doi.org/10.1145/566654.566629

Xiaohan Ma and Zhigang Deng. March 2009. Natural Eye Motion Synthesis by Modeling Gaze-Head Coupling. In *2009 IEEE Virtual Reality Conference*. IEEE, 143–150. https://ieeexplore.ieee.org/document/4811014

Aline Normoyle, Jeremy B. Badler, Teresa Fan, Norman I. Badler, Vinicius J. Cassol, and Soraia R. Musse. 2013. Evaluating Perceived Trust from Procedurally Animated Gaze. In *Proceedings of Motion on Games* (Dublin 2, Ireland) *(MIG '13)*. 141–148. https://doi.org/10.1145/2522628.2522630

Oyewole Oyekoya, William Steptoe, and Anthony Steed. 2009. A Saliency-Based Method of Simulating Visual Attention in Virtual Scenes. In *Proceedings of the 16th ACM Symposium on Virtual Reality Software and Technology* (Kyoto, Japan) *(VRST '09)*. 199–206. https://doi.org/10.1145/1643928.1643973

Tomislav Pejsa, Daniel Rakita, Bilge Mutlu, and Michael Gleicher. 2016. Authoring Directed Gaze for Full-Body Motion Capture. *ACM Transactions on Graphics* 35, 6, Article 161 (dec 2016), 11 pages. https://doi.org/10.1145/2980179.2982444

Christopher Peters and Adam Qureshi. 2010. A head movement propensity model for animating gaze shifts and blinks of virtual characters. *Computers & Graphics* 34, 6 (2010), 677–687. https://www.sciencedirect.com/science/article/pii/S0097849310001408 ID: 271576.

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, 103–108. https://doi.org/10.1109/ISMAR-Adjunct.2018.00044

K. Ruhland, S. Andrist, J. B. Badler, C. E. Peters, N. I. Badler, M. Gleicher, B. Mutlu, and R. McDonnell. 2014. *Look me in the Eyes: A Survey of Eye and Gaze Animation for Virtual Agents and Artificial Systems*. https://doi.org/10.2312/egst.20141036

Kerstin Ruhland, Katja Zibrek, and Rachel McDonnell. 2015. Perception of Personality through Eye Gaze of Realistic and Cartoon Models. In *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception* (Tübingen, Germany) *(SAP '15)*. 19–23. https://doi.org/10.1145/2804408.2804424

Sven Seele, Sebastian Misztal, Helmut Buhler, Rainer Herpers, and Jonas Schild. 2017. Here's Looking At You Anyway! How Important is Realistic Gaze Behavior in Co-located Social Virtual Reality Games?. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. 531–540. https://doi.org/10.1145/3116595.3116619

Teflpedia. [n. d.]. Teflpedia conversation questions. Retrieved July 19, 2021 from http://teflpedia.com/Category:Teflpedia_conversation_questions

Vinoba Vinayagamoorthy, Maia Garau, Anthony Steed, and Mel Slater. 2004. An eye gaze model for dyadic interaction in an immersive virtual environment: Practice and experience. In *Computer Graphics Forum*, Vol. 23. Wiley Online Library, 1–11. https://doi.org/10.1111/j.1467-8659.2004.00001.x

Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2022. The GENEA Challenge 2022: A Large Evaluation of Data-Driven Co-Speech Gesture Generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction* (Bengaluru, India) *(ICMI '22)*. 736–747. https://doi.org/10.1145/3536221.3558058