Eye Movement in a Controlled Dialogue Setting

David Dembinsky
david.dembinsky@dfki.de
RPTU Kaiserslautern-Landau
Kaiserslautern, Germany
DFKI GmbH
Kaiserslautern, Germany

Ko Watanabe
ko.watanabe@dfki.de
RPTU Kaiserslautern-Landau
Kaiserslautern, Germany
DFKI GmbH
Kaiserslautern, Germany

Andreas Dengel
andreas.dengel@dfki.de
RPTU Kaiserslautern-Landau
Kaiserslautern, Germany
DFKI GmbH
Kaiserslautern, Germany

Shoya Ishimaru
ishimaru@omu.ac.jp
Osaka Metropolitan University
Osaka, Japan
DFKI Lab Japan
Osaka, Japan

ABSTRACT

Designing realistic eye movements for animated avatars poses a challenge, as gaze behavior is predominantly unconscious. Accurately modulating those movements is crucial to avoid the Uncanny Valley. The human gaze exhibits different characteristics in conversations, depending on speaking or listening. Albeit these distinctions are known, data for synthesizing eye movement models suitable for avatars is scarce. This research introduces a novel dataset involving human gaze behavior during remote screen conversations. The data are collected from 19 participants, offering 4 hours of gaze data labeled as *Speaking* and *Listening*. Our data analysis substantiates prior knowledge of gaze behavior while providing new insights through higher precision. Furthermore, we demonstrate the dataset's suitability for machine learning algorithms by training a classifier, achieving 88.1% binary classification accuracy.

CCS CONCEPTS

• **Computing methodologies → Activity Recognition**; **Animation**.

KEYWORDS

Eye Movement, Gaze Synthesis, Data Collection, Animated Avatars

ACM Reference Format:
David Dembinsky, Ko Watanabe, Andreas Dengel, and Shoya Ishimaru. 2024. Eye Movement in a Controlled Dialogue Setting. In *2024 Symposium on Eye Tracking Research and Applications (ETRA '24), June 4–7, 2024, Glasgow, United Kingdom.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3649902.3653337

ETRA '24, June 4–7, 2024, Glasgow, United Kingdom

ACM ISBN 979-8-4007-0607-3/24/06…$15.00
https://doi.org/10.1145/3649902.3653337

1 INTRODUCTION

Conversational animated avatars are employed in a wide range of scenarios, appearing as teachers to boost educational success [Amemiya et al. 2022; Pataranutaporn et al. 2022], as physicians to raise patients' motivation for self-disclosure [Lucas et al. 2014; Moriuchi 2022], or as intermediaries to communicate with autistic children [Kellems et al. 2020].

An avatar's success usually depends on believable acting, implying that its movements must match its appearance. If their behavior is less human than their appearance suggests, they will suffer from the Uncanny Valley, resulting in less acceptance by human observers [Mori et al. 2012]. Since animation technologies make avatars look more and more realistic and humanoid robots' appearances get closer to real humans, it is essential to design even the most subtle movements accurately. Various approaches rely on statistical, rule-based models [Le et al. 2012; Lee et al. 2002], while others leverage neural networks [Cudeiro et al. 2019; Tian et al. 2019] or a combination of both [Edwards et al. 2020] to generate facial movements for avatars. Most previous research focuses on correctly dubbing the mouth for text or audio input, excluding eyes or leaving their control to simple models.

To account for the intricacies of human gaze behavior in general and especially during conversations, deep learning-powered generators are a promising technique. However, they require large amounts of curated data to train. To our knowledge, there is no dataset publicly available, that accurately captures fine-grained movements for conversations and is suitable for deep-learning models. In this work, we introduce a new dataset that serves to fill this gap [1]. We tracked the gaze of 19 participants while they talked and listened to their partner in a video call. We extract fixations and saccades from the data and analyze them statistically, comparing it to the previous work of Lee et al., who also analyzed the movements of eyes during conversation [Lee et al. 2002]. Further, we create a dataset for machine learning and prove its applicability by training neural network classifiers that predict whether a sequence of fixations was created when talking or listening and achieve an accuracy of 88.1%.

[1] available at https://github.com/psyberlab/eye-gaze-during-conversation-dataset

In this article, we will first show the significance of eye movement in communication and introduce previous work on conversational avatars. Then we provide details on how we collected and preprocessed the gaze data, before analyzing the data statistically and comparing it to previous work. Subsequently, the classification using neural networks is presented. Finally, we discuss all the significant limitations of the data we collected.

2 RELATED WORK

2.1 Significance of Eye Movements in Communication

During a conversation, the majority of communication takes place non-verbally between conversants [Watanabe et al. 2021], with eye contact being a fundamental behavior in social interactions [Kleinke 1986; Zhang et al. 2017]. The gaze may facilitate engaging in conversations by expressing shared attention [Wohltjen and Wheatley 2021] or is used to manage who speaks next in multi-person interactions [Jokinen et al. 2009]. Eye movement is triggered more frequently than head movement, especially during conversations, where fixation points are locally nearby [Duchowski and Jörg 2015; Vrzakova et al. 2016]. Previous experiments have indicated that humans receive avatars exhibiting realistic eye movements better [Garau et al. 2003; Heylen et al. 2002; Lee et al. 2002]. Heylen et al. showed that a very simplistic rule set deciding whether to look at the partner or to gaze away increases the perceived quality as long as it is designed to mimic realistic movements. Garau et al. demonstrated that a realistic appearance of the avatar comes with an increased demand for accurate eye movements, as random motions result in a less engaging and natural-looking avatar. Similarly, Lee et al. compared the acceptance of conversational avatars that did not move their eyes to those that used random motions or a statistical model. They found the latter to appear most interested, engaged, and lively to human observers.

2.2 Avatar Generation

Since the movement of the lower face is linked directly to the process of forming speech, many researchers focused on accurate lip and mouth movement, using a wide range of techniques like Variational Autoencoders or LSTM architectures, ignoring the eyes [Cudeiro et al. 2019; Richard et al. 2021; Tian et al. 2019; Villanueva Aylagas et al. 2022].

In contrast to the mouth, the movement of the eyes is not immediately linked to the words spoken, forcing researchers to concentrate on other sources that influence them [Le et al. 2012; Lee et al. 2002; Ma and Deng 2009; Masuko and Hoshino 2006]. Lee et al. collected statistical information on eye movements during conversation and used them to create a rule-based model that randomly selects actions to match the statistics collected. Similarly, Masuko and Hoshino used a single set of rules to simultaneously control head and eye movements. Ma and Deng take the underlying coupling between head and eye agitation to statistically predict the latter from the input of a head motion sequence. Le et al. stacked multiple statistical models that control the head and eyes but rely on expensive motion-capture data to fit their models. Noteworthy is the recent work of Edwards et al., who combined various models to control the avatar's entire face. It automatically synchronizes mouth movements to audio input in different languages and controls head, gaze, blinks, and further facial movements. The usage of their model in the video game *Cyberpunk 2077* has shown the effectiveness and adaptability of their model to different languages and speaking styles. On the flip side, the model requires a lot of manual labor to prepare the data, including a tagged transcript of the audio [Edwards et al. 2020].

Table 1: The statistical information gathered from the participants, based on their self-disclosure. We only display the information of those participants we included in the later analysis.

Gender		**Nationality/Ethnicity**	
Male	10 (53%)	Western Europe	6 (32%)
Female	7 (37%)	India	9 (47%)
Non-binary	2 (11%)	Other	4 (21%)
Wearing glasses		**Knew their partner**	
Yes	8 (42%)	Yes	7 (37%)
No	11 (58%)	No	12 (63%)
Eye diseases or mental health issues		**Experience with Eye-Tracking studies**	
None	16 (84%)	Yes	15 (79%)
ADHD	3 (16%)	No	4 (21%)

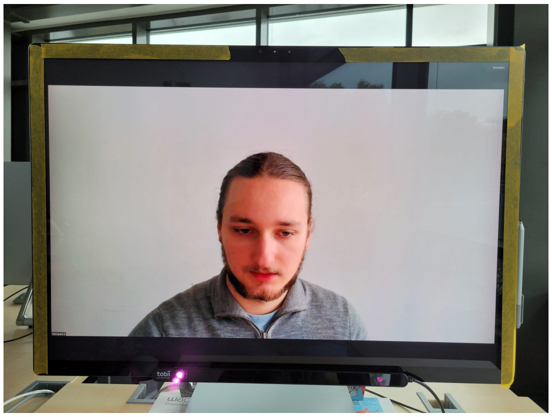


Figure 1: The data collection's setup, from one participant's view. The orange border is a reminder, not to look outside the screen.

3 DATASET

3.1 Participants

To conduct the study, we recruited 22 people from the university. We obtained consent from the participants in accordance with the General Data Protection Regulation (GDPR) and participants had the option to withdraw from the experiment at any point. To avoid unconscious bias among participants, a priori we asked them to exclusively focus on the conversation. After the data collection, we explained that their data could be used for the eye generation of an avatar. As compensation, every participant was offered a 10 Euro Amazon gift card, which one participant declined.

We had to exclude three participant's data from the analysis. One reported not having paid attention to the experiment, one reported suffering from a vision-impeding disease, and one reported suffering from a depressive disorder. The participants were paired randomly, which resulted in four pairs knowing their counterparts. In retrospect to the study, we collected some statistical information from the participants, presented in Table 1.

3.2 Study Setup

We collected the gaze behavior in a controlled lab condition where two participants talked to each other in an online video call. First, we placed each one in front of a large computer screen[2] in different rooms, centering their heads on the built-in camera and calibrating the Eye-Tracker at an approximate head-to-screen distance of 67.5cm. We used the Tobii 4c Eye-Tracker, which samples at 90 Hz and can track the gaze inside the boundaries of the PC screen and, to some extent, beyond. We marked the borders of the monitor using orange duct tape to remind the participants not to look outside the screen, as can be seen in Figure 1. We recorded the entire video call, both in video and audio, to ease the labeling process. To collect data, we asked the participants to engage with each other in three rounds, each announced by the supervisor. The three rounds invited the participants to introduce themselves, tell a story from their everyday lives, and finally discuss some projects they are working on. They took turns, letting one participant talk freely from 30 seconds to five minutes, while the other listened before switching roles. At the end of each round, the participants were allowed to enter a dialogue and exchange questions if desired. The participants should not monitor their speaking time, and they were allowed to deviate from the prescribed topic to feel as unrestricted as possible.

3.3 Methods

Using the recordings, we labeled the data by extracting the timestamps of each speaker's turn from the recordings. We assigned one of the three labels *Speak*, *Listen*, and *Dialogue*. The first two being opposites, describe prolonged sequences of speaking and listening without interruption. The latter is assigned to both participants simultaneously, describing phases where both participants converse in short succession, switching between speaking and listening frequently.

Extracting the gaze positions was limited by the Tobii Eye-Tracker's capabilities, as it only tracks the gaze reliably inside the screen. Looking too far outside or occluding the eyes through gestures or extreme head poses will result in missing positional data (denoted as NaN), rendering data points useless. We interpolate brief gaps of less than 50ms linearly, which is less than half a fixation. Following, we did Fixation-Saccade extraction by using *Dispersion-Based Identification (I-DT)*, where subsequent gazes close to each other are counted as one fixation [Salvucci and Goldberg 2000]. We used a sliding window, counting the window as a fixation if the calculated dispersion was below a threshold and included the following points, as long as their dispersion did not pass the threshold. We assume the Tobii 4c to have similar properties to the older EyeX [3], being found to have an accuracy of < 0.6° [Gibaldi et al. 2017]. We select a threshold of 30px (or 0.85cm) as diameter, corresponding to an eyeball rotation of ~ 0.72°, to stay above the resolution of the eye-tracker. This threshold allows us to argue about individual fixation points, ignoring micro-saccadic movement, which is defined below 0.5° [Poletti and Rucci 2016]. We require each fixation to last at least 0.1s. An example of the process can be seen in Figure 2.

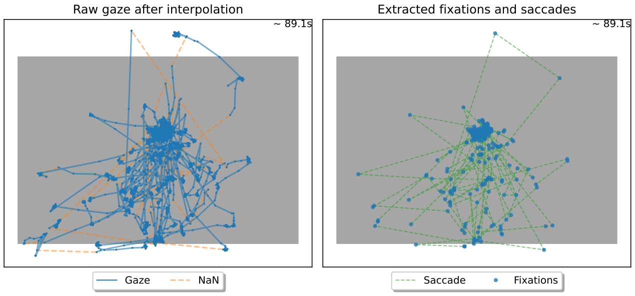


Figure 2: The transformation of the raw gaze data to fixation-saccade data. The raw data on the left has more small movements than the extracted data on the right. The gray box marks the dimensions of the screen.

4 RESULTS

When analyzing the data, we divided each recording into segments according to the label. However, the labeling process is only meaningful up to a certain granularity, as the beginning and end of speaking are characterized by thinking about what to say and waiting for the other participant or supervisor to seize the word. Because we want to extract the eyes while the conversation runs smoothly, we remove the beginning and end (3 seconds) of the data for each round. To further exclude missing gaze values, or NaNs, from our data, we divide the round into various segments, cutting and removing each NaN occurrence. With this data, we perform a statistical analysis and machine learning analysis.

4.1 Statistical analysis

4.1.1 Data. The analyzed data was built from the extracted fixations, removing the NaNs. Each segment contained a varying amount of fixations, with Table 2 exhibiting the exact dataset composition. We will not explicitly include an in-depth analysis for the *Dialogue* class, as it is, by construction, a middle-ground between speaking and listening.

4.1.2 Gaze Position. Comparing the distribution of fixations around the screen, we notice that the distribution is similar to previous findings [Lee et al. 2002], with the fixations being much more spread out when speaking than in listening mode, as can be seen in Figure 3. When listening, 90% of the fixations are less than 5.2cm apart from the center point, which is less than 4.4° eyeball rotation, while during speaking a circle with radius 6.8cm (5.7°) only contains 50% of the data. We assumed that the partner's face is the center of

[2] Microsoft Surface Studio 1. Gen with 637.35 × 438.90 mm screen size.

[3] as described in this post: https://developer.tobii.com/community/forums/topic/accuracy-of-tobii-4c/

Table 2: The composition of the statistical dataset, where snippets contain a varying amount of fixations.

	Speak	Listen	Dialogue
Total Number of Fixations	15209	17717	7007
Total Time covered (Fix.+Sacc.)	6715.5s	7405.6s	3200.9s
Min/Max/Mean Number of Fixations per Segment	1/75/3.44	1/61/4.42	1/52/3.51
Min/Max/Mean Number of gaze points per Fixation	2/144/20.05	10/284/23.60	10/157/22.62
Min/Max/Mean Time per Fixation	0.1/1.6/0.2s	0.1/3.2/0.3s	0.1/1.7/0.3s

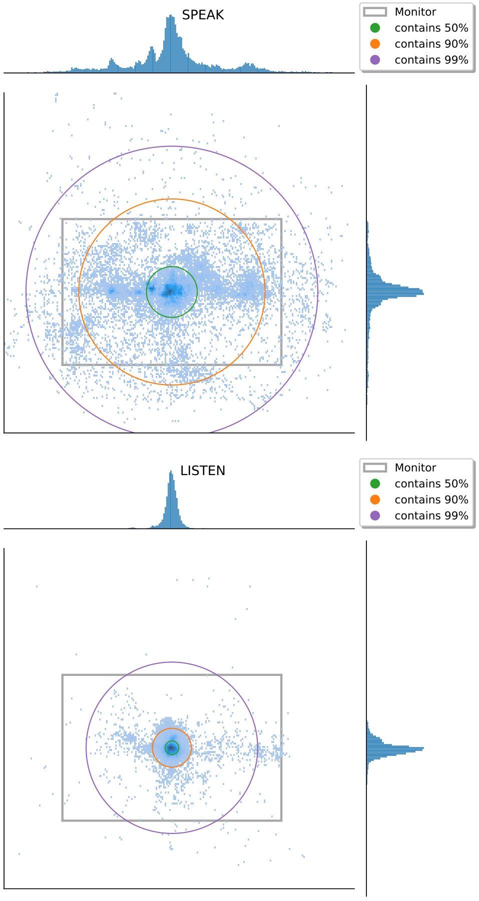


Figure 3: The heatmaps and histograms show the fixations' distribution. Each participant's gaze is centered individually to accommodate for varying positioning of the partner's face on the screen. The gray box marks the boundaries of the monitor. The circles indicate how much data lies inside of them. We can observe that the gaze extends out more during speaking compared to listening.

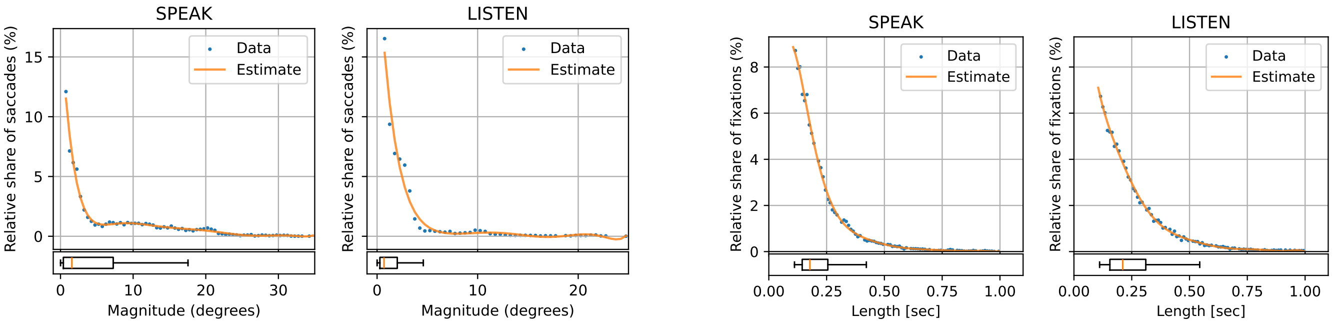
attention and for each participant individually centered the gaze, by calculating the mean of all fixational points per participant.

4.1.3 Fixation & Saccades. The statistics of the saccades between fixations reinforce these findings. For both classes, smaller saccades are most prominent. However, while there are few saccades above 10° when listening, there is a tail stretching up to around 20° in the *Speak* distribution. The listening distribution follows the same shape as the one reported by Lee et al., although we note a steeper downward slope in ours [Lee et al. 2002].

In opposition to their findings, we found only subtle differences in the fixation duration between the two modes. In general, they report longer fixation durations, differing between mutual gaze and gaze away. For speaking, the differences amount to a factor of two on the x-axis, corresponding to a one-second shift, and for listening, the factor totals 10 (around nine seconds). The visualizations for saccade magnitude and fixation duration can be seen in Figure 4, also denoting the bin width δ. We model the four distributions through polynomials, with the exact parameters given in the supplementary materials.

Additionally, we extract the directions in which saccades move from fixation to fixation. We sort them according to the direction of the saccade itself, independent of the monitor region the participant is looking at. We distinguish between horizontal, vertical, and diagonal movements. Figure 5 shows the relative frequency of different movement directions, opposing listening and speaking behavior. Further, we distinguish between big magnitudes with more than 2° (83 px) eyeball rotations and small ones. As seen previously, the fraction of relatively small saccades is higher when listening than when speaking. During listening and speaking, the small saccades are equally as likely in each of the eight directions. Larger saccades exhibit different behaviors for the two modes: When speaking, most large movements are in the horizontal direction, while there seems to be no difference between the frequency of vertical and diagonal movements. In listening mode large horizontal and vertical movements are more dominant than diagonal ones, but there is no difference between left-right and up-down movement. This later finding is in line with the results from Lee et al., whereas they did not report large horizontal movements when speaking.

4.1.4 Comparison. While the general statistics of our dataset tally those reported by Lee et al., the apparent differences can be traced back to various methodical differences. First, they used a worn eye-tracker, ignoring movements of the head, while we used a stationary one that extracts the gaze position on the screen. Hence, we extract more horizontal movement, which is associated with head movement, opposing to their method which ignored head movement. Second, we processed the raw data by extracting fixation-saccade movements through *I-DT*, while they used a median filter to stabilize the gaze positions. This smoothing may remove valuable information on subtle motion, combining gaze data into a single fixation, that we would have unraveled into several nearby ones. This explains why we find more of the short saccades since we can differentiate between different fixations even when close by.



(a) Saccade Magnitude ($\delta = 0.5°$)

(b) Fixation Duration ($\delta = 0.01$**s**)

Figure 4: The histograms show the magnitude of saccades and the duration of fixations, comparing speaking and listening modes. The orange line is a polynomial fitting curve. δ denotes the bin width. During speaking, the magnitude of saccades ranges up to around $20°$, while they stay below $10°$ during listening. The fixations' lengths show no significant differences.

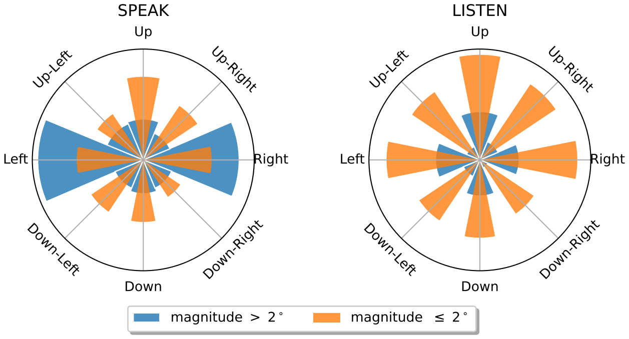


Figure 5: The distribution of saccade directions when speaking and listening. Blue shows big saccades and orange shows small ones. Small saccades are evenly distributed. Larger ones prefer horizontal movement when speaking and axis-aligned motion when listening.

If they accumulate many different fixations that are spatially and timely close into one single fixation, this grows its duration, leading to higher fixation times. This also explains their preference for axis-aligned movement, as it appears in bigger saccades. Hence, they miss the information of the equally distributed small saccades. Finally, they only analyzed data of one single conversation between two people, lasting nine minutes in total, while we had a total of 19 participants covering 4 hours of speaking and listening data, making our data more robust and statistically reliable.

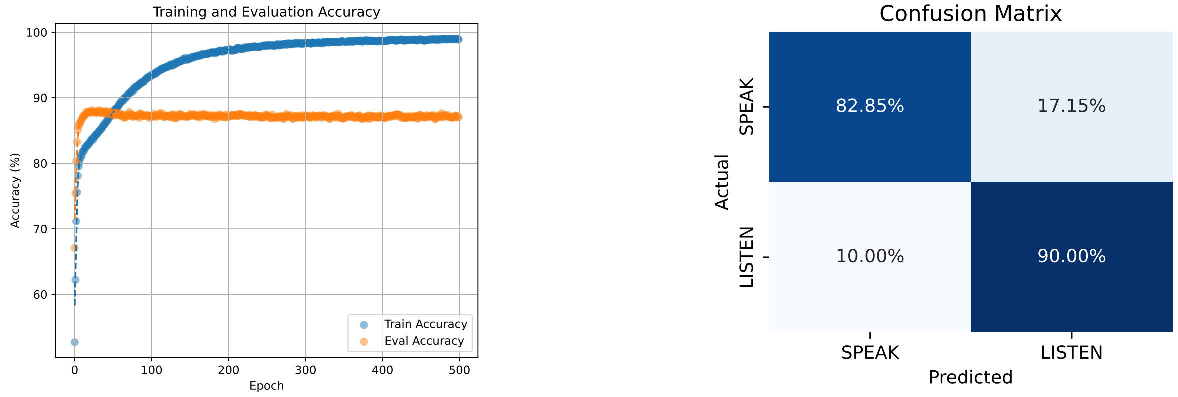
4.2 Machine Learning Analysis

To show the suitability of our collected data for machine learning, we transform it into a format usable for ML. Therefore, we assemble windows of $W = 10$ subsequent fixations, including their length, the gap before the next one starts, and its x and y position. This results in data points of the shape $(4, W)$. Starting from the first, we shifted each window one fixation ahead. We make sure that windows do not overlap between training and evaluation sets. This results in some randomness in the exact dataset's composition. We average over multiple runs, to achieve a more robust estimate of the performance. The datasets consist of 2698.5 ± 45.5 windows that span an average of 5.6 seconds per window.

We train a neural network classifier to distinguish between speaking and listening data. The architecture consists of two convolutional inputs with channel-wise and time-wise convolution side-by-side, followed by two linear layers of width 128 and 32. Dropout layers surround the second hidden layer. The model is trained over 500 epochs with a batch size of 40 windows, using the Adam optimizer with learning rate $\alpha = 0.0005$ and cross-entropy loss. The results are averaged over 20 iterations. The training accuracy converges towards 100%, while the evaluation accuracy converges towards 87%, maxing out after around 25 epochs at 88.1%. The behavior indicates insufficient generalization from this point on. Figure 6 presents the model's accuracy curve and confusion matrix. For increased interpretability, we normalized the confusion matrix over the actual label (row). It reveals a small difference between classifying the two different labels, being worse at speaking samples. Overall we have shown, that the dataset is suitable to train neural network classifiers.

5 LIMITATION

Our dataset presents a valuable contribution to future research on human gaze behavior during conversations. Nevertheless, we are aware of some limitations inherent to our data. The study setup itself was restricted and controlled, as participants were placed in separate rooms and had to talk through a video call rather than a more realistic face-to-face communication. This limits the participants' possibilities, to effectively use and react to gestures and other non-verbal cues. Additionally, the labeling process was not fine-grained and could be enhanced in multiple aspects. First, the time annotation relied on the timestamps in the recording of the video call, which had to be extracted by hand and were only given in seconds, so the partition always happens between two seconds. Second, the labels do not take any personal factors into account, like engagement, stress, and similar, which might influence eye movements.



(a) Accuracy Curve

(b) Confusion Matrix

Figure 6: The performance graphs for our trained classifier. The results are achieved by averaging over multiple runs. The model shows signs of overfitting, with speaking samples beeing slightly more difficult.

Regarding technical aspects, the data collected using the Tobii Eye Tracker contained many NaN values for positional data. To tackle those, we had to make several assumptions, interpolating gaps of less than 50ms linearly and segmenting the data on longer NaN gaps. Additionally, the Tobii Eye Tracker only recognizes the screen position the participant is looking at. We do not take into account, what exactly he is looking at, for example, the other participant's eyes, mouth, or other things. Since the other participant may have altered their position during the conversation, eye movements may be provoked by this, although our data does not take such scenarios into account.

6 CONCLUSION

This research presented the collection and preparation of a new dataset which consists of human gaze behavior during conversations. The data was collected from a heterogeneous group of young people in video-call talk using eye trackers. The tracker followed their gaze on the screen, which displayed their partner's face. We extracted the fixations and saccades that define the eyes' behavior from the raw data collected. Analyzing the data statistically, we found that the results are substantially consistent with previous research on eye behavior during conversations, enhancing the data available to this point. By training a neural network classifier on the fixation statistics, we proved the applicability of our data for machine learning tasks. We achieved an average binary classification accuracy of 88.1%. We finalize the data collection by pointing out several process hurdles and justifying assumptions. Those remarks may serve for future research to enhance the collected data. In summary, this research contributes a novel dataset for studying human gaze behavior during conversations and verifies the data's suitability for machine learning applications. It allows future researchers to generate realistic, fresh eye movements and visualize them using the collected data.

ACKNOWLEDGMENTS

This work was supported by DFG International Call on Artificial Intelligence "Learning Cyclotron" (Proj. Nr.: 442581111).

REFERENCES

Tomohiro Amemiya, Kazuma Aoyama, and Kenichiro Ito. 2022. Effect of Face Appearance of a Teacher Avatar on Active Participation During Online Live Class. In *Human Interface and the Management of Information: Applications in Complex Technological Environments*, Sakae Yamamoto and Hirohiko Mori (Eds.). Springer International Publishing, Cham, 99–110. https://doi.org/10.1007/978-3-031-06509-5_7

Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J. Black. 2019. Capture, Learning, and Synthesis of 3D Speaking Styles. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10093–10103. https://doi.org/10.1109/CVPR.2019.01034

Andrew T Duchowski and Sophie Jörg. 2015. Modeling physiologically plausible eye rotations. In *Proceedings of Computer Graphics International*.

Pif Edwards, Chris Landreth, Mateusz Popławski, Robert Malinowski, Sarah Watling, Eugene Fiume, and Karan Singh. 2020. JALI-Driven Expressive Facial Animation and Multilingual Speech in Cyberpunk 2077. In *ACM SIGGRAPH 2020 Talks* (Virtual Event, USA) *(SIGGRAPH '20)*. Association for Computing Machinery, New York, NY, USA, Article 60, 2 pages. https://doi.org/10.1145/3388767.3407339

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)*. Association for Computing Machinery, New York, NY, USA, 529–536. https://doi.org/10.1145/642611.642703

Agostino Gibaldi, Mauricio Vanegas, Peter J Bex, and Guido Maiello. 2017. Evaluation of the Tobii EyeX Eye tracking controller and Matlab toolkit for research. *Behavior research methods* 49 (2017), 923–946.

Dirk Heylen, Ivo Van Es, Anton Nijholt, and Betsy van Dijk. 2002. Experimenting with the gaze of a conversational agent. In *Proceedings international CLASS workshop on natural, intelligent and effective interaction in multimodal dialogue systems*. EU CLASS, 93–100.

Kristiina Jokinen, Masafumi Nishida, and Seiichi Yamamoto. 2009. Eye-Gaze Experiments for Conversation Monitoring. In *Proceedings of the 3rd International Universal Communication Symposium* (Tokyo, Japan) *(IUCS '09)*. Association for Computing Machinery, New York, NY, USA, 303–308. https://doi.org/10.1145/1667780.1667843

Ryan O Kellems, Cade Charlton, Kjartan Skogly Kversøy, and Miklós Győri. 2020. Exploring the use of virtual characters (avatars), live animation, and augmented reality to teach social skills to individuals with autism. *Multimodal Technologies and Interaction* 4, 3 (2020), 48. https://doi.org/10.3390/mti4030048

Chris L Kleinke. 1986. Gaze and eye contact: a research review. *Psychological bulletin* 100, 1 (1986), 78.

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. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques* (San Antonio, Texas) *(SIGGRAPH '02)*. Association for Computing Machinery, New York, NY, USA, 637–644. https://doi.org/10.1145/566570.566629

Gale M. Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior* 37 (2014), 94–100. https://doi.org/10.1016/j.chb.2014.04.043

Xiaohan Ma and Zhigang Deng. 2009. Natural Eye Motion Synthesis by Modeling Gaze-Head Coupling. In *2009 IEEE Virtual Reality Conference*. 143–150. https://doi.org/10.1109/VR.2009.4811014

Soh Masuko and Junichi Hoshino. 2006. Generating head–eye movement for virtual actor. *Systems and Computers in Japan* 37, 12 (2006), 33–44. https://doi.org/10.1002/scj.20513 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/scj.20513

Masahiro Mori, Karl F. MacDorman, and Norri Kageki. 2012. The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine* 19, 2 (2012), 98–100. https://doi.org/10.1109/MRA.2012.2192811

Emi Moriuchi. 2022. Leveraging the science to understand factors influencing the use of AI-powered avatar in healthcare services. *Journal of Technology in Behavioral Science* 7, 4 (2022), 588–602. https://doi.org/10.1007/s41347-022-00277-z

Pat Pataranutaporn, Joanne Leong, Valdemar Danry, Alyssa P. Lawson, Pattie Maes, and Misha Sra. 2022. AI-Generated Virtual Instructors Based on Liked or Admired People Can Improve Motivation and Foster Positive Emotions for Learning. In *2022 IEEE Frontiers in Education Conference (FIE)*. 1–9. https://doi.org/10.1109/FIE56618.2022.9962478

Martina Poletti and Michele Rucci. 2016. A compact field guide to the study of microsaccades: Challenges and functions. *Vision Research* 118 (2016), 83–97. https://doi.org/10.1016/j.visres.2015.01.018 Fixational eye movements and perception.

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

Dario D Salvucci and Joseph H Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*. 71–78.

Guanzhong Tian, Yi Yuan, and Yong Liu. 2019. Audio2Face: Generating Speech/Face Animation from Single Audio with Attention-Based Bidirectional LSTM Networks. In *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. 366–371. https://doi.org/10.1109/ICMEW.2019.00069

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* 41, 8, 255–265. https://doi.org/10.1111/cgf.14640 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14640

Hana Vrzakova, Roman Bednarik, Yukiko I Nakano, and Fumio Nihei. 2016. Speakers' head and gaze dynamics weakly correlate in group conversation. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications* (Charleston, South Carolina) *(ETRA '16)*. Association for Computing Machinery, New York, NY, USA, 77–84. https://doi.org/10.1145/2857491.2857522

Ko Watanabe, Yusuke Soneda, Yuki Matsuda, Yugo Nakamura, Yutaka Arakawa, Andreas Dengel, and Shoya Ishimaru. 2021. Discaas: Micro behavior analysis on discussion by camera as a sensor. *Sensors* 21, 17 (2021), 5719.

Sophie Wohltjen and Thalia Wheatley. 2021. Eye contact marks the rise and fall of shared attention in conversation. *Proceedings of the National Academy of Sciences* 118, 37 (2021), e2106645118. https://doi.org/10.1073/pnas.2106645118

Xucong Zhang, Yusuke Sugano, and Andreas Bulling. 2017. Everyday Eye Contact Detection Using Unsupervised Gaze Target Discovery. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (Québec City, QC, Canada) *(UIST '17)*. Association for Computing Machinery, New York, NY, USA, 193–203. https://doi.org/10.1145/3126594.3126614