© 2025 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Exploring Gaze Dynamics: Initial Findings on the Role of Listening Bystanders in Conversational Interactions

Jonathan Ehret[1] Valentin Dasbach[2] Jan-Nikjas Hartmann[2] Janina Fels[3]
Torsten W. Kuhlen[1] Andrea Bönsch[1,*]

[1] Visual Computing Institute, RWTH Aachen University, Germany
[2] Department of Computer Science, RWTH Aachen University, Germany
[3] Institute for Hearing Technology and Acoustics, RWTH Aachen University, Germany

(a) (b) (c)

Figure 1: (a) The living room with a male character used as a body avatar. Besides the user, the social group includes (b) two speakers (green, red) and three bystanders in P_{Listen}, and (c) one speaker (green) and four bystanders in P_{Act}. Agents are consistently colored across scenarios, with environment-based gaze aversion targets per agent marked accordingly.

Abstract

This work-in-progress paper investigates how virtual listening bystanders influence participants' gaze behavior and their perception of turn-taking during scripted conversations with embodied conversational agents (ECAs). 25 participants interacted with five ECAs – two speakers and three bystanders – across three conditions: no bystanders, bystanders exhibiting random gazing behavior, and social bystanders engaging in mutual gaze and backchanneling. Participants either observed the conversation or actively participated as speakers by reciting prompted sentences.

The results indicated that bystanders reduced the participants' attention to speakers, hindering their ability to anticipate turn changes and resulting in longer delays in shifting their gaze to the new speaker after an ECA yielded the turn. Random gazing bystanders were particularly noted for obscuring conversational flow. These findings underscore the challenges of designing effective and natural conversational environments, highlighting the need for careful consideration of ECA behaviors to enhance user engagement.

Index Terms: Virtual reality, eye tracking, social groups, turn taking, conversations, bystanders.

1 Introduction

In recent years, the development of embodied conversational agents (ECAs) [2] has gained significant attention in both research and practical applications. These agents are designed to interact with users naturally, mimicking human conversation through verbal and non-verbal cues. One critical aspect of conversational dynamics is turn-taking, which relies heavily on visual cues such as gaze direction, gestures, and body language (cf. [3]). Understanding how these elements influence user perception and engagement is essential for improving the design of ECAs.

*e-mail: boensch@vr.rwth-aachen.de

Our research objective is to investigate how virtual listening bystanders -— agents who participate in conversations without taking an active speaking role -— impact participants' gaze behavior and recognition of turn-taking cues during interactions, as understanding these dynamics is crucial for enhancing VR-based interactive systems where social interaction plays a pivotal role. While adding more ECAs may enhance realism, it is vital to understand potential side effects on user engagement and communication flow depending on whether they show context-sensitive gazing behavior. Previous research has explored various aspects of turn-taking and social presence among speaking agents (e.g., [15, 10, 3]), while Oertel et al. [11] specificially investigated multi-party listener behavior, focusing on implementing an attentive listening system that generates multi-modal listening behavior. Their findings indicate that appropriate modeling of listener behavior – including gaze patterns and feedback tokens – can positively affect perceptions of empathy, understanding, and rapport; conversely, inappropriate use may lead to negative consequences. However, there remains a gap in understanding how additional listening bystanders affect the participants' engagement with the conversation.

By investigating gaze dynamics in this context, we aim to uncover insights into how participants allocate their visual attention among speakers and bystanders, as well as how bystander behavior influences the understanding of the conversational flow. Preliminary findings from our study suggest that the presence of virtual bystanders may alter gaze patterns and potentially hinder users' recognition of turn-taking cues.

Through this work-in-progress paper, we seek to contribute to the ongoing discourse surrounding ECA design by providing initial insights into gaze behavior influenced by virtual listening bystanders. Our findings will demonstrate how adding bystanders to a conversational setting can further complicate or enhance user experience in VR-based interactive systems.

2 User Study and Methodology

To investigate the impact of listening bystanders in a conversational setting, we build on the validated framework established by Ehret et al. [3], which examined turn-taking cues within a limited social

J. Ehret, V. Dasbach, J.-N. Hartmann, J. Fels, T. W. Kuhlen and A. Bönsch, "Exploring Gaze Dynamics: Initial Findings on the Role of Listening Bystanders in Conversational Interactions," 2025 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), Saint Malo, France, 2025, pp. 748-752, doi: 10.1109/VRW66409.2025.00151.

(a) (b)

Figure 2: (a) Phase P_{Listen} with the male agent in the dark shirt speaking. (b) Phase P_{Act} with the female agent dressed in a purple sweater speaking, and the next sentences of the participant to be recited on the flip chart. For both, the three bystanders are in condition B_{Social}.

context involving two speaking ECAs. We enhance this work by incorporating three bystanders with distinct gazing behaviors. This enables a comprehensive examination of social dynamics and potential side effects of the additional ECAs. By maintaining the original two-phase structure -— P_{Listen}, where participants listen to a family story, and P_{Act}, where they narrate scripted stories -— we aim to elucidate how virtual bystander presence affects gaze behavior and interaction flow. Thereby, we hypothesize that, **H1**, the presence of social bystanders will improve participants' detection of turn changes compared to interactions with only speaking ECAs. Conversely, we except, that **H2**, bystanders exhibiting behaviors that deviate from typical social norms of visual engagement during conversations will hinder the detection of turn changes. This chapter details the methodology employed, including ECA and bystander behaviors, as well as data collection processes.

2.1 Virtual Setting

Following [3], we used the same living room environment[1] to provide a non-disturbing and private setting suitable for the family stories to be presented..

Participants were represented by a gender-matched body avatar, allowing them to feel physically present within the virtual space (Fig. 1(a)). To this end, we equipped each participant with two Valve Index Controllers and a Vive Pro Eye headset, utilizing the full-body inverse kinematics (IK) solver integrated into Unreal Engine to apply the tracked data to the body avatar. This setup ensures correctly synchronized upper-body movement of the avatar, which is crucial for our conversational setting. Although additional logic was implemented to simulate potential walking or side-stepping movements -— due to the absence of dedicated tracking points for the lower body -— we will not explore these aspects here further, as our participants remained stationary throughout the user study.

Accompanying the participant were five *MetaHuman*[2] agents, known for their realistic appearance and expressive capabilities. They were arranged in a circular formation, a common setup for stationary groups during conversations [1, 6, 9], which not only facilitates engagement but also provides participants equal visual access to all group members. Interpersonal distances adhered to standard norms, and internal testing ensured that standing in the circle felt comfortable.

Among these five virtual agents, two served as speakers, while the remaining three acted as listening bystanders. The **speaking ECAs** were programmed to exhibit naturalistic behaviors during conversations, including gestures, breathing, and gaze that aligned with their turn-taking cues. Their actions adhered to the social model defined by [3], ensuring that they periodically made eye contact with the current speaker and engaged in appropriate non-verbal communication throughout the interaction. In P_{Listen} the participant and the bystanders are considered to be addressees of the conversation, while in P_{Act} only one previously speaking ECA remains as a speaker, while the other becomes an addressee as well. Besides looking at the other agents or the participant, the speakers occasionally directed their gaze toward the environment, creating a more lifelike and believable interaction. To this end, we manually placed environment gaze targets as shown in Fig. 1(b) and Fig. 1(c). These were relocated compared to [3], ensuring clear averted gaze signaling and addressing a limitation identified in the initial work. However, for consistency, we maintained the remaining conversation dynamics and recorded facial and full-body animations.

The three **listening bystanders** were governed by a rule-based social model that guided their gaze behavior, determining how frequently they looked at other agents, the participants, or the environment during conversations, with the latter being manually defined environment gaze target as for the speaking ECAs, see Fig. 1(b) and Fig. 1(c). This approach created a dynamic setting where participants could observe variations in the attention of the bystanders between the speaking and listening agents. Based on the social model, the bystanders were categorized into three conditions:

In Condition B_{None}, only the two speakers were present, allowing for a focused pursuit of the conversation flow without any additional visual distractions. This condition serves as the baseline for our two hypotheses.

In Condition B_{Random}, the three listening bystanders exhibited random gazing behavior without actively acknowledging either the speakers or the participant. Their gaze targets – whether in the environment (Fig. 1(b) and Fig. 1(c), respectively) or at one of the group members – were selected to ensure variability, with no semantic meaning associated with looking at group members, while maintaining a weighted distribution similar to B_{Social}. This condition exemplifies unsocial behavior, characterized by the lack of active engagement and acknowledgment. While the presence of additional agents might enhance the realism of the environment, their unsocial behavior may have a negative impacting social interactions (cp. **H2**).

In Condition B_{Social}, the three listening bystanders adhered to a structured social model that guided their gaze, enabling them to actively acknowledge the speakers through context-sensitive mutual gaze and gaze following, thus providing additional valid cues for the conversation flow, easing its assessment (cp. **H1**). The distribution of their gazes among speakers, other listeners, and the environment was informed by prior research [14, 12]. Notably, during turn changes, the listening agents looked toward the next speaker with a 60% probability approximately 100 ms before the transition, signaling anticipation as an additional cue for turn-taking [8]. We chose minimal variances in this timing to enhance the believability of the interaction and ensure a more natural conversational flow, specifically $M = 0.1$ s and $SD = 0.03$, following a normal distribution. Additionally, the model guided the bystanders' behavior by enabling subtle smiles and backchanneling cues, such as nodding and vocalizations as proposed by [11] for an enhanced rapport and understanding.

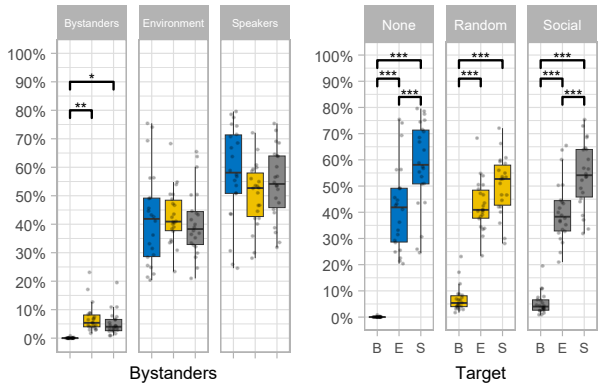
2.2 Study Procedure

The study was organized into two phases, each examining bystander behavior under the three varying conditions B_{None}, B_{Random}, B_{Social}. The study was set up and conducted using the *StudyFramework* [4]. Within each phase, each condition was conducted four times consecutively to increase exposure and thereby improve the reliability of gathered data. The overall order of the conditions was counterbalanced to minimize potential biases.

Listening Phase P_{Listen}: During this phase (Fig. 2(a)), participants listened to family stories narrated by two speaking ECAs while observing additional listening behaviors present in the three

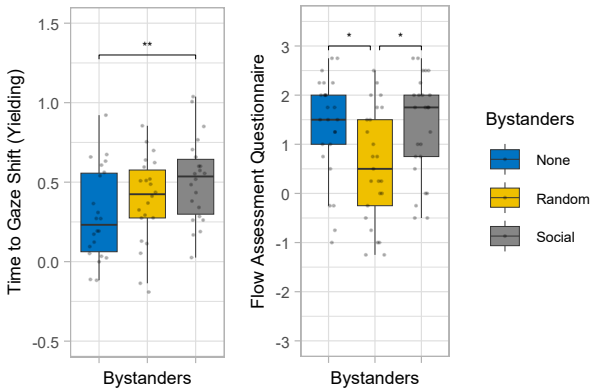
[1] Adapted from the free Unreal asset https://www.fab.com/listings/62e0fe0f-3fd7-4d40-993a-cae13e8199f4

[2] https://www.unrealengine.com/metahuman



(a) (b)

Figure 3: Boxplots of the percentage of time participants looked at the bystanders (B), the environment (E), and the speakers (S) ordered by the region of interest in (a) and the bystander conditions in (b). Significant pairwise differences are indicated by *** for $p < .001$, ** for $p < .01$, and * for $p < .05$, while all other differences are non-significant. The legend can be found in Fig. 4.



(a) (b)

Figure 4: (a) Time (in seconds) taken by participants to shift their gaze to the next speaker following a turn yield. (b) Mean answer scores for the conversation flow assessment questionnaire. Significant pairwise differences are indicated by ** for $p < .01$ and * for $p < .05$, while all other differences are non-significant.

aforementioned configurations (none, random, social). Participants were instructed to focus on the narrative content while being asked to look at the social group, avoiding them to close their eyes as a strategy to better remember the family stories. This is also true for the next phase.

Act Phase P_{Act}: In this phase (Fig. 2(b)), participants were instructed to take an active role as speakers by reciting sentences from family stories provided on a flip chart alongside the other speaker opposite them, which showed their next line in advance, requiring them to recognize when it was their turn to speak. The speaking ECA was of the opposite gender to the participant, mimicking the gender balance in speakers from P_{Listen}. The speaking ECA from phase P_{Listen}, who is no longer involved as a speaker, becomes an addressee. In terms of its behavior, it uses the behavior model of B_{Social}, which is comparable to the model of the speakers when listening. This approach also facilitates comparisons with the study by [3], as our B_{None} condition then mirrors the version they tested. To accommodate the introduction of the flip chart, we reordered the positions of the ECAs to ensure that the chart was clearly visible to participants. Consequently, the male and female speaker positions from P_{Listen} were flipped, leading to fewer occlusions while still maintaining a slight overlap for enhanced visual appeal. The same configurations for the presence and behavior of the bystanders were applied as in P_{Listen}. To avoid having same-gender individuals positioned next to each other, promoting a more diverse interaction dynamic, and reducing potential biases in participant responses, the listening bystanders were randomly shuffled around.

After each story presentation, participants answered comprehension questions related to what they had heard before verbally. The questions were displayed on the virtual TV screen in the living room, while the experimenter recorded which answers were correct and which were incorrect. After repeating each condition twice, they filled out questionnaires assessing their perceptions of conversation dynamics.

The family stories used are part of the established Heard Text Recall (HTR) paradigm [13], consisting of 34 German texts that provide information on three generations of family members, with nine questions per text. Voice recordings of the texts and the respective facial trackings are available in [7].

2.3 Data Collection

Although for the overall research objectives, we collected more data, we will focus only on a subset for this work-in-progress paper:

Gaze target distribution was analyzed across different regions of interest (ROIs) – *Listening Bystanders*, *Environment*, and *Speakers* – to assess where participants allocated their visual attention most frequently using a Vive Pro Eye headset for eye tracking. Therefore, we measured the total conversation time and the duration spent observing these ROIs to calculate gaze target percentages for all listening bystander conditions. Furthermore, **gaze switch times** in seconds were recorded to analyze the time it took participants to focus their gaze on a new speaker when an active speaking ECA yielded a turn. A time window was established to evaluate gaze switches, defined as occurring between 0.5 seconds before and 3 seconds after a change of turn, thus ignoring turn-changes where the speaker was not looked at during this time frame.

Participants also completed different questionnaires to provide qualitative insights into their prior experiences with virtual interactions and perceptions about turn-taking cues. However, in this paper, we will limit our discussion to our custom questionnaire on **conversation flow assessment**, comprising the four questions "It was easy to comprehend when I should speak.", "The behavior of the other persons was ambiguous.", "The behavior of the other persons confused me.", and "The other persons followed the conversation." All questions were rated on a 7-point Likert scale between -3 ("Do not agree") and 3 ("Agree"), while answers to the second and third item were reversed resulting in the final score in positive values representing a good comprehension of the conversation flow.

2.4 Participants

A total of 25 individuals participated in the study, comprising 13 males and 12 females. All participants were fluent in German, with ages ranging from 21 to 36 years (mean age: 26.16 years, SD: 3.50).

3 Results

This chapter presents our findings, focusing on gaze target distribution, gaze switching, and participants' subjective feedback gathered from the questionnaires. We tested for normality of the data using Shapiro-Wilk-tests and corrected for violated sphericity where applicable. For the gaze evaluations, we excluded three participants due to technical problems with the gaze tracking.

3.1 Gaze Target Distribution

Due to violations of normality in the gaze target distribution data, we employed the Aligned Rank Transform (ART) [5] method for non-parametric factorial analysis which allows analysis analogous to a two-way repeated-measures ANOVA while respecting the non-parametric nature of the data. We assessed the effects of the three listening bystander conditions on the participants' gaze allocation among the three ROIs. The analysis indicated significant main effects for ROIs ($F(2,168) = 253.46, p < .001$), as well as interaction effects between bystanders and ROIs ($F(4,168) = 5.66, p < .001$), showing that the presence of bystanders significantly influences where participants direct their attention during conversations (Fig. 3).

Tukey-corrected paired ART-C tests revealed several significant differences in the participants' gaze allocation when comparing the bystander conditions among the ROIs (Fig. 3(a)): Unsurprisingly, gaze allocation towards the *Listening Bystanders* was significantly higher when bystanders were present (B_Random and B_{Social}) compared to B_{None} (both $p's < .021$). However, there was no significant difference in gaze allocation between B_{Random} and B_{Social} ($p > .99$). No significant differences were observed in gaze allocation towards the *Environment* across the bystander conditions (all $p's > .99$) as well as towards the *Speakers* (all $p's > .60$).

Additionally, Tukey-corrected paired ART-C tests were conducted to assess participants' gaze allocation across the three regions of interest under the three varying bystander conditions (Fig. 3(b)): For all three listening bystander conditions, participants allocated significantly more gaze time ($p < .001$) to both the *Environment* and *Speakers* compared to the *Listening Bystanders*. Comparing *Environment* and *Speakers* across the bystander conditions revealed a significant difference in B_{None} ($p < .001$) and B_{Social} ($p < .001$), while no significant difference was found for B_{Random} ($p > .22$).

3.2 Gaze Switching

A repeated-measures ANOVA revealed a statistically significant difference in gaze switch times in P_{Listen} across the three different bystander conditions ($F(2,44) = 7.95, p < .001$). Pairwise t-tests indicated that this effect was due to a significant difference in switch times between B_{None} ($M = 0.29$ s, $SD = 0.28$) and B_{Social} ($M = 0.51$ s, $SD = 0.26$) ($p = .003$), while no significant difference ($p's > .08$) was found in the other two pairs with B_{Random} ($M = 0.44$ s, $SD = 0.27$). (Fig. 4(a))

We also looked at events, where the speaker "holds a turn", as these moments provide insight into how participants maintain attention on speakers during extended utterances. However, here a repeated-measures ANOVA did not reveal a significant difference ($F(2,44) = 1.64, p = .21$).

3.3 Conversation Flow Assessment

Cronbach's α indicated acceptable reliability at .68, allowing us to compute the mean ratings for the four custom assessment questions. A repeated-measures ART ANOVA showed a significant difference between the bystander conditions ($F(2,48) = 5.47, p = .007$). Tukey-corrected paired ART-C tests revealed significant differences in mean ratings between B_{Random} and B_{None} ($t(48) = 2.86), p = .017$) as well as between B_{Random} and B_{Social} ($t(48) = -2.87), p = .016$), with B_{Random} receiving the lowest scores, as shown in Fig. 4(b).

3.4 Participant Open Feedback

Furthermore, participants provided qualitative insights through written feedback at the end of the study. Verbal comments throughout the study were noted by the experimenter together with the related condition. All qualitative feedback was grouped manually, and key insights are reported anecdotally. Feedback indicated that while many found it generally easy to follow conversations and anticipate who would speak next, some expressed difficulty when listening bystanders were present. Three participants, for example, explicitly stated they recognized the next speaker based on listening agents' behavior, while six participants reported that they identified turn changes primarily through cues from the current speakers.

Overall responses suggested that while engagement levels varied based on agent configurations (particularly with random gazing), many participants felt more comfortable following conversations without bystanders as visual distractors.

4 Discussion

The findings of this study provide valuable insight into how virtual listening bystanders influence participants' gaze behavior and perception of turn-taking during interactions with ECAs.

The analysis revealed that the presence of social bystanders negatively impacted participants' ability to quickly switch their gaze to new speakers during conversations. Participants exhibited longer gaze switch times in conditions with social bystanders compared to no bystanders, suggesting that additional visual stimuli may have distracted users from identifying turn changes effectively. This aligns with previous research indicating that clarity in visual cues is essential for recognizing conversational dynamics ([15, 10, 11]). These results challenge our hypothesis **H1**, stating that social bystanders would enhance participants' detection of turn changes compared to only speaking ECAs. However, they are in favor of hypothesis **H2**, stating that unsocial behavior will worsen the conversation flow assessment. So, while social bystanders may improve the naturalness of interactions, they may also be distractors – such as the unsocial bystanders –, negatively impacting conversation dynamics, emphasizing the importance of further research on the resulting social dynamics.

The results also indicated a shift in attention away from speakers when bystanders were present, highlighting how additional agents can dilute the focus on primary conversational partners. Interestingly, while random bystanders tended to look more frequently at the environment, this behavior did not effectively redirect users' focus toward those areas. Instead, both random gazing and social bystanders primarily changed gaze distribution within the social group. While it is not surprising that the presence of other agents can draw attention – especially when they look at the user – this finding remains relevant as it underscores potential challenges in maintaining engagement within multi-agent interactions. The decreased gaze toward speakers in both random and social bystander conditions points to potential challenges in maintaining engagement within multi-agent interactions.

These findings suggest that while incorporating listening agents may aim to enhance social presence, it can inadvertently lead to confusion regarding or ignorance of turn-taking cues – even in such a simple setting with a limited amount of speaking agents. It is crucial to critically consider that if no listeners are present, users cannot engage visually with them; thus, their absence does not detract from interaction quality. However, when listeners are included, their behaviors must be meaningful; otherwise, it may be better to exclude them entirely. As such, careful consideration must be given to agent behaviors and their implications for user experience. Furthermore, feedback indicated that the unsocial bystanders particularly obscured the conversation flow due to their random gazing behavior, which may have detracted from the overall interaction quality.

There are a few **shortcomings** in our study. First, the sample size is relatively small, which may affect the generalizability of our findings. Second, we employed a challenging standardized psychology task designed specifically to focus participants on the ongoing conversation rather than the ECAs' behavior; while this approach aimed to elucidate subconscious social dynamics, it might limit

broader applicability. Finally, exploring alternative social models could provide further insights into optimizing ECA configurations for enhancing communication.

Still, our work-in-progress research highlights an important area for future research: understanding how various configurations of ECAs can be optimized to support clearer communication and enhance user interaction without overwhelming them with extraneous visual information or introducing distracting behaviors.

5 Conclusion

In summary, our study underscores the complexities involved in designing effective virtual conversational environments. While virtual listening bystanders can enrich interactions and enhance realism, their impact on visual user attention and perception must be carefully managed to ensure meaningful engagement with primary speakers. Designers should consider potential distractions from additional agents in multi-agent social encounters. Future research should explore their effects on auditory spatial attention and investigate how different configurations of listening agents can optimize communication flow without overwhelming users. Key areas for further investigation include examining socially appropriate behaviors of listening agents and their influence on turn-taking recognition, conversation recall, and perceived social presence.

Acknowledgments

This research was funded by the German Research Foundation (DFG) within the project "Listening to, and remembering conversations between two talkers: Cognitive research using embodied conversational agents in audiovisual virtual environments", which is part of the DFG Priority Program "AUDICTIVE" (SPP 2236).

References

[1] A. Bönsch, A. R. Bluhm, J. Ehret, and T. W. Kuhlen. Inferring a User's Intent on Joining or Passing by Social Groups. In *Proc. of the 20th ACM International Conference on Intelligent Virtual Agents*, 2020. doi: 10.1145/3383652.3423862 2

[2] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill. *Embodied Conversational Agents*. MIT Press, 2000. 1

[3] J. Ehret, A. Bönsch, P. Nossol, C. A. Ermert, C. Mohanathasan, S. J. Schlittmeier, J. Fels, and T. W. Kuhlen. Who's next? Integrating Non-Verbal Turn-Taking Cues for Embodied Conversational Agents. In *Proc. of the 23rd ACM International Conference on Intelligent Virtual Agents*, 2023. doi: 10.1145/3570945.3607312 1, 2, 3

[4] J. Ehret, A. Bönsch, J. Fels, S. J. Schlittmeier, and T. W. Kuhlen. Studyframework: Comfortably setting up and conducting factorial-design studies using the unreal engine. In *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW): Workshop "Open Access Tools and Libraries for Virtual Reality"*, 2024. doi: 10.1109/VRW62533.2024.00087 2

[5] L. A. Elkin, M. Kay, J. J. Higgins, and J. O. Wobbrock. An aligned rank transform procedure for multifactor contrast tests. In *Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology*, pp. 754–768, 2021. doi: 10.1145/3472749.3474784 4

[6] C. Ennis and C. O'Sullivan. Perceptually Plausible Formations for Virtual Conversers. *Computer Animation and Virtual Worlds*, 2012. doi: 10.1002/cav.1453 2

[7] C. A. Ermert, C. Mohanathasan, J. Ehret, S. J. Schlittmeier, T. W. Kuhlen, and J. Fels. AuViST - An Audio-Visual Speech and Text Database for the Heard-Text-Recall Paradigm, 2022. doi: 10.18154/RWTH-2023-05543 3

[8] J. Holler and K. H. Kendrick. Unaddressed Participants' Gaze in Multi-Person Interaction: Optimizing Recipiency. *Frontiers in Psychology*, 2015. doi: 10.3389/fpsyg.2015.00098 2

[9] A. Kendon. *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Cambridge University Press, 1990. doi: 10.2307/2075490 2

[10] B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, and N. Hagita. Footing in Human-Robot Conversations: How Robots Might Shape Participant Roles Using Gaze Cues. In *Proc. of the 4th ACM/IEEE International Conference on Human-Robot Interaction*, 2009. doi: 10.1145/1514095.1514109 1, 4

[11] C. Oertel, P. Jonell, D. Kontogiorgos, K. F. Mora, J.-M. Odobez, and J. Gustafson. Towards an Engagement-Aware Attentive Artificial Listener for Multi-Party Interactions. *Frontiers in Robotics and AI*, 2021. doi: 10.3389/frobt.2021.555913 1, 2, 4

[12] R. Rienks, R. Poppe, and D. Heylen. Differences in Head Orientation Behavior for Speakers and Listeners: An Experiment in a Virtual Environment. *ACM Transactions on Applied Perception*, 2010. doi: 10.1145/1658349.1658351 2

[13] S. J. Schlittmeier, C. Mohanathasan, I. S. Schiller, and A. Liebl. Measuring Text Comprehension and Memory: A Comprehensive Database for Heard Text Recall (HTR) and Read Text Recall (RTR) Paradigms, with Optional Note-Taking and Graphical Displays, 2023. doi: 10.18154/RWTH-2023-05285 3

[14] R. Vertegaal, R. Slagter, G. van der Veer, and A. Nijholt. Eye Gaze Patterns in Conversations: There is More to Conversational Agents than Meets the Eyes. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, 2001. doi: 10.1145/365024.365119 2

[15] Z. Wang, J. Lee, and S. Marsella. Multi-Party, Multi-Role Comprehensive Listening Behavior. *Autonomous Agents and Multi-Agent Systems*, 2013. doi: 10.1007/s10458-012-9215-8 1, 4