Towards creating a conversational memory for long-term meeting support predicting memorable moments in multi-party conversations through eye-gaze

Tsfasman, Maria; Fenech, Kristian; Tarvirdians, Morita; Lorincz, Andras; Jonker, Catholijn; Oertel, Catharine

DOI
10.1145/3536221.3556613

Publication date
2022

Document Version
Final published version

Published in
ICMI 2022 - Proceedings of the 2022 International Conference on Multimodal Interaction

Towards creating a conversational memory for long-term meeting support: predicting memorable moments in multi-party conversations through eye-gaze

Maria Tsfasman
Delft University of Technology
Delft, The Netherlands
m.tsfasman@tudelft.nl

Kristian Fenech
Eötvös Loránd University
Budapest, Hungary
fenech@inf.elte.hu

Morita Tarvirdians
Delft University of Technology
Delft, The Netherlands
m.tarvirdians@tudelft.nl

Andras Lorincz
Eötvös Loránd University
Budapest, Hungary
lorincz@inf.elte.hu

Catholijn M. Jonker
Delft University of Technology
Delft
Leiden University
Leiden, The Netherlands
c.m.jonker@tudelft.nl

Catharine Oertel
Delft University of Technology
Delft, The Netherlands
c.r.m.m.oertel@tudelft.nl

ABSTRACT

When working in a group, it is essential to understand each other's viewpoints to increase group cohesion and meeting productivity. This can be challenging in teams: participants might be left misunderstood and the discussion could be going around in circles. To tackle this problem, previous research on group interactions has addressed topics such as dominance detection, group engagement, and group creativity. Conversational memory, however, remains a widely unexplored area in the field of multimodal analysis of group interaction. The ability to track what each participant or a group as a whole find memorable from each meeting would allow a system or agent to continuously optimise its strategy to help a team meet its goals. In the present paper, we therefore investigate what participants take away from each meeting and how it is reflected in group dynamics.As a first step toward such a system, we recorded a multimodal longitudinal meeting corpus (MEMO), which comprises a first-party annotation of what participants remember from a discussion and why they remember it. We investigated whether participants of group interactions encode what they remember non-verbally and whether we can use such non-verbal multimodal features to predict what groups are likely to remember automatically. We devise a coding scheme to cluster participants' memorisation reasons into higher-level constructs. We find that low-level multimodal cues, such as gaze and speaker activity, can predict conversational memorability. We also find that non-verbal signals can indicate when a memorable moment starts and ends. We could predict four levels of conversational memorability with an average accuracy of 44 %. We also showed that reasons related to participants' personal feelings and experiences are the most frequently mentioned grounds for remembering meeting segments.

ICMI '22, November 7–11, 2022, Bengaluru, India

ACM ISBN 978-1-4503-9390-4/22/11.
https://doi.org/10.1145/3536221.3556613

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in collaborative and social computing.*

KEYWORDS

conversational memory, multi-modal corpora, social signals, multi-party interaction

ACM Reference Format:
Maria Tsfasman, Kristian Fenech, Morita Tarvirdians, Andras Lorincz, Catholijn M. Jonker, and Catharine Oertel. 2022. Towards creating a conversational memory for long-term meeting support: predicting memorable moments in multi-party conversations through eye-gaze. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3536221.3556613

1 INTRODUCTION

Many things can go wrong when working with a team. If some people are more dominant than others, they end up the only ones talking and other participants can feel left out. Sometimes people think about their own contribution more than listening to others. This can be illustrated by the fact that humans remember more of the information they said than the information they listened to [10]. This is less true for people with better interpersonal skills [15], but not everyone has those. For a productive meeting, participants need to listen to each other and build on each other's ideas [9]. Here, conversational agents can be of help. Agents giving feedback on (non-)verbal context of the meeting already show great promise in improving meeting productivity [11, 26]. To further tackle group meeting support, such agents need to be able to automatically understand which moments are most important, build upon them throughout the interaction, and identify the moments of needed intervention. To do this, an agent needs to have a conversational memory model. Ideally, an agent possessing such a memory would use verbal or non-verbal signals received from participants of the conversation to identify which moments are most important to memorise and refer back to. While studies that focus on the textual modality of conversational memory show promising results (e.g.

[3, 8, 27, 31]), participants alternate in speaking turn, and therefore textual data is sparse. Non-verbal signals, such as gaze, on the other hand, can be continuously used to track participants' involvement in conversation with very minor interruptions [18]. In the present paper, we pioneer a data-driven approach to conversational memory modelling. We analyse a multi-modal corpus of group conversations annotated with first-party reports of most memorable moments and provide evidence that group gaze behaviour can serve as a predictor of memorable moments in multi-party conversations.

In the present paper, we define conversational memory, following the accepted classification of memory types [2], as explicit (i.e. retrievable in a verbal free-recall task), long-term (retained for longer than minutes or seconds after the interaction), episodic (moments of life rather than semantic knowledge), and specific to moments of conversation.

2 BACKGROUND

Capturing what people find important and remember is a complex task for several reasons. Humans do not always reveal their affective responses and often keep their inner thought processes to themselves. Therefore, for an observer, it is difficult to understand which moments are memorable for other participants in the interaction. Conversations are particularly challenging to study in the context of memory because of a lack of control over the topics brought up and missing contextual knowledge of the participant's autobiographic experiences.

Although there is little research on conversational memory yet, it appears that there are multiple factors that can affect what is remembered. For example, the relational history of interlocutors with one another, as well as their participatory role in the conversation, might influence their conversational memory. Specifically, participants tend to remember conversational information differently from passive observers: participants' memories are more detailed, episode-specific, and centred around non-verbal communication [1, 29]. People also remember more information from conversations with friends than strangers [25] and when they have more common ground [14]. Participants' individual traits, such as interpersonal skills can also influence what they remember from conversations [15]. People with better interpersonal skills memorise more information about their conversational partner than the ones with poorer communication skills [15]. All these factors (participants' role, personal traits, and relationship between them) are more or less static throughout a conversation, but if we are given all these variables, it would still be difficult to guess which moments are most likely to be remembered. There needs to be a continuous measure that can track how salient each moment is throughout the interaction.

One way salience can be tracked in a conversation is by how engaged or involved participants are at a particular moment [4, 17, 19, 20, 24]. In meeting summarisation, participants' involvement has been connected to moment importance and has been shown to improve the quality of the resulting meeting points [32]. Involvement is usually annotated by third-party observers and can be predicted via speech, prosody [12, 32] or group gaze behaviour [18]. Although the direct connection between group involvement and memory has not been investigated, individual attention has been known to be a major factor in memorisation on a physiological level in the human brain [16, 28, 30]. In a conversational context, a study found that participants are more likely to recall the part of the conversation in which they were most active [10]. Although not on a group level, this indicates once again how non-verbal features connected to involvement could potentially be indicative of memorable moments.

Therefore, we hypothesise that it is possible to predict which moments will be more memorable for the group using features previously connected to group involvement. Our research questions are formulated as:

RQ1: Do humans non-verbally encode moments they are more likely to remember? Do features that distinguish between different levels of group engagement [18] also discriminate for different degrees of group conversational memory?

RQ2: Which reasons do participants give for remembering a moment?

RQ3: How can we use these features to predict the memorability of conversational moments?

Based on the literature discussed above we formulate the following hypothesis:

Hypothesis: The signals previously connected to involvement and engagement are predictive of conversational memorability.

Contribution. The present study pioneers a largely unexplored topic of conversational memory modelling. This study provides the first step on the path of characterising the multimodal features that are predictive of conversational memory.

3 METHODS

3.1 Corpus

3.1.1 General Description. In this study, we are using a multimodal group discussion corpus that is part of a larger data collection called the "MEMO corpus". The MEMO corpus consists of video-call discussions in groups of 3-6 participants over three consecutive sessions distanced 3-4 days apart. Throughout 45 minutes long sessions, participants discussed COVID-19 and their experiences in the pandemic. To facilitate an active discussion, each group was paired with a moderator. We recruited moderators who had experience in moderating meetings, facilitating creative sessions, and conducting interviews. Moderators were confederates instructed to keep the conversation going and encourage participants to express their opinions and emotions as much as possible.

Overall, 53 participants (28 F, 25 M; 18-76 y. o.) and 4 moderators (3 M, 1 F; 24-45 y.o.) took part in the data collection. All participants were fluent English speakers and resided in the UK. Each participant and moderator signed a consent form for video, audio and survey data collection and filled up a pre-screening survey before the experiment. When recruiting participants, we tried to maximise the diversity of in-group opinions. For that reason, we tried to have representatives from various COVID19-affected demographics in every group: parents with young children, older adults (50+), students, (ex)-business owners. Participants were divided into 15 groups assigned with the principles of maximal in-group diversity and participants' scheduling preferences (3-5 participants and 1 moderator per group). To control for previous relationships,

participants have never met each other or the group moderator before the first discussion.

The resulting data comprises 34 hours of group discussions (45 sessions). On average, the sessions were 45 minutes long (with a standard deviation of +- 6.6 minutes). Before and after each session, participants and moderators filled in a series of surveys. The surveys included a wide range of perceptual measures, for conciseness, we only mention the ones used in the current analysis which are described in the following subsections.

3.1.2 Memory Labels Calculations. All the feature annotations were calculated using 5-second sliding windows. For each time window, there was one value for each group feature. This applies to memory, gaze, and speaker annotations.

First-party memory annotation In order to capture memorable moments from the interaction and collect ground-truth labels of when they occurred, the memory annotation consisted of two stages illustrated by Figure 1:

(1) **Free recall reports in the post-session questionnaire:** straight after every session participants had to complete a survey which started with free-recall memory questions. The task description was following: "Recall and describe moments of the most recent discussion session in as much detail as you can remember. Any details are great - for example, about the content, other participants, the moderator, you, your feelings, the reaction of others, your words, others' words, timing, or anything that happened throughout the discussion. Recall at least 3 moments. If you remember more, the fields will show up as you go until you leave one of them empty."
(2) **First-party timing annotation:** After the post-session survey was completed, participants were asked to annotate the memorable moments they described in their free recall description in the session video recording. Each moment had to be indicated by a start and end time. Participants had an option of leaving timing blank if the moment could not be attached to a particular timestamp (for example, a general impression of the entire session). Additionally, we asked participants to indicate the reasons for remembering the moment. Timing annotations served as ground-truth labels for all further analysis.

We considered individual memorable moments as consecutive in case they overlapped in time unless one of the moments lasted longer than half of the discussion session. All annotations encompassing more than half a session were discarded as they did not apply to a particular moment but rather "an overall feeling" of discussion. All individual memorable moment annotations of the group were brought together for analysis. Overall, there were 633 memorable moments (mean= 143.5, standard deviation = 183.6 seconds).

Figure 2 illustrates the process of creating the memory labels step-by-step. We employed a 5-second sliding window approach and considered the memorable moment as a binary variable. For **individual memory annotations,** each moment was represented as an array of the time slice t per participant i. It was 1 if at least half of the time slice t was included in the moments remembered by the participant i, and it was 0 otherwise. After that, we computed the **group-level memory index** as the proportion of participants that considered each time slice t memorable. We then divided the group memory indices into four **memorability level labels**: *zero* - if nobody remembered a slice; *low* - if > 0 and < 30 % remembered a slice; *middle* - if 30-70% considered a slice memorable; *high* - >70% reported a slice as remembered (see the last line in Figure 2). We used these memory level labels in the classification and for other analyses further on.

3.1.3 Speaker Activity Identification. From the recorded audio of the discussions, we extracted active speaker information using Kaldi Speech Recognition Toolkit [23]. After conducting speaker diarization, we extracted an active speaker array per time window t containing binary values of each participant i speaking at that time interval ($s_i(t)$). The value was 1 if the participant i spoke for at least half of the slice t and 0 otherwise. We then calculated the active speaker index per time window t using the following equation:

$$S(t) = \frac{\sum_{i=1}^{N} max(s_i(t), s_i(t-1), s_i(t-2))}{N} \quad (1)$$

Simply put, we calculated the number of individual participants (i) that were speaking in each time slice (t) or in two time slices proceeding to it ($t-1$ and $t-2$) and divided that sum by the overall number of participants N in the session.

3.1.4 Gaze Annotation. **Eye gaze target extraction.** Point of gaze was estimated with GazeSense software [6]. For each participant a grid matching the gallery layout was defined based on their provided screen capture. At the start of each session, a calibration stage was performed: participants were required to fix their gaze to the screen segment containing the current target participant. We estimate the target calibration point of the i-th segment of the grid $\mathbf{p}_{grid,i}$ to be the coordinates in the centre of the segment. Point of gaze estimates $\mathbf{p}_{gaze}$ were then obtained for all remaining frames beyond the final calibration frame. The final gaze target T_{gaze} for each frame was determined as

$$T_{gaze} = \begin{cases} \arg\min_i \|\mathbf{p}_{gaze} - \mathbf{p}_{grid,i}\|, & \text{if } \mathbf{p}_{gaze} \text{ detected} \\ -1, & \text{otherwise} \end{cases} \quad (2)$$

Because of recording imperfections and problems with some screenshots participants uploaded, the resulting gaze data had 40 participants (14 groups) and 16248 individual time windows (23 hours).

Group gaze features. [18] has shown a connection between group eye gaze behaviour and participants' conversational involvement. Specifically, there was a series of group-level eye-gaze features that have been shown to correlate with perceived involvement: presence, maxGaze, entropy, and symmetry. In this paper, we use the first three of these, since the gaps in data made the symmetry feature unreliable. All the features are calculated from the gaze matrix g with $N \times K$ dimensions: N being the number of participants with valid gaze data and K - the number of targets (number of participants and an additional label for when they look away from other participants or the screen).

Individual gaze matrix g_{ij} consisted of binary measures of gaze for each time slice t. It was 1 if participant i looked at participant j for at least half of the time window t, it was 0 otherwise. Unlike [18], a participant can gaze at themself on the screen, so there are

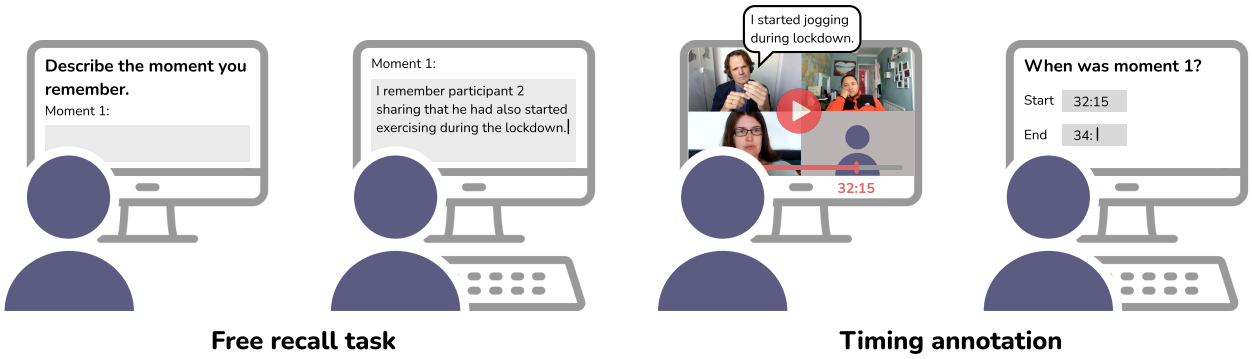


Figure 1: First-party memory annotation straight after the discussion session. Free-recall reports on the right and timing annotation on the left. The moment mentioned on the screen is an example from the data: "I remember participant 2 sharing he had also started exercising during lockdown."

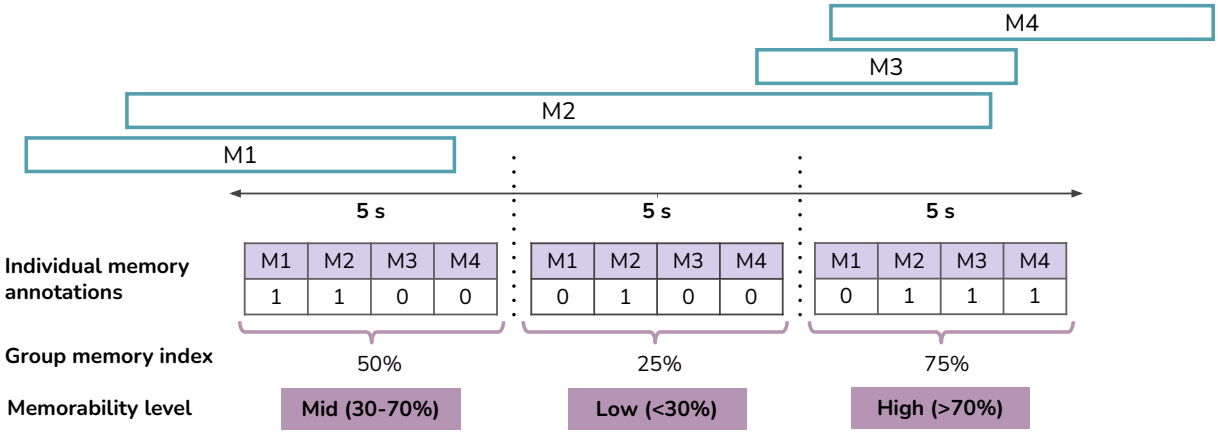


Figure 2: Memorability level annotation. The blue frames on the top illustrate moments remembered but 4 different participants. There are 3 consequent time windows on the x axis (5 seconds each). The memorability is considered True (=1) if the moment lasts for half or more of the specific window. Therefore, M3 is 0 for the second window.

no limitations to the value of g_{ii} in this regard. Nevertheless, since each participant could only gaze at one target at a time, following equation applies:

$$\sum_{i=1}^{N}\sum_{j=1}^{K} g_{ij}(t) = N, \forall t \tag{3}$$

The speaker-directed gaze feature $f_s(t)$ was calculated to see how many participants are looking at the active speaker at any time t. It was based on matrix $s_j(t)$, which also consisted of binary measures - it was 1 id j was an active speaker at that time slice t and 0 otherwise. For each participant i and target participant j we then computed a speaker-directed gaze value S_{ij} - it was 1 if participant i was gazing towards j ($g_{ij}(t) = 1$) and j was an active speaker ($s_j(t) = 1$) at that time slice t:

$$S_{ij}(t) = \begin{cases} 1, & \text{if } s_j(t) = 1 \ \ \& \ \ g_{ij}(t) = 1 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

To compute the final speaker-directed gaze feature f_s we then computed a fraction of participants looking at the active speaker for each time slice t:

$$f_s(t) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} S_{ij}(t)}{N} \tag{5}$$

The gaze presence feature $f_p(t)$ from [18] is the proportion of participants looking at other participants as opposed to looking away:

$$f_p(t) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} g_{ij}(t)}{N} \tag{6}$$

MaxGaze feature f_m computes the maximal number of participants looking at the same target at a particular time window t:

$$f_m(t) = \frac{max_{j\in[1,K]}\sum_{i=1}^{N} g_{ij}(t)}{N} \tag{7}$$

Entropy measure indicates the probability of each target being looked at by all others at each particular time:

$$P(\text{target} = j|t) = \frac{\sum_{i=1}^{N} g_{ij}(t)}{N} \quad (8)$$

To compute the final entropy measure $f_e(t)$ the probability is then normalised as follows:

$$f_e(t) = \frac{\sum_{j=1}^{K} P(\text{target} = j|t) log(P(\text{target} = j|t))}{log(K)} \quad (9)$$

Therefore, it is the lowest ($f_e(t) = 0$) when all participants are looking at the same target. It is the highest ($f_e(t) = 1$) if all participants are looking at different targets.

3.1.5 Memory reason annotation. After each memory timing annotation question, participants were also asked to self-report the reason for which they think they remembered each particular moment. To divide the reported reasons in separate types by their content, the dataset was manually annotated by two third-party annotators in multiple layers based on the given reason. Figure 3 shows the multi-layer annotation scheme. To assess the inter-rater reliability, we randomly selected 145 samples, and the annotators were not given information about the other annotator's labels. The inter-annotator agreement was measured using Fleiss' kappa statistics, which was found to be 0.60. According to [13], it is considered a moderate agreement.

3.2 Classification

3.2.1 Model architectures. For classification of memorability levels, we trained various supervised machine learning models: logistic regression, support vector machine (with RBF kernel), random forest classifier, and a multi-layer perceptron (MLP or "neural network" further).

The neural network consisted of 1 input, 3 hidden (256,128,32 neurons) and 1 output layers with 2 drop-out layers (0.2). The network was L2-regularised to reduce the effective size of the model. The model used cross-entropy loss and Adam optimiser, and ReLU activation function. Every iteration of the model was trained over 300 epochs with early stopping, to ensure that the model trains properly and does not overfit. The batch size was 32 and the learning rate was 0.005. The models were then compared using test and train balanced accuracy scores.

All scripts were written in Python using scikit-learn 0.24.1 [22] and pytorch toolkits 1.11.0 [21].

3.2.2 Features. For the input features, we used all the speaker and gaze features mentioned above. In addition to the features calculated for a given time window, we also added features for the previous and subsequent two time steps. We included the feature's mean, max and min over the included time steps. The resulting input vector consisted of 40 continuous features for each instance. The output label was one of four classes of memorability levels: zero, low, middle and high.

3.2.3 Training samples. The train and test sets were divided 80% and 20% respectively for all the models, except for MLP. For MLP we also had a validation set: 80% train, 10% test, and 10% validation. All the models were trained in 20 iterations with different random samples.

Since the class distribution was severely unbalanced (zero: 8501, low: 2355, middle: 4865, high: 527 instances), we took an under-sampling approach. The training set was under-sampled after being separated from the test and validation sets to have an equal class representation. In order to take into account as much data as possible, we randomly sampled the data this way 20 times, trained the model on each of these samples and approximated all the results over these iterations.

3.2.4 Feature selection and ablation study. In order to understand which features were most important for the model's predictions, we computed the **permutation importance** for the best-performing model. We also conducted a **feature ablation study** for the neural network. This included removing the feature sets connected to each of the main features (entropy, presence, maxGaze speaker-directed gaze, and active speaker index) one at a time and training the models on the remaining features. For example, when removing the entropy feature, we would also remove all other features connected to it: entropy plus-minus 2 time-stamps, entropy min, max, and mean. Therefore, the resulting model contained 28 input features at a time. We trained these models using the same procedure as the main model and using the same architecture and hyperparameters. The results were approximated over the same 20 samples as other models. We then used balanced accuracy results to compare the performance of each ablated model.

4 RESULTS

4.1 Memory level Analysis

This section explores the relationship between the four levels of conversational memorability with the gaze features of presence, entropy, max-gaze, and speaker activity. Figure 4 illustrates their means and confidence intervals.

The presence feature differed significantly across the four levels of memorability. According to the Kruskal-Wallis H test, there is a significant difference between the different memorability levels ($\chi^2(3) = 750.96$, $p < 0.001$). Dunn's Multiple Comparison post hoc test revealed that the presence is significantly different between all the memorability levels ($p<0.001$). The higher the level of memorability, the fewer people looked at other people.

Gaze entropy (blue in Figure 4) follows a similar trend as presence. A Kruskal-Wallis H-test showed a significant difference in entropy between memorability levels ($\chi^2(3) = 420.65$, $p < 0.0001$). Dunn's post hoc test revealed significant differences between all memorability levels ($p<0.001$). The higher the memorability level, the lower the entropy. This means participants were more likely to look at the same target in moments of high memorability.

A Kruskal-Wallis H-test revealed that maxGaze (green in Figure 4) is significantly different between the different memorability levels ($p<0.001$). A Dunn's post hoc test revealed that all differences between levels are significant except for moments of mid-level memorability to highly memorable ($p=0.062$) and zero to low ($p=0.004$).

A Kruskal-Wallis H-test revealed that speaker-directed gaze proportion (pink in Figure 4) is significantly different between the different memorability levels ($p<0.001$). A Dunn's post hoc test revealed that all different levels except for moments of mid-level memorability to highly memorable ($p=0.5$) and zero to low ($p=0.038$).

Labels	Sub-labels	Description
fact_about_others	view	agreement or disagreement with the view of another participant
	social_facts	speaker describes something about his/her activities, family or friends
	unexpected_info	speaker describes something which is unexpected from the annotator's perspective
fact_about_world	entities	view about entities such as lockdown, vaccination, etc.
	people	view about people
self_perception	annotator_feelings	feelings of the annotator such as happy, sad, embarrassed, etc.
	annotator_stories	stories or views of the annotator
shared_experience	shared_story	annotator has a similar experience or feeling with the speaker
meta_behaviour_of_other	emotional_moment	emotional moments of other participants
	behaviour	behaviour of the participants during the session such as laughter, anger, etc.
cognitive	cognitive_empathy	annotator sees the situation from the perspective of the speaker in a logical way
time_label	first	annotator remembers the first moments of the session
	last	annotator remembers the last moments of the session

Figure 3: Multi-layer annotation scheme for memorability reasons with corresponding descriptions. "Annotator" is a participant for whom the segment was memorable. "Speaker" is the main speaker of the segment.

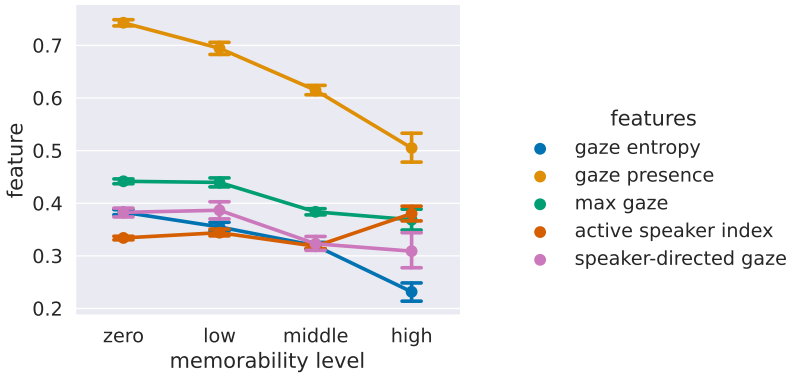


Figure 4: The differences between gaze and speaker features in relation to group memorability levels. On the y-axis: points are means of the feature for specific memorability levels and 95% confidence intervals as bars. On the x-axis: "zero" is for time slices that no one in the group recalled after the discussion; "low" are moments remembered by less than 30% of participants in the group; "middle" applies to slices remembered by 30-70% of participants; "high" - moments that 70 % or more of participants recalled)

Regarding the active speaker index (orange-red in Figure 4), the proportion of active speakers is significantly higher in highly memorable moments (Kruskal-Wallis $H(3) = 124.89$ and $p < 0.001$) with a mean active speaker index score of 0.37 for zero memorability, 0.38 for low memorability, 0.35 for mid-level memorability and 0.42 for high memorability. Dunn's post hoc test postulates significant differences ($p<0.001$) for the active speaker index for all pairs of levels, except for low vs zero memorability ($p = 0.011$). The active speaker index is significantly higher in high memorability moments than in middle, low and zero memorability ones. This means that there are more active speakers in highly memorable moments than in moments of lower memorability.

4.2 Memory-level Analysis across time

We also investigated whether any contextual cues might signal a memorable moment coming up or some changes that occur directly after the moment. For that, we compared how features changed in the two time windows before the memorable moment("BM" in Figure 5), within memorable moments ("M"), two time windows after each memorable moment ("AM"), and all other windows outside the mentioned groups ("NM"). In this case, a memorable moment is a moment remembered by at least one participant. Therefore, the comparison between the windows that fall into the categories "NM" and "M" was somewhat similar to the results described in Section 4.1. However, the difference between "NM"/"M" vs "BM"/"AM" is of

greater interest, since it sheds some light on whether there might be a cue that indicates the start or the end of a memorable moment.

There were no significant differences in group gaze entropy for different timing as indicated by the post hoc Dunn's test (p>0.001). For maxGaze and speaker-directed gaze, there was a significant difference between during vs. before, during vs. after, outside vs. within memorable moments (p<0.001 in all three pairs judging by Dunn's test) but there were no significant differences between outside vs. before (p=0.2 for maxGaze, p=0.9 for speaker-directed gaze) and outside vs. after (p=0.1 for maxGaze, p=0.8 for speaker-directed gaze). This can mean that while the lower max or speaker-directed gaze features do indicate memorable moments, there are no distinct predictive cues of a beginning or an end of the memorable moment within these features.

For the group presence measure, there was a gradual decrease in it from outside to right before the memorable moment and an increase from the end of the memorable moment to further outside the memorable segments. Although we can see this trend in Figure 5, the differences were significant only in the following pairs: during vs after/before/outside, outside vs after (p< 0.001, Dunn's post hoc). Differences between outside vs. before are insignificant (p=0.3, Dunn's post hoc).

The most promising candidate for being a cue in signalling a memorable moment was the active speaker index (fourth subplot in Figure 5). In the time window directly preceding a memorable moment window, the proportion of active speakers significantly increases (p>0.001, post hoc Dunn's test). Although there was also a slight increase in the subsequent time window, this increase was not significant (p=0.006, Dunn's test). Interestingly, the proportion of speakers within the memorable moment did not differ from moments further away ("M" vs "NM" p=0.5). This finding might serve as an additional indication that, in this case, what matters is how many participants are actively involved in the discussion directly before the moment becomes particularly memorable.

4.3 Memory reason analysis

In addition to the free recall of memorable moments we also asked participants to provide the reason why they remembered the moment. We then annotated the reasons according to an annotation scheme we devised and are describing in Table 3 with the inter-rater reliability metrics of 0.60 Fleiss' kappa.

The distribution of the memory-reason analysis is shown in Figure 6a. The most common reason was self-perception (250 of 633 memorable moments). The next frequent reason-label captured facts about other participants in the group (186). Other labels were considerably less frequent: shared experience (52), facts about the world (46), meta-behaviour of other participants in the group (44), time label (31), and cognitive empathy (24).

The sub-level distribution is shown in Figure 6b. Self-perception labels included more sub-labels related to the participant's feelings (199), than life experiences (51 reasons labelled "stories" in Figure 6b). Fact-about-other label had the majority of moments with the "view" sub-label (110 out of 186). This means that the reason for remembering the moment was related to the views of other participants in the group (for example, agreeing or disagreeing with their point of view). The second most frequent sub-label in fact-about-other reasons was unexpected information (52 out of 186), and the least frequent was social facts (52 out of 186).

4.4 Classification

We used four different classification methods: logistic regression (LogisticRegression in Figure 7), support vector machine with RBF kernel (NuSVC), random forest (RandomForest) and a multi-layer perceptron neural network (NeuralNetwork). Figure 7 shows balanced accuracy scores aggregated over models trained on 20 random samples (see 95% confidence intervals as error bars over these iterations). All the models performed better than chance (see a random classifier for comparison - "DummyClassifier" in Figure 7) with mean balanced test accuracy of 32% for logistic regression (LogisticRegression in Figure 7), 33% for support vector machine with RBF kernel (NuSVC), 43% for random forest classifier (RandomForest) and 44% for a multi-layer perceptron neural network (NeuralNetwork). We should note that the low test accuracy of the DummyClassifier is connected to it learning the distribution between classes on the under-sampled training test and when applied to the original class distribution in the test set performs considerably lower than chance. It should be noted that when dividing groups used for training and testing, the prediction accuracy considerably decreases (LogisticRegression - train 0.36, test 0.27, NuSVC - train 0.39, test 0.27; RandomForest train 0.43, test 0.26; NeuralNetwork - train 0.44, test 0.31).

Since the random forest and neural network models performed the best, judging by train and test accuracy, we analysed feature importance for these models to see if there are features that are more predictive of memorability level than others. For that, we computed the permutation importance of input features for the random forest model and conducted a feature ablation study for the neural network. For the random forest model, we averaged the permutation importance over the primary features (maxGaze, speaker-directed gaze proportion, entropy, active speaker index, presence). Averaged **permutation scores** in Figure 8 show that all features except maxGaze were important for model prediction - gaze presence being the most important one, then speaker-directed gaze, entropy and active speaker index. We also performed a **feature ablation** study on the neural network model to verify the inclusion of each feature in the training set for the model. Removing primary features one at a time showed a slight decrease in test accuracy compared to the model trained on all the features (see Figure 9). The only significant decline in test and train accuracy was when removing entropy or speaker-directed gaze (post hoc Dunn's p<0.001). Train accuracy was also significantly lower when removing the active speaker index feature if compared with the model trained on all the features (post hoc Dunn's p<0.001).

5 DISCUSSION

The first research question (**RQ1**) we aimed to answer was whether humans encode memorable moments in their non-verbal behaviour. The overarching hypothesis was that the most remembered moments are encoded with similar features to high group involvement (from [18]) since higher attention has previously been connected to better memorisation of information [16, 28, 30]. Specifically, in

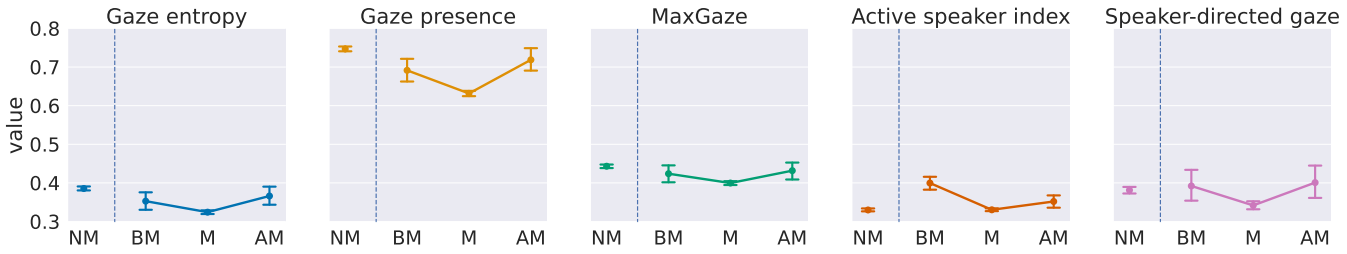
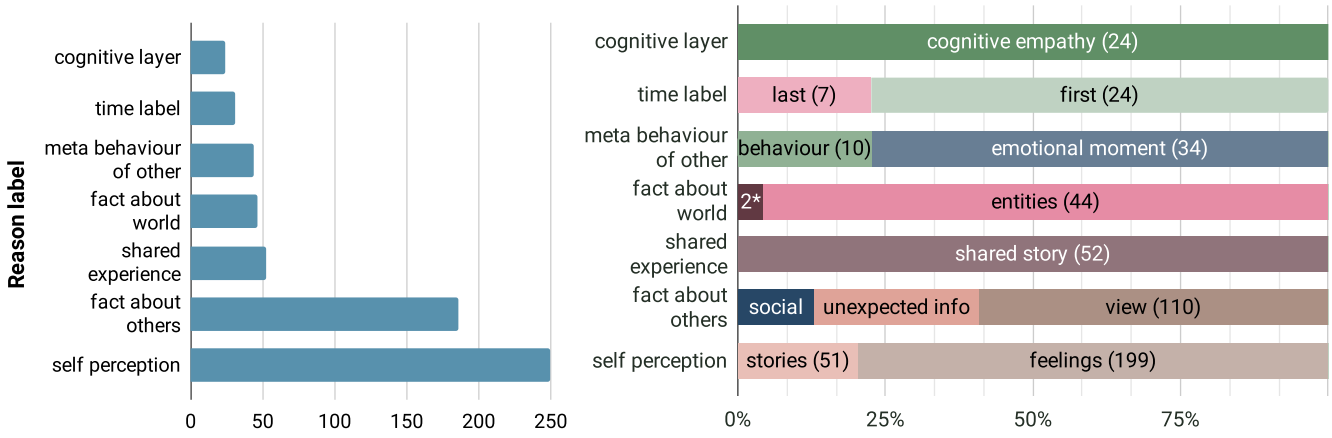


Figure 5: A comparison of gaze and speaker features in different moments in relation to their timing in relation to moments remembered by at least one participant in a group. The windows within such memorable moments - "M" on x-axis, two time slices before these moments - "BM", two time slices after M intervals - "AM", and all remaining time slices not included in the above - "NM".



(a) Number of memorable moments per label

(b) Sub-labels distribution within main reason labels

Figure 6: Visualisation of reasons label distribution: main label distribution over the whole data set (plot a), and sub-labels within the main labels (plot b). Important to note that a "memorable moment" in this context is the entire memorable interval, rather than a time slice as in the statistics for the gaze and speaker features. *"people"

Figure 7: A comparison of the performance of different classification models: balanced accuracy averaged over instances trained on 20 random samples of data (95% confidence intervals as error bars).

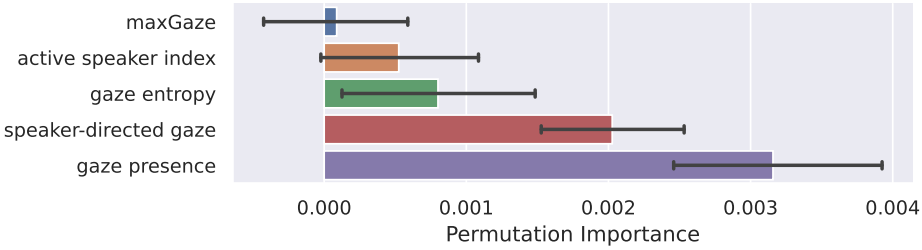


Figure 8: Permutation importance for the random forest classifier with 95% confidence intervals over the models trained on different balanced samples of data. The permutation importance score for each feature shown in the figure is averaged over 20 Random forest model iterations trained and tested on different random samples.

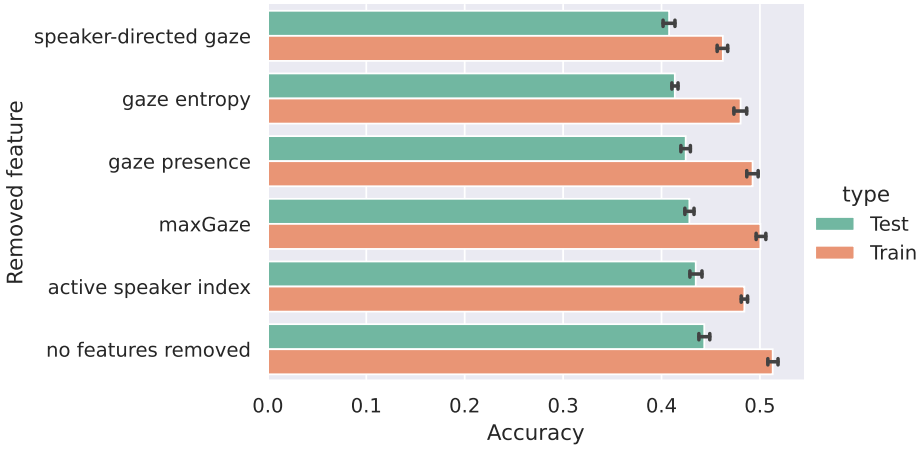


Figure 9: Ablation study results: accuracy of models trained on different sets of features with 95% confidence intervals over models trained on different random samples as error bars. "No features removed" below on the y-axis shows a baseline model with all features included. Other labels mean a feature being removed from the training sample.

previous research [18] higher involvement in the interaction has been related to a higher degree of group presence and maxGaze and to a lower degree of group gaze entropy.

The feature of group gaze **entropy** was in fact lower in highly memorable moments than in middle, low-level and non-memorable moments. This means that participants had more agreement in their gaze targets at moments most remembered by the group. Therefore, our hypothesis was confirmed in relation to the entropy feature.

The hypothesis that group involvement can be connected to memorised information was also confirmed in relation to the measure of **speaker activity**. A higher proportion of speakers were actively involved in the discussion in segments of high memorability. This is also consistent with the previous results of [14] and [10] on ego-centric bias in conversational memory. Specifically, [14] showed that speakers remember more of the produced content than their listeners. [10] has also shown that people remember better the parts of the conversation in which they were most active. Another trend we see in our data is that the active speaker proportion is at its peak at the moment preceding the remembered interval. This finding might be related to the fact that humans remember the reaction in response to their own words better than a reaction to other people's words [33].

However, the findings in the **presence** and **maxGaze** features were not in line with our hypothesis. In our case, presence and maxGaze were lower in highly memorable intervals than in non-memorable ones. The **speaker-directed gaze** feature was also lower in highly memorable moments. A potential explanation could be connected to cognitive load. Higher cognitive load has previously been connected to gaze aversion: when asked difficult questions people tend to avert their gaze [5] which has also been reported to facilitate remembering [7]. For our data, this could mean that memorable moments encourage more deep thinking, which highlights those moments over non-memorable ones.

Our second research question (**RQ2**) investigated participants' reasons for remembering memorable moments. The most common reason was self-perception. This means that people remembered a moment because of the feelings they experienced during the remembered moment or because of a view or experience they expressed themselves throughout the moment. This goes in line with the previous research on ego-centric bias [10] and means that people remember things that were personally distinct for them because

of their feelings rather than the factual content of the moment. The fact-about-world category was assigned less frequently than fact-about-other and self-perception.

The third research question (**RQ3**) posed explored whether group memorability of conversational moments could be predicted using our gaze and speaker-related features. The models that performed the best were the random forest and the neural network classifiers with 43-44% test accuracy for a 4-class classification of group memorability levels. In other words, the classification models were able to achieve nearly twice above chance performance (44% accuracy as compared to 25% chance of 4-class classification), although trained solely on gaze and speaker features. This can be considered an indicator of the strong predictive power of the selected features. We also wanted to investigate whether individual performance affects the model performance. This resulted in a considerable decline in the performance of our models, though still showing above chance accuracy (31%). This might mean that there are some group-related specifics which influence the performance and initial memorability distribution. The permutation importance scores of the random forest classifier identify that the most important features are presence, speaker-directed gaze, and entropy. MaxGaze and speaker-directed gaze were less important. The feature ablation study showed significant changes in model performance only when removing two features - speaker-directed gaze and gaze entropy. Removing other features from the training set of the model did not show significant changes to the original model. This can be an indication that the selected features are correlated with each other, which gives the neural network enough information about the removed feature from the remaining ones.

Giving a closer look at the examples of correct and incorrect predictions of the best-performing model (the neural network), it seems that the misclassification might be connected to several reasons. First, technical issues in the videos that result in an incorrect prediction of eye-gaze behaviour - for instance, participants wearing glasses or not enough lighting on a participant's face at a certain time segment. Second, following a thin-slicing approach, we are dividing memorable moments into smaller segments, where each segment has the same memorability rating as the other segments of that memorable moment. This means, we are treating memorability as a constant in each moment annotated as memorable. This is not always the case: for example, if there is a pause or a moment of hesitation in a longer memorable segment it would also be classified as memorable, while the non-verbal signals would indicate participants' disengagement. Last, the neural network might have misclassified some instances because of ambiguities in non-verbal signals. For example, a segment where most participants avert their gaze would be classified as memorable, since statistically speaking it is a signal indicative of memorability. However, it could also be a signal of disengagement and, therefore, the lack of attention needed to memorise the moment. In this case, the segment would be incorrectly classified as highly memorable. This highlights the need for a wider context for accurate predictions of conversational memorability. Specifically, introducing additional modalities, such as speech or prosody, along with constructs such as engagement or affect could help to solve these ambiguities.

6 CONCLUSIONS

In the present paper, we investigated whether it was possible to predict conversational memory from non-verbal multi-modal cues. We could show that gaze and speech activity features were able to distinguish between 4 levels of memorability on a group level. Highly memorable moments were significantly different from the low memorability moments for all the features. In highly memorable moments, participants were looking away from other participants (lower presence) more, they looked at the same participants (lower entropy) to a higher degree, and more speakers actively participated in the discussion in comparison to non-memorable segments. An important distinction between memorable moments and moments of high involvement seems to be gaze aversion (lower presence) in highly memorable moments. The memorable moments were also preceded and followed by specific cues. For example, more participants were actively speaking before and after the moment than within or further away from the remembered moment. The most common reasons the participants mentioned for recalled each specific moment were related to participants' personal feelings and experiences. The second most common reason was related to information on other participants in the session. This highlights the importance of analysis of conversational memory since participants did not recall factual information about the world, but rather the moments that were important for their social image and knowledge about other participants in the group. We could also automatically distinguish between 4-classes of memorability with an average accuracy of 44 % for the neural network.

Future research will include investigating the usability of additional modalities, such as text, prosody, turn-taking, and group-specific characteristics to predict conversational memorability. Other, more complex signals such as affect could also be an interesting area of future research. Adding a temporal component to memorability prediction models could also be promising to further investigate how memorable moments unravel in time. Further investigation of memorability reasons and their relation to non-verbal cues could also lead to a deeper understanding of conversational memory.

ACKNOWLEDGMENTS

The data collection of MEMO corpus was supported by the Design for Values institute at TU Delft. We thank Chirag Raman, David Tax, Marco Loog for the indispensable advice and feedback on the paper.

REFERENCES

[1] William L. Benoit, Pamela J. Benoit, and James Wilkie. 1996. Participants' and observers' memory for conversational behavior. *Southern Communication Journal* 61, 2 (March 1996), 139–154. https://doi.org/10.1080/10417949609373007

[2] Eduardo Camina and Francisco Güell. 2017. The Neuroanatomical, Neurophysiological and Psychological Basis of Memory: Current Models and Their Origins. *Frontiers in Pharmacology* 8 (2017). https://doi.org/10.3389/fphar.2017.00438

[3] Joana Campos, James Kennedy, and Jill F. Lehman. 2018. Challenges in Exploiting Conversational Memory in Human-Agent Interaction. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '18)*. International Foundation for Autonomous Agents and Multiagent Systems, Stockholm, Sweden, 1649–1657. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.5555/3237383.3237945

[4] Kevin Delgado, Juan Manuel Origgi, Tania Hasanpoor, Hao Yu, Danielle Allessio, Ivon Arroyo, William Lee, Margrit Betke, Beverly Woolf, and Sarah Adel Bargal.

2021. Student engagement dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3628–3636.

[5] G. Doherty-Sneddon and F. G. Phelps. 2005. Gaze aversion: A response to cognitive or social difficulty? *Memory & Cognition* 33, 4 (01 Jun 2005), 727–733. https://doi.org/10.3758/BF03195338

[6] GazeSense. 2022. 3D eye tracking software for depth-sensing cameras. https://eyeware.tech/gazesense/

[7] Arthur M. Glenberg, Jennifer L. Schroeder, and David A. Robertson. 1998. Averting the gaze disengages the environment and facilitates remembering. *Memory & Cognition* 26, 4 (01 Jul 1998), 651–658. https://doi.org/10.3758/BF03211385

[8] Zerrin Kasap and Nadia Magnenat-Thalmann. 2012. Building long-term relationships with virtual and robotic characters: the role of remembering. *The Visual Computer* 28, 1 (Jan. 2012), 87–97. https://doi.org/10.1007/s00371-011-0630-7

[9] Simone Kauffeld and Nale Lehmann-Willenbrock. 2012. Meetings Matter: Effects of Team Meetings on Team and Organizational Success. *Small Group Research* 43, 2 (2012), 130–158. https://doi.org/10.1177/1046496411429599 arXiv:https://doi.org/10.1177/1046496411429599

[10] Dominique Knutsen and Ludovic Le Bigot. 2014. Capturing egocentric biases in reference reuse during collaborative dialogue. *Psychonomic Bulletin & Review* 21, 6 (01 Dec 2014), 1590–1599. https://doi.org/10.3758/s13423-014-0620-7

[11] Olga Anatoliyivna Kulyk, Jimmy Wang, and Jacques Terken. 2006. *Real-Time Feedback Based on Nonverbal Behaviour to Enhance Social Dynamics in Small Group Meetings*. Number 10/3869 in Lecture Notes in Computer Science. Springer, Netherlands, 150–161. https://doi.org/10.1007/11677482_13 10.1007/11677482 ; null ; Conference date: 11-07-2005 Through 13-07-2005.

[12] Catherine Lai, Jean Carletta, and Steve Renals. 2013. *Detecting Summarization Hot Spots in Meetings Using Group Level Involvement and Turn-Taking Features*. https://doi.org/10.13140/2.1.4721.8564

[13] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[14] Geoffrey L. McKinley, Sarah Brown-Schmidt, and Aaron S. Benjamin. 2017. Memory for conversation and the development of common ground. *Memory & Cognition* 45, 8 (Nov. 2017), 1281–1294. https://doi.org/10.3758/s13421-017-0730-3

[15] Judi Beinstein Miller and Patricia Ann de Winstanley. 2002. The Role of Interpersonal Competence in Memory for Conversation. *Personality and Social Psychology Bulletin* 28, 1 (2002), 78–89. https://doi.org/10.1177/0146167202281007 arXiv:https://doi.org/10.1177/0146167202281007

[16] Klaus Oberauer. 2019. Working Memory and Attention - A Conceptual Analysis and Review. *Journal of cognition* 2, 1 (08 Aug 2019), 36–36. https://doi.org/10.5334/joc.58 31517246[pmid].

[17] Catharine Oertel, Kenneth A Funes Mora, Joakim Gustafson, and Jean-Marc Odobez. 2015. Deciphering the silent participant: On the use of audio-visual cues for the classification of listener categories in group discussions. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 107–114.

[18] Catharine Oertel and Giampiero Salvi. 2013. A Gaze-Based Method for Relating Group Involvement to Individual Engagement in Multimodal Multiparty Dialogue. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction* (Sydney, Australia) *(ICMI '13)*. Association for Computing Machinery, New York, NY, USA, 99–106. https://doi.org/10.1145/2522848.2522865

[19] Catharine Oertel and Giampiero Salvi. 2013. A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. 99–106.

[20] Catharine Oertel, Stefan Scherer, and Nick Campbell. 2011. On the use of multimodal cues for the prediction of degrees of involvement in spontaneous conversation. In *Twelfth annual conference of the international speech communication association*.

[21] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[23] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* (Hilton Waikoloa Village, Big Island, Hawaii, US). IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

[24] Steve Rathje, Jay J Van Bavel, and Sander Van Der Linden. 2021. Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences* 118, 26 (2021), e2024292118.

[25] Jennifer A. Samp and Laura R. Humphreys. 2007. "I said what?" Partner familiarity, resistance, and the accuracy of conversational recall. *Communication Monographs* 74, 4 (2007), 561–581. https://doi.org/10.1080/03637750701716610

[26] Samiha Samrose, Daniel McDuff, Robert Sim, Jina Suh, Kael Rowan, Javier Hernandez, Sean Rintel, Kevin Moynihan, and Mary Czerwinski. 2021. MeetingCoach: An Intelligent Dashboard for Supporting Effective & Inclusive Meetings. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 252, 13 pages. https://doi.org/10.1145/3411764.3445615

[27] Avinash Saravanan, Maria Tsfasman, Mark Neerincx, and Catharine Oertel. 2022. Giving Social Robots a Conversational Memory for Motivational Experience Sharing. In *Proceedings of 31st IEEE International Conference on Robot & Human Interactive Communication*. IEEE.

[28] Tali Sharot and Elizabeth A. Phelps. 2004. How arousal modulates memory: Disentangling the effects of attention and retention. *Cognitive, Affective, & Behavioral Neuroscience* 4, 3 (01 Sep 2004), 294–306. https://doi.org/10.3758/CABN.4.3.294

[29] Laura Stafford, Vincent R. Waldron, and Linda L. Infield. 1989. Actor-Observer Differences in Conversational Memory. *Human Communication Research* 15, 4 (June 1989), 590–611. https://doi.org/10.1111/j.1468-2958.1989.tb00200.x

[30] Geoffrey Underwood. 2013. *Attention and memory*. Elsevier.

[31] Piek Vossen, Selene Baez, Lenka Bajcetić, and Bram Kraaijeveld. [n. d.]. Leolani: A Reference Machine with a Theory of Mind for Social Communication. In *Text, Speech, and Dialogue (Lecture Notes in Computer Science)*. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-00794-2_2

[32] Britta Wrede and Elizabeth Shriberg. 2003. Spotting "hot spots" in meetings: human judgments and prosodic cues. In *INTERSPEECH*.

[33] E. Zormpa. 2020. *Memory for speaking and listening*. [S.l.] : [S.n.]. https://repository.ubn.ru.nl/handle/2066/227383