# An Automatic Evaluation Framework for Social Conversations with Robots

Maike Paetzel-Prüsmann
maike.paetzelprusmann@disney.com
Disney Research
Zurich, Switzerland

Jill Fain Lehman
jfl@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Celia J. Gomez
celia.gomez@disney.com
Disney Research
Glendale, California, USA

James Kennedy
james.kennedy@disney.com
Disney Research
Glendale, California, USA

## ABSTRACT

When deploying social robots in the wild, it is crucial for developers to gain an understanding of how the interactions between the robot and its human conversational partners are progressing. Unlike in traditional task-based settings in which a human and a robot work on a tangible outcome that can serve as a proxy for how well a conversation is going, social settings require a deeper understanding of the underlying interaction dynamics. In this paper, we assess a set of recorded features of a robot having social conversations in a multi-party, multi-session setting and correlate them with how people rated their interaction. We then propose a framework that combines the features into a model that can automatically assess an ongoing conversation and determine its performance.

## CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**; *User models*; HCI theory, concepts and models; • **Computing methodologies** → **Discourse, dialogue and pragmatics**; • **Computer systems organization** → Robotic autonomy.

## KEYWORDS

Human-Robot Interaction Evaluation, Conversational Quality, Social Robotics, Conversational Dialogue Systems

**ACM Reference Format:**
Maike Paetzel-Prüsmann, Jill Fain Lehman, Celia J. Gomez, and James Kennedy. 2024. An Automatic Evaluation Framework for Social Conversations with Robots. In *2024 International Symposium on Technological Advances in Human-Robot Interaction (TAHRI 2024), March 9–10, 2024, Boulder, CO, USA*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3648536.3648543


*TAHRI 2024, March 9–10, 2024, Boulder, CO, USA*

ACM ISBN 979-8-4007-1661-4/24/03
https://doi.org/10.1145/3648536.3648543

## 1 INTRODUCTION

Robots designed for social interactions face different challenges than those created for logistical tasks. In the latter, the outcome of the task serves both as the main incentive and the core metric for evaluation. When the human user has a clear need for support in a task with a tangible outcome, e.g., booking a ticket, there are numerous automatic measures that can help evaluate whether the interaction was successful [28]. For social robots, on the other hand, success is predicated on the experience being engaging. Interactions with social tasks like cultivating relationships or entertainment hence require defining more complex and nuanced measures.

A systematic review of humans interacting with social robots identified four main evaluation techniques used in these settings, with two (questionnaires and interviews) relying on self-reporting from the users, one requiring post hoc annotations of video recordings, and the last and smallest considering biometrics as more objective measures [16]. Each of these techniques has a significant barrier to its application in long-term deployments. In the case of surveys and interviews, most users do not opt in to report, and if they do, they are unlikely to do so repeatedly. In the case of recordings and biometrics, users can have privacy concerns or be unwilling or unable to wear additional hardware equipment.

For developers, understanding the success of an interaction and the user's perception of it is essential, particularly in long-term interactions. Even if a robot was well-tested before the final deployment, system components can fail, or user behavior and expectations can change over time. As an example of the former, consider a social robot that is expected to make and understand pop-culture references and so is given access to a popular movie database to query. A typical system component failure would occur if a database becomes unavailable, a severe error but one that is easy to detect automatically. Our work is concerned with the latter, more elusive issue; specifically, the decline in user satisfaction because the user and their expectations are changing how they approach the system over time. In this case, a robot able to reference meta-data from a movie database might have been sufficient while the system was tested and smart speakers were still people's main point of reference. But as more people gain access to advanced AI technologies, they might come to expect the robot would be able to talk about the content of movies, something the extant robot is not able to do, leading to a decline in user satisfaction with its interactivity.

In cases where a deployed robot has hundreds of interactions each day, it is not feasible for the developers to read system logs to assess manually the success of interactions on the level described above. Automatic flagging of problematic experiences or more systematic failures over time would allow system designers to understand when system updates are necessary and what areas would bring the most improvement to the experience.

In this paper, we present ongoing work in developing an automatic model to judge conversational quality in a multi-party, multi-session experience between human users and a social robot. While most prior work in the area of social Human-Robot Interaction (HRI) has focused on engagement tracking using visual measures of affect, our system does not have video input available. As there is limited research regarding what conversational features correspond to user engagement and satisfaction in social dialogue [13], we focus our discussion on the relationship between features we extracted from a previous deployment of a social robot and the user's rating of the experience. We then discuss how the features are currently used within our larger evaluation framework and our plans to extend its use in the future.

## 2 RELATED WORK

Robots interacting with humans have been applied in many areas, from the hospitality industry to healthcare and entertainment [3]. Measuring the successful use of a robot typically involves task-specific metrics. For example, in an educational context a robot's success would be measured through increasing the knowledge of the human partner [5], while in healthcare, the improvement of symptoms is a likely measure of interest [18, 20].

At the same time, robots need to solve similar underlying tasks, like retrieving information, and metrics that work across different contexts can measure part of the interaction success [25], independent of task domain. However, social aspects play an important role when humans interact with robots, as robots are inherently seen as social actors [31]. In the following, we review the evaluation of robots developed for purely social tasks, like entertainment, focusing on measures specific to the social aspects of the interaction, as well as evaluation of social dialogue systems more generally.

### 2.1 Evaluation of social robots

Jung et al. [16] organized evaluation techniques of the social aspects in HRI into four main categories: questionnaires, video analysis, interviews, and biometrics. The number of questionnaires developed for social robotics is growing, covering broad scales of user perception [4, 11] to more specific aspects like trust [30]. While questionnaires can be helpful instruments to understand human behavior, data from questionnaires are usually not available once a robot is deployed, as users are typically not willing to invest this time. More importantly, questionnaires cannot assess the interaction while it is ongoing without disturbing the experience; they do not provide a continuous measure throughout a conversation.

Video analysis of interactions can be performed using human coders or automatic metrics. In both cases, human affect is the main measure used to evaluate a robot, most commonly through visual input [22]. A systematic review of affective computing concluded that visual signals are most effective in recognizing humans' emotional states [29]. Another technique that has shown to serve as a proxy for engagement is the analysis of gaze and attention [2, 7]. However, the interpretation of gaze in social settings may vary substantially with and without a common frame of visual reference [24].

If visual input is not available, as is common in deployed systems for privacy reasons, audio features paired with text transcriptions of the interactions can provide successful classifications. However, the context of many audio datasets is distinct from natural interactions, either because humans are asked to act emotions, or because they are taken from scripted contexts like movies [17]. In natural settings, emotions are often more subtle and harder to detect [10, 27]. Moreover, especially in entertainment settings, the correlation between emotions and engagement can be complex as even emotions traditionally considered to be negative can be desired in certain parts of a narrative. If the conversational content is taken into consideration, as in Bohus and Horvitz, the focus is mostly on the current dialogue state and the time spent in a particular state [7].

Tian and Oviatt developed a taxonomy of social errors in HRI which relies on socio-affective competencies [26]. While these provide useful considerations, many (e.g., 'Synthesizing inappropriate or absent non-verbal expressions') are difficult to automatically detect during an interaction. For this paper, we incorporate a subset of their features appropriate for an automatic assessment given the current state of the art. Honig and Oron-Gilad also discuss understanding errors in HRI with a focus on automatically detecting and resolving them [15]. While they touch on communication failures, the main point of interest in our work, they note that not a lot of literature is available on that topic, and so focus on system failures instead. Andrist et al. examined situated interaction failures in the wild [1]. Similar to our approach, they examined how certain error classes influence the interaction score with their social robot. Their analysis, however, focuses on component failures for an automatic assessment and relies on manual coding for timing and content analysis. Our work aims to extend this by analyzing features that can be applied in automatic tracking of engagement.

Finally, using social signals as automatic rewards in human-robot interaction is of increasing interest to the reinforcement learning community. However, most related work discusses explicitly giving reward feedback, either verbally or non-verbally [19], which is not a natural behavior for humans to exhibit in social conversations.

### 2.2 Evaluation of dialogue systems

As this paper is focused on verbal interactions with social robots, work related to the evaluation of (disembodied) dialogue systems is relevant. One of the earliest frameworks for evaluating spoken dialogue agents is PARADISE [28]. It considers increasing user satisfaction by maximizing task success and minimizing the user's cost to achieve the task. Their main contribution is measures related to task efficiency, e.g., number of utterances and dialogue time. While this work remains foundational for evaluating dialogue systems, its focus on solving logistical tasks poses a challenge when it comes to adapting measures for social use cases. In Sec. 4, we discuss potential re-interpretations of these features in social settings.

Another traditional approach relies on overhearers of the conversation to judge its quality [8]. While this may come closest to the judgement of the situated interlocutor, this approach fails to

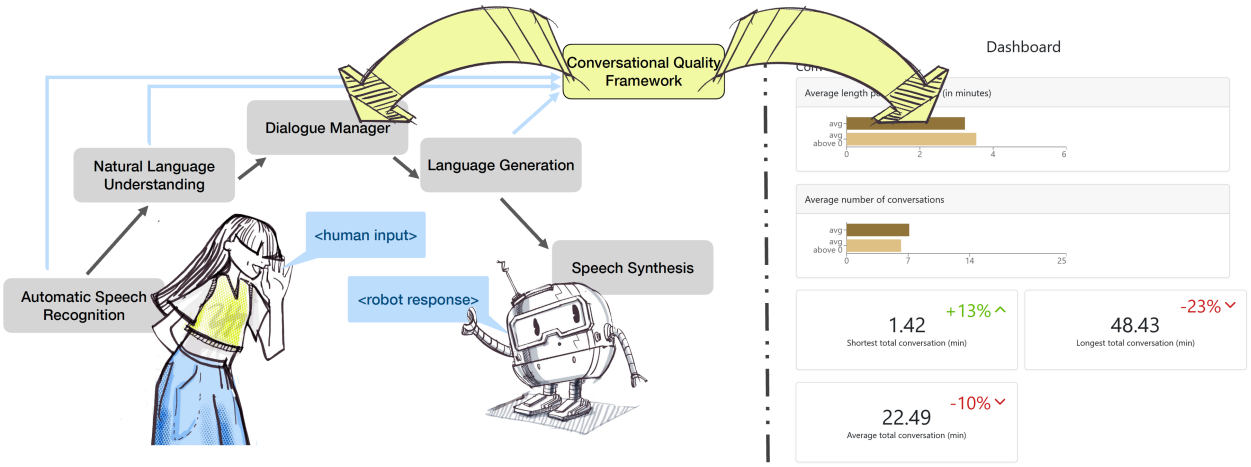


**Figure 1: The main components of a robot's dialogue system (left) and how our proposed framework component (yellow, center top) will listen to messages across components in the system (blue arrows), serving as an input to the dialogue manager as well as the basis for the Conversational Quality Dashboard (right). From the system input, the framework component calculates features as described in Sec. 5. The component and the dashboard are described in detail in Sec. 6.**

scale. A recent review by Deriu et al. [13] on the evaluation of dialogue systems revealed that there are still major gaps in automatic evaluation techniques for conversational dialogue that go beyond merely judging the quality or relevance of text generated by the system. Curry, Hastie and Rieser [12], for example, discuss traditional scores like BLEU and ROUGE, and shortcomings in how they can be applied to generating social dialogue. Other prominent work in the area of generative dialogue focuses on dialogue breakdown, which are lines produced by an agent after which it is either challenging or impossible for a user to smoothly continue the conversation [14]. While all of these can individually relate to a decrease in user engagement with a social agent, we agree with Deriu et al. [13] that a broader set of evaluation metrics is required to assess the conversational dialogue holistically. This paper aims to make a contribution in this direction.

## 3 RESEARCH QUESTION AND SCOPE

In this paper, we aim to answer the following research question and contribute to our long-term goal of enabling a social robot to autonomously judge the quality of an ongoing conversation (Fig. 1).

(RQ) Which features that are automatically extractable from social spoken interactions contribute to a high-quality interaction?

We focus on social settings in which a robot interacts with one or more individuals. Features specific to multiple conversational parties and multiple sessions are discussed separately and while our proposed framework has only been tested in such settings, we have confidence our techniques are applicable to interactions with single sessions and individual users. Given the current state of the art, we believe that in order to create truly social and engaging interactive settings, human experience designers must carefully curate the interaction. Hence, settings in which the robot plays an entirely passive role and is not at least partially driving the experience forward are outside of the scope of this paper. Moreover, we focus on interactions that are primarily verbal.

## 4 IDENTIFYING FEATURES TO EVALUATE SOCIAL CONVERSATIONS

In the following, we describe features that can be tracked automatically in social conversations and how they relate to features used in more traditional settings involving a logistical task (Sec. 4.2). We then propose different techniques to evaluate these features and their meaning for user satisfaction (Sec. 4.3).

### 4.1 Methodology

In addition to considering features from the related work in social context, we also conducted an interviews with 28 professional team members that all have experience working on at least one deployed social robot interaction. Interviewees were recruited from a diverse background, ranging from computer scientists to experience designers, writers and quality assurance team members. Each interview lasted an hour and was semi-structured around what conversational quality means to them. Two of the authors then conducted a thematic analysis on the interview notes to compile the feature overview presented in the next section.

### 4.2 Feature Overview

*System Component Performance.* In most dialogue engines that follow a modular structure, the performance of the components can be tracked individually. Traditionally, this is done by logging *component errors* as well as *latency*. For any component that involves stochastic decision-making or classification, the *confidence* in the decision can be recorded. Especially for Automatic Speech Recognition (ASR) and Natural Language Understanding (NLU) components, most evaluations to date require a manual correction of the component output. As this is slow, time consuming and may be infeasible in certain settings given privacy concerns, another option is to compare several ASR and NLU software systems after the interaction. This has the advantage that the ground-truth can be determined in a majority vote between the software systems if their

results show high overlap. Moreover, a high divergence between their results can be used to lower the confidence that the original transcription or intent classification of the system was correct.

*Generic Conversational Features.* Many of the conversational measures that are applied in evaluations of conversations can be applied in purely social contexts [28]. However, their interpretation may need adaptation to the different standards that apply in social conversations. For example, a lower *interaction length* in a successfully resolved task-based interaction is usually considered better as this indicates increased efficiency in solving a task. In social interactions, the ideal conversation length depends on the amount of content that is available for the selected activity. For example, in a game-based interaction where the players have not yet found a certain clue, a short interaction is to be expected. If the ideal length of conversation for a given activity is unknown, comparing against the average conversational length of the same interaction type can be a good proxy. The same is true for the expected *turn length* of the user. If the robot asked a yes/no question, a brief response is of no concern. However, if they engage in collaborative story-telling, then one-word responses could be a sign of disengagement. If the input of the designers on the intended input length is missing, a comparison to the average response length can yield good results. Another component of turn-taking is the presence of *overlapping speech*. This is dependent on the social context - in some conversations interruptions and overlapping speech can show engagement, while in others simultaneous speech would violate social norms. Even humans are not perfect in interpreting turn-taking cues, which means some overlapping speech is expected, so a comparison against the average is beneficial. Both in task-based and social settings, the conversation can either be *terminated* by the system (if a task is completed or all content explored), or by the user if they have no further need or time to interact. If more content is available and the user still disengages, this could be a potential sign of frustration. If the user is actively requesting more conversations but the robot disengages, this could also lead to dissatisfaction.

*Conversational Content.* In dialogue systems engaging in logistical tasks, content is often judged by task progression, requests for help, fallback trigger frequency, and user request rejections. With adaptation, these same measures can be applied in social dialog. If, for example, *fallback* triggers a generic response when the NLU cannot match the user input to available output, then the number of times a fallback occurs will likely correlate to how well the user feels understood. Similarly, if the robot knows its limitations and needs to *reject* a request for content, this will likely decrease engagement with the system. However, tracking these in social settings is challenging due to the broader range of potential user requests compared to task-based settings.

We propose two techniques for capturing content progression in social settings. The first is simpler and tracks the percentage of *seen versus new content* that is used in the conversation. Which one leads to more engagement will depend on the specifics of the social setting. In certain moments, sharing previous memories may be desired, while in others novelty is key to a good experience [9]. For a more narrative-based experience, available content can be captured in an experience graph. This graph represents the available content on different levels, from the abstract topics down to the individual lines that were uttered by the robot. We can track content progression using *graph-traversal algorithms* that capture how many unseen parts of the graph were explored. Unlike other techniques, this also gives the opportunity to detect *conversational loops* users got stuck in, and affords analysis of the *order of conversational content* being explored. In the latter case, we can match this against an ideal vision from the designers, or the average experience of different users. If certain parts of the graph are only accessible when certain *system beliefs* apply, we can track how often the beliefs are updated and whether the updates happen in the expected order and at the expected moment.

Finally, we propose *reactions to moments of interest* in the conversation as an important novel feature. Such moments of interest can be the offer of additional content, or moments that are expected to carry emotional impact. Imagine a robot in an escape room scenario that acts as a character in the game. It offers background information on the scenario and asking it the right questions at the right time will lead to it revealing clues required to progress in the game. If you approach the robot at a moment that is not time critical and where it does not have new clues to reveal, it could offer you stories about the place or items in it. Opting to hear this content even though it is not integral to progress can be a sign of being invested. If one of the background stories told by the robot covers topics such as loss, then an empathetic reply from the user is to be expected and could show their engagement with the robot.

While these moments can be determined after the experience has been designed, they can also be explicitly implemented as an intentional measure. For example, a moment where a quick reaction is required can be written into an experience, and the reaction time can then be used as a proxy for engagement in the moment.

*Robot Personality.* In social interactions, the personality of a robot is of special importance as it sets people's expectations about how the robot will react in certain moments. As a result, both the *amount of personality-driven responses* and the *coherence in the personality* are important features in these settings. If the content is hand-authored and pre-annotated for personality by the experience designers, the annotations can be used directly. If the content is at least partially generated autonomously, a post hoc evaluation by language models, for example, can provide an estimate. As a robot's personality shows also in its body language, annotations or post hoc classifications of its non-verbal behavior and how well it matched the content are important as well.

*Multi-Party Settings.* In cases with several conversational partners, we can track the *individual speech time* of the different interlocutors. Moreover, *cross-talk* between the human users, as well as *side conversations* and comments about the experience will be more prevalent in multi-party settings.

*Multi-Conversation Settings.* If users have the option to approach the robot for additional conversations, *the number of interactions* carries meaning for engagement. If the robot can indicate the availability of new content or its own desire to talk, the time between the delivery of the message and the user initiating the conversation can be captured. If the robot can initiate the conversation, the *rejection rate* of the user can be tracked. If the robot is only part of a larger experience, or is situated in an environment where users

can engage in other activities (like people's homes, or puzzles in an escape room), the *relative amount of time spent with the robot* and whether it declines over time can indicate a change in engagement.

### 4.3 Assessing Features

When using the above features to evaluate a social robot, both the availability and the importance of each feature will vary depending on the exact robot and social setting. Hence, selecting features and assigning weights is key to developing a model that can assign a conversational score on a large scale. As such, ground-truth annotations of conversations within the setting are required initially to allow the model to then run independently of direct user feedback. If extensive user testing with the real environment can be carried out pre-deployment, targeted questionnaires will give the best results. However, the importance of certain features may change over time and extensive testing under realistic conditions is often not feasible. Below, we discuss three techniques that can be applied to automatically assign and continuously update feature weights while requiring minimal involvement of the end users.

*(1) Quantitative Experience Rating.* While questionnaires cannot be used in real deployments without disturbing the experience itself, rating the overall experience post-hoc on a single scale is common. Such scales often use generic terms or pictorials to assess user satisfaction and can be presented when leaving the physical space of an interaction. In the case of an at-home experience, the rating can be elicited via email, app or the device itself, such as the Alexa smart speaker asking for user rating as part of the Amazon Alexa Challenge. In a long-term deployment, users can be approached infrequently to update their experience rating. Quantitative experience ratings are used directly in our framework to assign weights to features or feature combinations that best predict the ratings.

*(2) Qualitative Experience Rating.* Users can be asked to provide a statement about their experience. We can use automatic sentiment tracking to discern positive and negative comments, and then use intent recognition via keyword matching or more sophisticated algorithms to extract areas of concern. For example, if many users mention long response times in conjunction with a negative sentiment, then component latency can be given a higher weight.

*(3) Meta-Comments.* In multi-party settings, users often engage in meta-conversations about the experience with each other. This can either be in the form of small comments ("This is so cool!"), or as a direct criticism of certain features ("I don't think it really understands us"). If utterances are automatically classified as system-directed or not, then a technique similar to (2) can be used to extract meaning. Although such comments may overlap with the kind of statements given in post hoc interviews, the execution context - the system state at the moment the comment is made - provides important additional data, e.g., an NLU confidence score associated with utterances prior to complaints about language understanding.

## 5 FEATURE EVALUATION IN TEST DEPLOYMENT

To better understand the utility of these features, we applied them to a multi-party experience where individuals or groups could engage with a social robot. The experience was narrative-based and all responses were authored by experience designers. We annotated the parts of the interaction intended to carry special value in the story arc with the designers after the test was performed. Additional annotations, e.g., about the intended response length or how much of the robot's personality was shown in individual lines or parts of the experience were not available. People could decide to engage with the robot as often as they chose, but due to the nature of the narrative, the overall available content was limited and the experience ended when no further content was available. At certain moments in time during the experience, the robot did not have new content available because users had not finished other parts of the experience. At those times, if users approached the robot, it would reject their request to communicate.

After completion, all user groups were asked to rate the experience on a five-point Likert scale with higher numbers indicating higher satisfaction, and optionally give an explanation of their rating in written form. Post hoc evaluation was optional and most users left without providing a rating. For privacy reasons, no demographic information was collected about the users. Users were informed that audio during the experience would be recorded.

In the following, we discuss some of the features outlined above by describing automated techniques for extracting the features and how they correlate with the rating of user's overall experience. We then perform a principle component analysis (PCA) in R to understand the correlation between features and how much they contribute to explain the variance[1] in our data sample.

### 5.1 Feature Assessment

A Shapiro-Wilk test of normality on each feature discussed below found that all violated the assumption of normality. Hence, we used Spearman's rank correlation analysis on the data to account for the non-normal distribution and to treat the user ratings as ordinal rather than continuous. An overview of the feature correlations and their significance is shown in Table 1.

*System Component Performance.* To understand the reliability of the ASR and NLU accuracy, we generated ground truth using human annotators so as not to introduce potential confounds from errors of other systems. To correct the ASR transcripts, the annotators received access to a randomly selected sample of the recordings. Overall, we have about 25% of our data manually annotated. The ASR accuracy is calculated using the JiWER library[2] by considering the Word Error Rate (WER) on the uncased text, with punctuation removed. The WER is 10.4%, i.e., around 1 in 10 words is erroneous, which is comparable to human-level performance of 5%-11% on similar conversational transcription tasks [21]. The same set of annotators corrected the classified intent where they believed the intent assigned by the system did not fit what the user said. The average accuracy was 88.7%, i.e., around 1 in 10 classifications was incorrect, based on the text transcript and the options available at that conversational turn. This ground-truth annotation requires deep understanding of the options at each turn of the conversation. Due to the highly contextual nature of the annotation, a meaningful baseline from other datasets or scenarios is difficult to identify.

[1] Performed using the factoextra package in R, https://cran.r-project.org/web/packages/factoextra/index.html
[2] https://pypi.org/project/jiwer/

**Table 1: Correlation between the user rating and features. Significant values are indicated using * ($p < .05$) and *** ($p < .001$)**

| | ASR | NLU | Total Conv. Time | Avg Conv. Time | Max Conv. Time | Min Conv. Time |
|---|---|---|---|---|---|---|
| Rating | -0.009 | 0.068* | 0.21*** | 0.183*** | 0.21*** | 0.03 |
| | Number of Conv. | Avg Turn Length | Number of Overlaps | Percent Fallbacks | Story Reaction | |
| Rating | 0.109*** | 0.007 | 0.066*** | -0.121*** | 0.022 | |

The relationship between ASR accuracy and ratings of the experience was not significant, $p = .74$ and accuracy ranged between 0.87 and 0.91. The relationship between the NLU accuracy is weak, $\rho_S = 0.068$, but significant, $p = .016$ and ranged between an accuracy of 0.8 and 0.87. This suggests that our system performs the tasks of ASR and NLU similarly well across all interactions, which means accuracy alone is not a good indicator of user experience.

*Generic conversational features in a multi-conversation setting.* Our system tracks when a user wakes up the robot, as well as when the user sends the robot to sleep or the robot makes the decision to end the conversation itself. We define the time in between these two points as the *conversation time*. For each group that does the experience, all of their interaction times with the robot are summed to a total conversation time. From this, we calculate the average, shortest, and longest length conversations with the robot.

There was a significant positive relationship between the total conversation time with the robot and human experience ratings, $\rho_S = 0.21, p < .001$, as well as between the average conversation time and the rating, $\rho_S = 0.183, p < .001$. Users with the lowest experience rating spent on average $M = 14.2$ minutes talking to the robot, while those rating it the highest spent $M = 29.7$ minutes with it. The length of each conversation varies between $M = 2.6$ minutes on average for those giving it the lowest rating and $M = 3.9$ minutes for those giving it the highest rating.

While the maximum individual length of a conversation was positively and significantly correlated with the experience rating, $\rho_S = 0.21, p < .001$, the correlation of the minimum length was not significant, $\rho_S = 0.03, p = .086$, suggesting that people who have a potentially very long and hence successful conversation rate the robot favorably, while the decision to end a conversation quickly can be for many reasons, independent of the robot behavior.

We can pull back from the level of individual conversations, to look at the experience as a whole. In doing so, we found a significant positive relationship between the number of times people approached the robot and their rating of the experience with it, $\rho_S = 0.109, p < .001$. Users giving the experience the lowest rating approached it on average $M = 6.2$ times, while those giving it the highest rating conversed with it $M = 9.4$ times.

Similarly, we can drop down a level to examine turn-level phenomena. For each utterance that was transcribed by the ASR, we extracted how many words were spoken by the user. The average length of the user's turn was not significantly correlated with the user rating, $p = .637$. However, the user's amount of overlapping speech with the robot, which we calculate as the number of times the robot was in a speaking state and there was ASR input transcribed within the same time range, had a significant but weak positive correlation. This suggests that more overlapping speech occurs in users rating the robot experience higher, $\rho_S = 0.066, p < .001$.

*Conversational content.* The percentage of fallbacks can be extracted directly from the NLU messages by dividing the number of unrecognized intents by the number of total requests made to the NLU. Analysis shows a negative correlation, suggesting that users giving the robot the highest rating experienced a lower number of fallbacks ($M = 15.72\%$) compared to users rating the robot the lowest ($M = 20.87\%$), $\rho_S = -0.121, p < .001$.

Using the post hoc manual annotations of the conversational graph with the input from the content designers, we picked one moment that the designers identified as a key emotional moment in the narrative. Our expectation was that highly engaged users would wake the robot up immediately after that story moment to resolve the cliffhanger. We designated a time period for the next initiation of a conversation to be considered immediate, tracked how often users approached the robot in this time period, and correlated the result to the user's experience rating. Interestingly, we did not find a significant correlation between users approaching the robot and their rating, $p = .257$ for this moment.

As a final feature, we picked one moment in which the robot offered two optional story elements. From a manual and unstructured analysis of the user's qualitative ratings of the experience we understood this story element to be one of the user's favorite. The story element was entirely optional and did not influence the main experience progression. We automatically extracted whether a user group was offered the story content by the robot (coded 0 if not), and then if they rejected the content offer (coded as 1) or if they made one of two story choices in that moment (coded as 2 and 3). Our hypothesis was that the story content users chose would not matter, but seeing this story episode would influence their overall rating. A Kruskal-Wallis rank sum test demonstrated that the choice was significantly related to people's experience rating, $p < .001$. A Dunn's post hoc test corrected for multiple comparisons using the Benjamini-Hochberg method showed that users who were not offered this choice at all gave the experience a significantly lower rating than all other user groups (code 0 vs. 1, 2 and 3; $M = 4.21$), users that rejected that story episode gave the robot experience as a whole a significantly lower rating than those that chose a content (code 1 vs. 2 and 3; $M = 4.41$), but the choice of content did not make a significant difference (code 2 vs. 3; $M = 4.56$ and $M = 4.61$).

## 5.2 Feature Selection

To understand the relative importance of the features discussed above and estimate their predictive capacity within the set of all available features, we first normalized our features and then checked the correlation between them (see Fig. 2). We see, for example, that the total number of interactions and the total length of all conversations is highly correlated. This is unsurprising, as more conversations also allow for more overall interaction time. We then performed a PCA, which revealed that the first component could

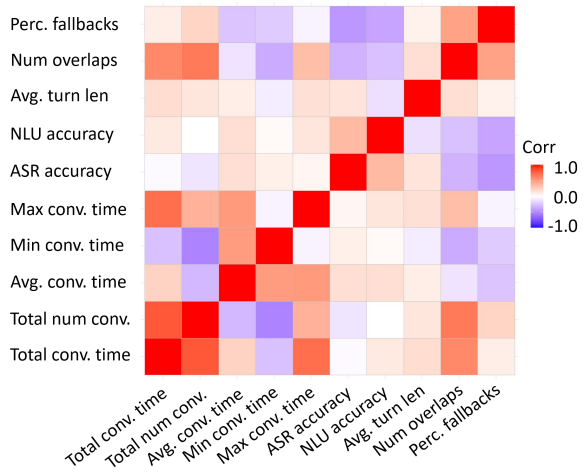


**Figure 2: Correlation matrix of the subset of features examined in this paper.**

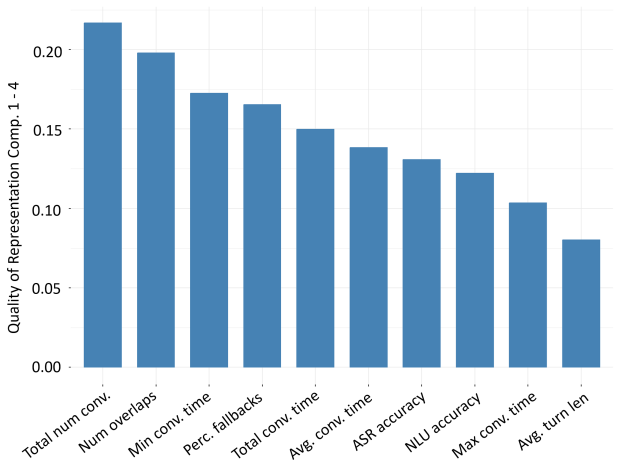


**Figure 3: Contribution of the individual features to the first four principle components.**

explain 58.66% of the variance, the second 18.97%, the third 11.73% and the fourth 6.52% for a cummulative proportion of 95.88%. All remaining components explained less than two percent each.

Fig. 3 shows that the total number of conversations contributes most to the first four principle components, followed by the number of overlaps in robot and human speech, the minimum conversation length, and the percentage of fallbacks triggered. We excluded two of the features assessed in Sec. 5.1 from this analysis (reaction to the story moment and the story choice), as the features cannot be normalized on a comparable scale.

The results of the PCA analysis combined with the correlation analysis in Sec. 5.1 argues for the exclusion of some features from the analysis of this particular experience going forward. As the user's turn length had both the lowest contribution to the data variance and little correlation to the user rating of the interaction, this feature can be removed. Despite having a high correlation with the user rating, the maximum length of an individual interaction can be excluded as it also had a high correlation with the total time of interaction while explaining little of the variance in the data. For our experience, both ASR and NLU accuracy could potentially be excluded from the final set of features as their PCA contribution was low and their correlation to the user rating small. However, as component updates could change this in the future, we decided to keep them in regardless. Finally, based on its low correlation to the user rating, we can remove the story reaction from the feature set.

# 6 THE CONVERSATIONAL QUALITY FRAMEWORK

We are developing a framework to take advantage of the kind of feature analysis demonstrated in Sec. 5, to serve two purposes. First, to visualize system performance in a way that gives developers the ability to see changes over time, particularly with respect to features that differ from the expected norm. Second, to provide a data source for the robot's dialogue system to facilitate adapting dialogue content based on the quality of the conversation. An overview is depicted in Fig. 1, which shows how the framework captures messages from components across the dialogue system to analyze the features throughout an interaction. Our framework is meant to augment spoken dialogue systems, and it can be implemented as a component in both custom dialogue frameworks as well as standard ones such as Microsoft psi [6] or retico [23], among others. Its main components are described below.

## 6.1 A Dashboard for Explainable AI

We have developed a Conversational Quality Dashboard that allows experience developers to monitor the quality of the robot's interactions over time (see Fig. 1, right side). The dashboard reports both cumulative statistics of all interactions in a selected time frame (e.g., a day or a week) and depicts its trend in comparison to the global average since the start of the test (see green and red percentages in Fig. 1). The dashboard also makes it possible to inspect individual interactions. Apart from aiming to explain AI systems, privacy is at the core of our implementation. While the dashboard does allow inspection of the system statistics and extracted features of individual conversations, identifiable information is hidden.

Having identified and selected an initial subset of features, we are now in the process of training models to automatically assess the overall quality of the conversation and assign it a value between 0 (indicating low conversational quality) and 1 (indicating optimal conversational quality). In the future, the dashboard will showcase this score to make it easier for developers to find conversations of interest to inspect. As we acquire more data, we intend to refine and adapt the feature set relevant for this social robot experience. We intend for the feature set to vary across the different experiences to which the working framework can be applied. In each case, we imagine that a feature evaluation like the one presented above will provide evidence to determine which features are best suited to capture and evaluate conversational quality.

## 6.2 Improving Robot Decision-Making

While the dashboard is currently regenerated after each group has concluded all interaction sessions and serves the sole purpose of

generating insights post hoc, other ongoing work is centered around giving the robot's dialogue system a quantification of how well the interaction is going in real-time. For this, our dialogue framework component continuously listens to all system messages that relate to the features of interest. It then keeps a set of the latest features for the model to assign a score to. The frequency of running the model is variable and supports different levels of incrementality, e.g., after every user turn, or after a certain amount of time has passed. The features used for the model have different update cycles, which poses a challenge for the component composing the latest set of features. Certain features are tied to the same turn, e.g., the length of a turn and whether the NLU could recognize an intent in it. Others, like the amount of cross-talk recorded in the background are turn-independent. By assigning unique identifiers to the messages flowing through the system, we are able to trace them back to the same turn. This allows the feature set to be released only after all measures related to a new turn have been recorded, which ensures the integrity of the features made available to the model. For future experiences, our goal is provide special content to the robot's dialogue system to use when the conversational quality framework's scores decrease. With this, we hope to be able to better recover from users disengaging from the experience.

## 7 DISCUSSION AND FUTURE WORK

In this paper, we discussed ongoing work to develop a framework to automatically assess conversational quality in social HRI. While our final system, and particularly the selection of a model to combine the features into a single conversational quality score is still under active development, we believe that the features we are extracting and how we are using them to judge conversational quality makes three contributions that are of interest to the HRI community.

First, we *discuss a set of well-known features in the dialogue community and analyze how they apply to social conversations.* We find that some behave differently with respect to task-based interactions than we would expect. Longer interactions correlate with higher ratings of the interactions in a social setting, whereas a shorter time is favorable in comparable settings of solving logistical tasks. We found that other features did not carry the anticipated meaning when examined in situ. For example, when considering the minimum conversation time in a multi-session setting, we hypothesized it would negatively correlate with user satisfaction, either because it indicates users giving up in frustration, or because there was no new content available, which again could lead to dissatisfaction. The data revealed, however, that while the minimum conversation length was of importance in explaining the variance we saw in the data, it had little correlation with the user satisfaction rating.

The second contribution of this work is the proposal of *using decisions in and reactions to episodes within a narrative as a feature for conversational quality in social contexts.* While we found that one episode that was believed to carry special meaning was not correlated with a user's experience rating, another moment had a significant influence on the user rating. What is interesting is that the first moment plays an important role in the main narrative, while the second one serves more as a filler if no immediate content is available. This highlights a strength of our framework: Showing system designers which parts of the interaction to pay attention to in future iterations. We plan to further explore how to reengage users dynamically by re-prioritizing moments in the dialogue manager automatically.

Finally, this work *contributes to the area of explainable AI by showcasing how we are visualizing our features for system designers to inspect* by highlighting differences in comparison to previous interactions. This allows developers to track system performance over time, and detect potential problems more easily. In the future, it will also help to explain the decision of the model we are developing to automatically combine the features into a single conversational quality score. The current set of implemented features are drawn from the larger set in Sec. 4 and were restricted by an analysis of their predictive power in a multi-conversation, narrative-driven social setting. We are in the process of validating the features in a different robot scenario; more work is necessary to fully understand the transferability to other social domains. We also believe that certain sets of features will differ depending on the implementation of the robot's conversational engine. For example, the ASR and NLU accuracy may carry more meaning if these components differ more in their performance between users or if the complexity of the input was higher than in our setting. Due to privacy regulations, we are unable to release our data. However, we believe that the contribution of this work is the feature description, their interpretation, and combination into the framework and will hopefully accelerate the work of other researchers wishing to adapt it to their context.

To weight our features, this work relied solely on ratings given by users. We acknowledge that, in not requiring every user to give feedback, the sample of ground-truth annotations we have may be biased, as users tend to give feedback when they either had very high or very low satisfaction with their experience. Indeed, we found responses reflecting very poor experiences were heavily under-represented. We believe, however, that identifying features that correspond to high user satisfaction is a first important step to learn more about conversational quality. In the future, we are planning to manipulate the conversations and artificially create negative examples for our users to collect a more even sample. We note that the rating of users may not correlate with an expert rating of what a conversation of high quality looks like. However, we argue that capturing the subjective conversational quality of users is of higher value as it is likely predictive of a future engagement.

This work is currently limited to settings where the robot's interactive purpose does not rely on physical manipulation. Our framework could be adapted to such settings, for example by tracking objects in the robot's surroundings. We also believe that our framework can be adapted to robots with varying autonomy and could bring valuable input to operators of teleoperated robots.

In the future, we are planning to test different model implementations that combine our features into a single score of conversational quality. We will also test retraining the feature selection and models after each new rating that was given by a user, or infrequently after a certain amount of new ground-truth ratings have been collected. We believe that updating the models over time is integral to ensure the system accurately reflects conversational quality over time.

## ACKNOWLEDGMENTS

We thank Samantha Ho for the artwork in Fig. 1.

# REFERENCES


[1] Sean Andrist, Dan Bohus, Ece Kamar, and Eric Horvitz. 2017. What went wrong and why? diagnosing situated interaction failures in the wild. In *Social Robotics: 9th International Conference, ICSR 2017, Tsukuba, Japan, November 22-24, 2017, Proceedings 9*. Springer, 293–303.

[2] Salvatore M Anzalone, Sofiane Boucenna, Serena Ivaldi, and Mohamed Chetouani. 2015. Evaluating the engagement with social robots. *International Journal of Social Robotics* 7 (2015), 465–478.

[3] Christoph Bartneck, Tony Belpaeme, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Šabanović. 2020. *Human-robot interaction: An introduction*. Cambridge University Press.

[4] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics* 1 (2009), 71–81.

[5] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. 2018. Social robots for education: A review. *Science robotics* 3, 21 (2018), eaat5954.

[6] Dan Bohus, Sean Andrist, Ashley Feniello, Nick Saw, Mihai Jalobeanu, Patrick Sweeney, Anne Loomis Thompson, and Eric Horvitz. 2021. Platform for Situated Intelligence. arXiv:2103.15975 [cs.AI]

[7] Dan Bohus and Eric Horvitz. 2014. Managing human-robot engagement with forecasts and... um... hesitations. In *Proceedings of the 16th international conference on multimodal interaction*. 2–9.

[8] Okko Buß, Timo Baumann, and David Schlangen. 2010. Collaborating on utterances with a spoken dialogue system using an isu-based approach to incremental dialogue management. In *Proceedings of the SIGDIAL 2010 Conference*. 233–236.

[9] 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*. 1649–1657.

[10] Houwei Cao, Ragini Verma, and Ani Nenkova. 2015. Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech. *Computer speech & language* 29, 1 (2015), 186–202.

[11] Colleen M Carpinella, Alisa B Wyman, Michael A Perez, and Steven J Stroessner. 2017. The robotic social attributes scale (RoSAS) development and validation. In *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction*. ACM, New York, NY, USA, 254–262. https://doi.org/10.1145/2909824.3020208

[12] Amanda Cercas Curry, Helen Hastie, and Verena Rieser. 2017. A review of evaluation techniques for social dialogue systems. In *Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents*. 25–26.

[13] Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review* 54 (2021), 755–810.

[14] Ryuichiro Higashinaka, Luis F D'Haro, Bayan Abu Shawar, Rafael E Banchs, Kotaro Funakoshi, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, and João Sedoc. 2021. Overview of the dialogue breakdown detection challenge 4. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*. Springer, 403–417.

[15] Shanee Honig and Tal Oron-Gilad. 2018. Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in psychology* 9 (2018), 861.

[16] Minjoo Jung, May Jorella S Lazaro, and Myung Hwan Yun. 2021. Evaluation of methodologies and measures on the usability of social robots: A systematic review. *Applied Sciences* 11, 4 (2021), 1388.

[17] Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). 2020. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online.

[18] Maria Kyrarini, Fotios Lygerakis, Akilesh Rajavenkatanarayanan, Christos Sevastopoulos, Harish Ram Nambiappan, Kodur Krishna Chaitanya, Ashwin Ramesh Babu, Joanne Mathew, and Fillia Makedon. 2021. A Survey of Robots in Healthcare. *Technologies* 9, 1 (2021). https://doi.org/10.3390/technologies9010008

[19] Jinying Lin, Zhen Ma, Randy Gomez, Keisuke Nakamura, Bo He, and Guangliang Li. 2020. A review on interactive reinforcement learning from human social feedback. *IEEE Access* 8 (2020), 120757–120765.

[20] Arne Maibaum, Andreas Bischof, Jannis Hergesell, and Benjamin Lipp. 2022. A critique of robotics in health care. *AI & society* (2022), 1–11.

[21] Courtney Mansfield, Sara Ng, Gina-Anne Levow, Richard A. Wright, and Mari Ostendorf. 2021. Revisiting Parity of Human vs. Machine Conversational Speech Transcription. In *Proc. Interspeech 2021*. 1997–2001. https://doi.org/10.21437/Interspeech.2021-1908

[22] Derek McColl, Alexander Hong, Naoaki Hatakeyama, Goldie Nejat, and Beno Benhabib. 2016. A survey of autonomous human affect detection methods for social robots engaged in natural HRI. *Journal of Intelligent & Robotic Systems* 82 (2016), 101–133.

[23] Thilo Michael. 2020. Retico: An incremental framework for spoken dialogue systems. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 49–52.

[24] Giulia Perugia, Maike Paetzel-Prüsmann, Madelene Alanenpää, and Ginevra Castellano. 2021. I can see it in your eyes: Gaze as an implicit cue of uncanniness and task performance in repeated interactions with robots. *Frontiers in Robotics and AI* 8 (2021), 645956.

[25] Aaron Steinfeld, Terrence Fong, David Kaber, Michael Lewis, Jean Scholtz, Alan Schultz, and Michael Goodrich. 2006. Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. ACM, 33–40.

[26] Leimin Tian and Sharon Oviatt. 2021. A taxonomy of social errors in human-robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)* 10, 2 (2021), 1–32.

[27] Thurid Vogt and Elisabeth André. 2005. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *In Proceedings of the 2005 IEEE International Conference on Multimedia and Expo*. IEEE, IEEE, 474–477. https://doi.org/10.1109/ICME.2005.1521463

[28] Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics (ACL)*. Association for Computational Linguistics, Madrid, Spain, 271–280. https://doi.org/10.3115/976909.979652

[29] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, et al. 2022. A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion* 83 (2022), 19–52.

[30] Rosemarie E Yagoda and Douglas J Gillan. 2012. You want me to trust a ROBOT? The development of a human–robot interaction trust scale. *International Journal of Social Robotics* 4 (2012), 235–248.

[31] James E Young, JaYoung Sung, Amy Voida, Ehud Sharlin, Takeo Igarashi, Henrik I Christensen, and Rebecca E Grinter. 2011. Evaluating human-robot interaction: Focusing on the holistic interaction experience. *International Journal of Social Robotics* 3 (2011), 53–67.