A Systematic Cross-Corpus Analysis of Human Reactions to Robot Conversational Failures

Dimosthenis Kontogiorgos*†
diko@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

Minh Tran*
minhntra@usc.edu
University of Southern California
Los Angeles, USA

Joakim Gustafson
jkgu@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

Mohammad Soleymani
soleymani@ict.usc.edu
University of Southern California
Los Angeles, USA

ABSTRACT

In this paper, we analyze multimodal behavioral responses to robot failures across different tasks. Two multimodal datasets are examined in which humans interact with guided-task robots in task-oriented dialogues. In both datasets, the robots simulated failures of conversational breakdown and miscommunication typically observed in human-robot interactions. We closely examine human reactions to these failures looking at facial and acoustic features. Our analyses identify the significant behavioral features for automatic detection of such failures in interaction. We also examine human responses to different types of robot failures and if failures occurred early or late in the interaction cause variation in the responses. Our findings indicate that several nonverbal behaviors are consistently present in responses to robots' failures, e.g., gaze and speech prosody, whereas, linguistic features appear to be task-dependent. We discuss how these findings may generalize to other tasks, and how autonomous robots may identify opportunities to detect and recover from failures in interactions with humans.

CCS CONCEPTS

• **Human-centered computing → Natural language interfaces**.

KEYWORDS

social signal processing, behavioral responses, miscommunication

ACM Reference Format:
Dimosthenis Kontogiorgos, Minh Tran, Joakim Gustafson, and Mohammad Soleymani. 2021. A Systematic Cross-Corpus Analysis of Human Reactions to Robot Conversational Failures. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3462244.3479887

*Both authors contributed equally to this research.
†Work conducted during a research visit at University of Southern California.

ICMI '21, October 18–22, 2021, Montréal, QC, Canada

ACM ISBN 978-1-4503-8481-0/21/10…$15.00
https://doi.org/10.1145/3462244.3479887

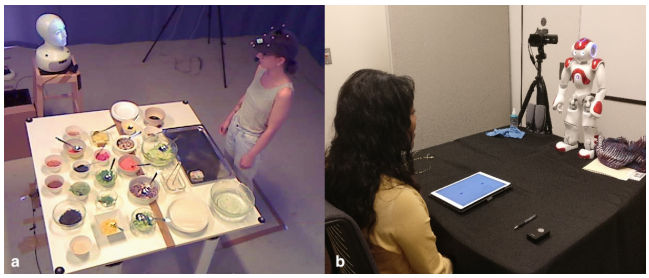

Figure 1: Two datasets were used in this paper on the examination of failure behaviors from robots: a) In the *Instruction* corpus, participants are guided by robots to complete a cooking task, b) In the *Negotiation* corpus, participants need to negotiate with a robot on a decision-making task.

1 INTRODUCTION

For many of us, a large part of our everyday life consists of interacting with technology in some form. Computing machinery has evolved from punch-card programming to interactions through voice, the most familiar mode of communication for humans. While machine failures have been part of our interactions with technology, there is something fundamentally different once voice is the main interface, as failures need to be resolved via dialogue. Humans use various channels to detect and mitigate miscommunication, and adhering to social protocols in failure mitigation is essential for an effective conversational interface.

State-of-the-art conversational systems do not incorporate sophisticated failure detection and failure recovery mechanisms, as they often conduct transactional interactions [27], with no opportunities to resolve failures. While computers are considered to be social actors [30], it is unknown whether responses to miscommunication with machines are similar to the ones among humans. It also remains unexplored whether responses to conversational failures are task-dependent, *i.e.,* do multimodal failure behaviors differ in open-ended dialogues, task-oriented interactions, or within different tasks with socially assistive robots? From the computational point of view, adapting to failures in different domains is challenging. However, utilizing online failure detection models brings opportunities to convey adaptive robot behaviors and affords interactions beyond single-turn information retrieval tasks. Such models will inevitably become more advantageous in the future, at the service of coherent human-robot interactions.

Table 1: Multimodal channels used in HRI literature on robot failure analysis. Work that does not analyze behavioral signals but focuses on robot perception is not included. Work that has conducted failure detection indicated in bold.

Paper	Body/Motion features	Head/Eye movement	Facial features	Speech/Acoustic features
Gehle et al. (2015) [17]	✓	✓	✓	
Giuliani et al. (2015) [18]	✓	✓	✓	✓
Hayes et al. (2016) [20]	✓	✓	✓	
***Andrist et al. (2017)** [1]*			✓	✓
***Trung et al. (2017)** [41]*	✓	✓	✓	
***Short et al. (2018)** [37]*	✓			✓
Cahya et al. (2019) [10]		✓	✓	
Flook et al. (2019) [16]	✓			
Aneja et al. (2020) [2]			✓	✓
***Kontogiorgos et al. (2020)** [25]*	✓	✓		✓

In this work, we investigate reactions to failures in two datasets in which robots make deliberate conversational mistakes (Figure 1). We analyze how individuals respond to non-anticipated violations of social protocols of interaction [9] (*i.e.,* not responding, providing incorrect answers, repeating oneself several times). We utilize automatically tracked behaviors to understand human behavioral responses to conversational failures, with the goal of their automatic detection. Such an automated failure detection system can inform the machine to recover from a conversational failure.

We build a machine learning model to identify the instances of failures in the interaction using multimodal behaviors, *i.e*, linguistic, facial and acoustic features. We demonstrate that despite the variations in reactions to failures, their elicited responses are detectable. An examination across datasets shows that certain features may be task-dependent. We find consistent reactions to failures regardless of when they happen, *i.e.*, early on or late in the interaction. Additionally, we conduct a within-corpus and between-corpus evaluation to further examine the consistency of reactions to failures.

We hypothesize that individuals react to robot failures with an intensity that makes them detectable, meaning that certain facial, linguistic or acoustic features are activated regardless of the task and context. With the aforementioned methodology in failure and corpus analysis, this article contributes to literature in multimodal human-robot interaction by answering the following questions: **RQ 1:** What are the changes in linguistic, facial and acoustic signals from humans in different displays of robot failures across different task contexts? **RQ 2:** How well can a robot detect whether it has violated social protocols of interaction through a failure with the observed behaviors? **RQ 3:** Do reactions through multimodal signals shift in different types of failures or throughout the interaction?

2 RELATED WORK

This article is related to studies examining miscommunication and robot failures in situated contexts. Prior work in dialogue systems has examined human reactions to dialogue breakdowns through analyzing verbal and acoustic features [19, 42]. The focus of this work is on multimodal embodied dialogue systems, typically in the form of a social robot, and human multimodal behaviors during conversations. Past HRI work has used a variety of multimodal signals for behavioral analysis that investigates reactions to failures in different contexts (Table 1). For a comprehensive review on failure taxonomies, mitigation, and reactions in HRI we refer the reader to the work of Honig and Oron-Gilad [22].

The exchange of turns in human-robot dialogue can be viewed as adjacency pairs - once the first turn is on the floor, the next turn is conditionally relevant [35]. This creates an expectation of socially acceptable exchange of turns, once voice is the main interface of interaction, that contributes to understanding among the interaction partners [12, 40]. Violations to these expectations, such as not responding, may result in reactions in human speech [15, 23, 34]. In interactions among humans, individuals typically detect quickly when the conversational partner does not understand and adjust their message construction to the recipient (known as recipient design). Misunderstanding is also a common type of failure in conversational interfaces [8], that can be detected early on at the speech recognition state [21, 28, 39]. Repetitions or explicit responses, *e.g.*, uttering "I do not understand", also reflect misunderstanding in user input [29]. When misunderstanding is detected, failure recovery strategies need to be applied, a functionality that the state-of-the-art systems have a limited capacity for [7, 38].

Humans tend to have a number of predictable behaviors in response to failures [22], including variations in gaze and head motion [17, 20, 25, 41], facial expressions [10, 20, 41], body movements [1, 16, 37, 41], and speech or acoustic features [2, 18, 25, 37]. Some of these studies have approached failures as a detection problem, often in real time, when anomalies are detected in social contingent signals in robot behaviors. Some research has also used audiovisual features to detect failures by measuring changes in signals. Shi *et al.* [36] for example modeled environmental changes in a robot distributing flyers to pedestrians to detect interactional failures. Short et al [37] used acoustic and video features to detect deviations from contingent expected human responses to robot requests. Grounding sequences or implicit behavioral responses to failures have also been examined in investigations of robot failures [24], including unexpected responses in human-robot interactions in-the-wild [1].

Overall, robot failures have been classified into technical failures and social failures (or interaction failures), shortcomings either caused by technical issues or inappropriate responses that violate social protocols of interaction [22]. Failures vary in nature, including not completing requested tasks, not responding or timing speech improperly, giving incorrect answers, repeating statements, or producing unexpected erratic behaviors. Many of these failures are present in the datasets used in this paper, representing some of the most commonly reported robot failures in HRI literature so far.

Table 2: Summary of the two robot failure datasets.

Data	Instruction	Negotiation
Number of interactions	88	103
Number of segments	824	1939
Number of failures/no-failures	382/442	879/1061
Total duration of all segments	103.9 min.	127.1 min.
Total duration failure/no-failure	60.5/43.4 min.	34.8/92.3 min.
Avg. duration of all segments	7.5 sec.	3.9 sec.
Avg. duration failure/no-failure	9.4/5.6 sec.	2.5/5.5 sec.
Avg. # user turns	5.7	14.3
Avg. # words per segment	6.0	10.5
Avg. # words failure/no-failure	7.3/4.4	7.5/14.0

3 DATA

The aim of this work is to understand the verbal and nonverbal behaviors in responses to failures in human-robot interactions. We perform an analysis on two robot failure datasets to study multimodal behaviors in response to miscommunication and dialogue breakdown. Both datasets contain dialogues of collaborative interactions between humans and robots on different tasks, namely cooking and item-ranking. We compare reactions to failures with linear mixed-effects models (LMM) and train machine-learning models to detect the occurrence of failures. Both datasets are labeled at human utterance level. Additionally, both datasets are designed to induce robot failures in particular times in the interaction, and therefore labels are extracted by experiment design. To simulate an autonomous robot system, the utterances are automatically extracted using the IBM Watson Speech-To-Text service. During the human utterances, we extract verbal and nonverbal behavioral cues to represent human responses to robot failures.

3.1 Instruction Corpus [Ins]

This dataset was collected in a study in which human users were instructed by robots to cook various recipes [26]. The corpus contains 88 interactions between human users and robot instructors. Participants were brought in a room with a table containing ingredients and a cooking preparation area, and a robot was placed next to the table to guide them in cooking. The dialogue was driven by participants asking the robot instructor which ingredient to pick next until the recipe is completed, and designed to examine grounding behaviors in robot instructions. Each participant (N=44) interacted twice with two different types of robots (a Furhat robot and an Amazon Echo), preparing different types of spring roll recipes. Participants did not know the recipe so they needed the robot's instructions in order to complete the cooking task.

Failures were designed to happen in various moments in the interaction including: delayed response to a participant (DL, 77 instances); giving incorrect guidance (WR, 102 instances); repeating previous instructions (RP, 68 instances); restarting the interaction (RS, 72 instances); and timing speech improperly (II, 63 instances). Failures happened in predetermined moments in the interaction, from which we derived the failure and no-failure instances. Participants' behaviors were captured by a front-facing camera (at 30fps) and a microphone, placed on the robot (Figure 1a). Dialogues lasted from 4 to 10 minutes (Table 2), and a human wizard was operating

```
USER : [FINISHED-ACTION] Yeah! I'm done.
ROBOT: [TRIGGERS-NORESPONSE-FAILURE]
USER : [WAITS-FOR-RESPONSE]
USER : [LOOKS-AT-ROBOT] I'm done!
USER : [FROWNS]
USER : [CLARIFICATION] Hello? I'm done!
USER : [QUESTION] What should I do next?
USER : [LOOKS-AT-TIMER]
ROBOT: [INSTRUCT] Take one spoon of red cabbage
       and put it in the spring roll.
```

Figure 2: Dialogue excerpt from the Instruction dataset.

```
USER : What's third?
ROBOT: [WRONG ANSWER] I ranked the oxygen first.
USER : Okay. Which one is third?
ROBOT: I ranked the stellar map third.
USER : Why?
ROBOT: I ranked the stellar map third because we will need it
for navigation when we walk.
USER : Okay. What's next?
ROBOT: [SILENCE]
USER : Which one do you want to do next?
ROBOT: I ranked the food concentrate fourth.
USER : Okay. Why fourth?
ROBOT: [REQUEST_REPETITION] I'm sorry, I didn't hear you.
```

Figure 3: Dialogue excerpt from the Negotiation dataset.

the robots according to a protocol indicating when robots should fail and what should it respond to in accordance with human input. A dialogue sample can be seen in Figure 2.

3.2 Negotiation Corpus [Neg]

This dataset is extracted from the *Niki and Julie* corpus [3]. The corpus contains 103 dialogues of interactions between human participants and a Nao robot, engaged in a collaborative item-ranking task designed to measure robot's influence on humans. For each interaction, participants were brought to a room and sat in front of their robot partner, named Niki, with an iPad to help them in their item-ranking task. Participants were asked to participate in one of the two classical survival ranking tasks, in which they have to rank the importance of items for survival in a hostile environment (desert or moon) or rank the importance of art pieces to save when a fire is approaching. To complete their tasks, participants needed to collaborate and negotiate ranking with the robot.

During the interaction, a human Wizard guided the robot's responses and behaviors. By study design, the robot made deliberate conversational failures. The failures included asking the participant to repeat themselves (AR, 468 instances); giving a wrong answer to a participant's question (WR, 142 instances); ignoring the participant when a response is expected (IG, 133 instances); repeating while interrupting the participant (RP, 46 instances); making an irrelevant or unintelligible remark (IR, 90 instances). Data was collected from a camera facing the participants (at 60fps) and placed next to the robot and a microphone (Figure 1b). To match the Instruction Robot Failure dataset, which has a balanced number of Failure and No-Failure segments, we sub-sampled the No-Failure segments in the Negotiation dataset such that it becomes balanced.

Each dialogue lasted from 2 minutes to 15 minutes (Table 2). Figure 3 shows a sample dialogue from the dataset.

3.3 Interaction Paradigms

In both datasets conversational failures occurred by design; while the reported failures represent common causes of dialogue breakdown in human-robot interactions, it is unclear whether subjects exposed to failures are aware that these failures were intentional and how detrimental they are to the task completion in each dataset. Additionally, both datasets represent task-oriented interactions. The Instruction corpus falls into the category of joint construction tasks, in which a human and a robot collaborate to assemble a structure, in this case a cooking recipe. The constrained nature of the task can be decomposed into information retrieval, reference identification, and assembly, with little questioning over the robot's knowledge. The Negotiation corpus however is meant to create discussion and question the decisions taken by both parties, in order to measure the influence each agent makes to the other. We therefore expected the grounding sequences to differ among corpora, on how subjects react to resolve failures, in observed verbal and nonverbal channels.

4 AUTOMATIC BEHAVIOR EXTRACTION

4.1 Feature Extraction

Using automatic speech recognition (ASR) we collected participant responses automatically and extracted features that represent verbal and nonverbal behaviors. In each segment, data is captured during participants' spoken utterances with timings derived from the ASR output. This system simulates a real-time input of user behavior to infer if the robot has conducted itself appropriately [NF] or if a dialogue breakdown has occurred [F], ensuring that all features can also be extracted in real-time. In the segments, the following lexical, visual and acoustic features are mean-aggregated during a given utterance and used for the statistical analysis and the failure detection machine learning models.

Lexical. We use the pre-trained Sentence-BERT model [33] to convert each utterance into a vector for the classification models. Sentence-BERT is a variation of the established BERT model [13], which utilizes Siamese architecture with triplet loss to produce semantically meaningful sentence embeddings. For statistical analysis, we use features extracted from the dictionary-based Linguistic Inquiry and Word Count (LIWC) toolkit on psychological processes such as tone, affect, positive and negative emotions [31].

Visual. We use the OpenFace toolkit [4], to extract visual features from the videos. We extract the intensities of all 17 supported facial action units (FAUs) along with eye gaze (x and y angles) and head pose (position, roll, yaw and pitch) features from the OpenFace outputs. We compute the mean and standard deviation of these features for the statistical analyses and classification models.

Acoustic. The OpenSMILE toolkit is used to extract acoustic features from the collected audio [14]. We use a set of features from *eGeMAPS* (*i.e.*,, the extended Geneva Minimalistic Acoustic Parameter Set) and Mel-frequency Cepstral Coefficients (MFCC) for the classification models. The extended set of GeMAP consists of 23 features characterizing voice production including fundamental frequency, loudness and formants, while the MFCC with the first and second order derivatives generate a vector of 39 features per utterance. We also compute the mean and standard deviation of these features for statistical analyses and the classification models.

4.2 Behavioral Analysis

Statistical Analysis. Statistical analyses on the behavioral measures are conducted in R [32], using the aforementioned lexical, visual and acoustic features as predictors. The *lme4* package [5] is used to run linear mixed-effects models on the relationship between robot failures (as a fixed-factor with two levels) and the independent variables. Analyses are conducted independently per dataset, and as random effects we have intercepts for participants in each corpus. Chi-square tests and p-values are derived from maximum likelihood estimation tests of the full models against intercept-only models. Pairwise comparisons are conducted with the *emmeans* package with Tukey post-hoc tests. We also run LMM with the order of the utterances in the interaction as a random factor, to determine if any of the behavioral predictors are associated with how early or late failures appear in the interaction. While we believe this to be appropriate to capture the variance in features across time in the interaction, we also conduct a correlation analysis to determine if there is a linear relationship between any of the features and the utterance index in each moment in the interaction.

Failure Classification. Both datasets are relatively small (<2000 samples). A bidirectional LSTM followed by a fully connected layer showed promising performance on temporal features in both datasets. However, best performance was achieved with mean-aggregated features per utterance and Gradient Boosted Tree (XGBoost) classifiers [11], which we report in the remaining of this paper. We choose XGBoost for classification due to its robustness to overfitting and effectiveness when handling imbalanced datasets. Features are concatenated per modality in early-fusion and the binary choice classifiers are evaluated in subject-independent 10-fold cross-validation. To evaluate the predictions, we report mean accuracy of all folds, within- and cross-corpus, and also perform multi-class classifications to determine the consistency of behaviors in each failure. We hypothesize that certain failures may cause consistent reactions, depending on how detrimental they may be to the interaction. We finally calculate the F-scores (determined by the F-statistic) to examine which features are most informative in failure detection, measuring the ratio of explained and unexplained feature variance in the construction of the gradient boosted trees.

5 RESULTS

5.1 Statistical Analysis

Lexical. Table 3 presents the results from the LMM analysis on the lexical features among corpora ([Ins] & [Neg]) and classes ([NF] & [F]). Statistical significance is obtained by maximum likelihood ratio tests, with the fixed-effect models tested against the null model. A number of cognitive and affective processes such as Tone, Affect, Positive and Negative emotion are associated with how humans respond to robot failures in lexical features, regardless of dataset, indicating a shift in language toward negative tone and emotion. The number of words spoken in the Instruction dataset is lower in no-failure (4.41 ± 0.53) than in failures (7.30 ± 0.53) with a statistically significant difference: $\chi^2 = 46.56, p < .001$. Accordingly, utterance duration is lower in no-failure (5.61 ± 0.55) than in failures

Table 3: LMM on lexical features (LIWC) on the Instruction corpus [Ins] and the Negotiation corpus [Neg]. RF. Variance represents the random intercept variance (participant). P-value indicators: $*p \leq 0.05, **p \leq 0.01, ***p \leq 0.001$.

Predictor	Mean NF	Mean F	RF. Variance	Coef. Estimate	t-value	chi-square	p-value
Ins - Tone	46.30 ± 2.01	40.40 ± 2.09	59.81	−5.83 ± 2.31	-2.52	6.32	*
Ins - Affect	13.29 ± 1.24	8.31 ± 1.30	16.38	−4.97 ± 1.54	-3.21	10.25	**
Ins - Posemo	12.11 ± 1.25	6.46 ± 1.31	20.27	−5.65 ± 1.49	-3.77	14.05	***
Ins - Negemo	1.21 ± 0.35	1.85 ± 0.37	0.87	0.64 ± 0.46	1.37	1.89	
Neg - Tone	45.60 ± 1.40	32.70 ± 1.42	62.79	−12.94 ± 1.66	-7.78	59.21	***
Neg - Affect	13.33 ± 0.77	5.63 ± 0.79	16.17	−7.69 ± 0.95	-8.04	60.80	***
Neg - Posemo	9.48 ± 0.65	3.30 ± 0.67	10.69	−6.17 ± 0.82	-7.47	53.57	***
Neg - Negemo	3.86 ± 0.42	2.28 ± 0.43	3.68	−1.58 ± 0.54	-2.92	8.33	**

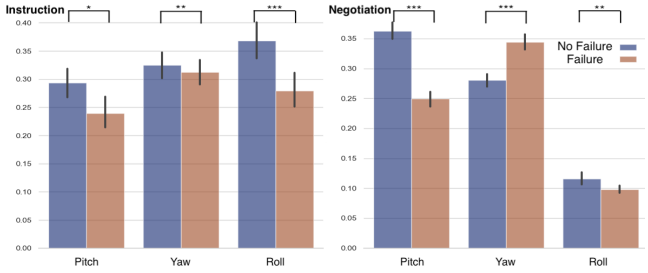

Figure 4: Normalized pose estimation per dataset. Values closer to 0 indicate gaze at the robot (where the cameras are placed). In both datasets (Yaw in Instruction, Pitch in Negotiation), users tend to gaze at the robot when failures occur.

(9.44 ± 0.57): $\chi^2 = 37.14, p < .001$. On the contrary, the number of words spoken in the Negotiation dataset is higher in no-failure (14.09 ± 1.15) than in failures (7.57 ± 1.17): $\chi^2 = 55.83, p < .001$, and utterance duration is higher in no-failure (5.51 ± 0.47) than in failures (2.57 ± 0.48): $\chi^2 = 61.52, p < .001$. This may indicate that utterance construction post-failure may be task-dependent.

Visual. Concerning visual features, linear mixed-effects models show significant differences in gaze behavior with users consistently gazing at the robot after a failure (in both tasks). LMM results are reported in Table 4 and Figures 4 and 5. We also find that in instruction tasks, users expressed higher intensity in smiling (AU06 and AU12) in failures; on the contrary in negotiation tasks, users smiled more when there were no failures. In both tasks, AU10 (lip raiser) intensity is higher in failures, potentially indicating exaggerated articulation of vowels during post-failure utterance construction. The same effect can be seen in AU25 (lips part) for the Instruction corpus and AU26 (jaw drop) for the Negotiation corpus, both associated with vowel hyper-articulation. The increase in intensity of AU04 (brow lowerer) for Negotiation and decrease in AU02 (outer brow raiser) for Instruction, related to surprise, are also associated with failures. This is in accordance with the expected behavior that people may be surprised when robots fail.

Acoustic. We find statistical differences in reactions to failures in acoustic features, in particular loudness appears to be higher in both datasets in utterance construction after a robot failure. Spectral flux, associated with the timbre of the acoustic signal is also higher in failures in both datasets. On the contrary, F0, representing pitch, appears to be lower in failures in Instruction tasks and higher in Negotiation tasks. Table 5 presents the linear mixed-effects model analysis in acoustic features.

5.2 Failure Detection

Table 6 presents the prediction accuracy of the two-class failure classification (failure vs. no-failure) and multiclass classification (5 types of failures and no-failure) for both datasets. It is important to note that the datasets do not contain the same failures even though they both have 5 types of robot failures (some of them are similar). We choose hyper-parameters for the XGBoost models using the grid-search cross-validation method. We also use balanced accuracy as the evaluation metric in Table 6 due to class imbalance on both datasets for multiclass classification. As a result, the random guessing baseline is 50% for binary classification and 16.7% for multiclass. Additionally, for multiclass classification on both datasets, we use the SMOTEENN [6] algorithm to resample the training sets and reduce the effects of class imbalance.

Results on both datasets and classification tasks (binary and multiclass) demonstrate the significance of lexical features in failure detection. Multimodal models achieve the best performance suggesting that the combination of verbal and nonverbal behaviors contain valuable information in failure classification. For binary classification, the models yield good results with an accuracy of 75.09% on Instruction data and 77.37% on Negotiation data. However, identifying the specific type of robot failure is a relatively challenging task. Although the models' prediction performances (59.10% for Instruction and 34.43% for Negotiation) are far better than random, the results demonstrate the challenges of classifying different types of miscommunication with robots (Figure 6).

We finally report the models' performance with cross-dataset validation. $A \rightarrow B$ represents training on dataset A and testing on dataset B. We can see that the classifiers' cross-dataset performances are close to chance level, which indicates the challenges in transferring failure reaction representations across datasets.

Feature Analysis We first examine the variance caused by the utterance sequence in each interaction for all of the predictors in the linear mixed-effects models performed. Most parameters do not indicate any significant deviations of increased variance caused by

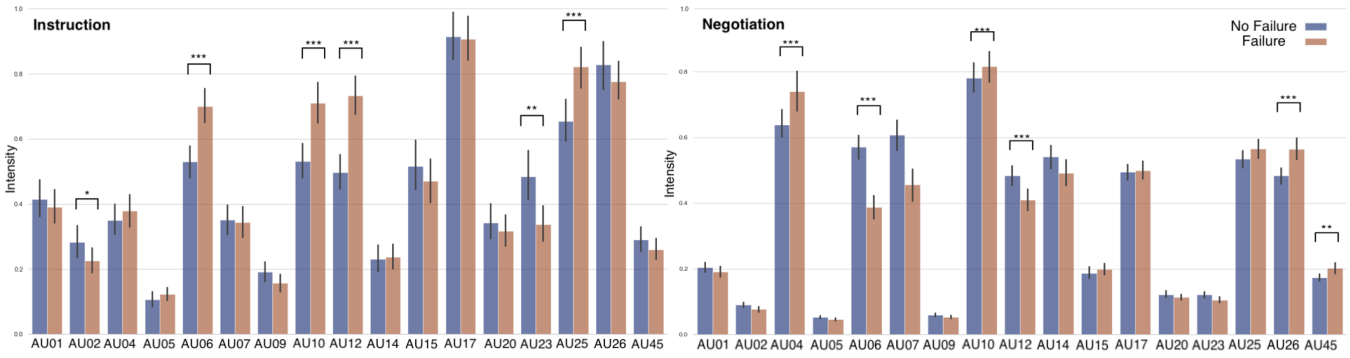

Figure 5: Mean Facial Action Unit activation intensity. Instruction dataset on the left, and Negotiation dataset on the right.

Table 4: LMM on visual features (head and AU) on the Instruction corpus [Ins] and the Negotiation corpus [Neg]. Only predictors with significant differences are reported with P-value indicators: $*p \leq 0.05, **p \leq 0.01, ***p \leq 0.001$.

Predictor	Mean NF	Mean F	RF. Variance	Coef. Estimate	t-value	chi-square	p-value
Ins - Pitch	0.20 ± 0.02	0.15 ± 0.02	0.00	0.04 ± 0.02	-1.99	3.95	*
Ins - Yaw	−0.17 ± 0.02	−0.23 ± 0.02	0.01	−0.06 ± 0.02	-2.71	7.29	**
Ins - Roll	0.33 ± 0.02	0.23 ± 0.02	0.01	−0.10 ± 0.02	-4.05	16.28	***
Ins - AU02 Outer brow raiser	0.28 ± 0.02	0.22 ± 0.02	0.00	−0.06 ± 0.03	-1.99	3.94	*
Ins - AU06 Cheek raiser	0.52 ± 0.03	0.69 ± 0.03	0.03	0.16 ± 0.03	4.74	22.19	***
Ins - AU10 Upper lip raiser	0.55 ± 0.05	0.70 ± 0.05	0.07	0.14 ± 0.03	3.94	15.39	***
Ins - AU12 Lip corner puller	0.50 ± 0.04	0.71 ± 0.04	0.05	0.21 ± 0.03	5.79	32.76	***
Ins - AU23 Lip tightener	0.48 ± 0.03	0.33 ± 0.03	0.02	−0.14 ± 0.04	-3.19	10.16	**
Ins - AU25 Lips part	0.65 ± 0.04	0.82 ± 0.03	0.02	0.17 ± 0.04	3.79	14.23	***
Neg - Pitch	0.33 ± 0.01	0.25 ± 0.01	0.01	−0.08 ± 0.01	-7.18	49.90	***
Neg - Yaw	0.30 ± 0.01	0.33 ± 0.01	0.02	0.03 ± 0.00	3.94	15.38	***
Neg - Roll	−0.03 ± 0.01	0.00 ± 0.01	0.00	0.02 ± 0.00	3.06	9.37	**
Neg - AU04 Brow lowerer	0.77 ± 0.06	0.89 ± 0.06	0.41	0.11 ± 0.02	4.45	19.71	***
Neg - AU06 Cheek raiser	0.45 ± 0.04	0.37 ± 0.04	0.15	−0.08 ± 0.02	-3.40	11.49	***
Neg - AU10 Upper lip raiser	0.72 ± 0.04	0.84 ± 0.04	0.18	0.12 ± 0.03	4.07	16.46	***
Neg - AU12 Lip corner puller	0.47 ± 0.03	0.38 ± 0.03	0.09	−0.08 ± 0.02	-3.62	13.08	***
Neg - AU26 Jaw Drop	0.47 ± 0.02	0.56 ± 0.02	0.02	0.08 ± 0.02	3.75	14.04	***
Neg - AU45 Blink	0.16 ± 0.00	0.19 ± 0.00	0.00	0.03 ± 0.01	2.99	8.93	**

Table 5: LMM on acoustic features on Instruction [Ins] and Negotiation [Neg]. P-values: $*p \leq 0.05, **p \leq 0.01, ***p \leq 0.001$.

Predictor	Mean NF	Mean F	RF. Variance	Coef. Estimate	t-value	chi-square	p-value
Ins - Loudness	0.51 ± 0.02	0.59 ± 0.02	0.00	0.08 ± 0.02	3.13	9.74	**
Ins - SpectralFlux	0.42 ± 0.02	0.55 ± 0.02	0.01	0.13 ± 0.03	4.03	16.09	***
Ins - F0	17.70 ± 0.51	14.70 ± 0.53	4.56	−2.92 ± 0.57	-5.11	25.58	***
Ins - Jitter	0.02 ± 0.00	0.02 ± 0.00	0.00	0.00 ± 0.00	-1.87	3.48	
Ins - Shimmer	0.82 ± 0.02	0.69 ± 0.02	0.00	−0.13 ± 0.02	-4.74	22.14	***
Neg - Loudness	0.42 ± 0.04	0.48 ± 0.04	0.16	0.05 ± 0.01	3.95	5.52	***
Neg - SpectralFlux	0.27 ± 0.04	0.37 ± 0.04	0.17	0.09 ± 0.01	5.31	28.01	***
Neg - F0	17.70 ± 0.38	18.50 ± 0.38	11.38	0.72 ± 0.26	2.73	7.23	**
Neg - Jitter	0.01 ± 0.00	0.01 ± 0.00	0.00	0.00 ± 0.00	-0.88	0.77	
Neg - Shimmer	0.69 ± 0.01	0.70 ± 0.01	0.02	0.01 ± 0.01	1.45	2.10	

the order of the spoken utterances. We further use the parametric Pearson correlation coefficient to determine if there is a linear association between any of the predictors with the index of spoken utterances (indicating their order in the interaction). This is done

Table 6: Contribution of modalities in failure classification based on gradient boosted trees. Values in bold indicate highest estimation accuracy per modality (A: acoustic, V: visual, L: lexical). Binary classification baseline is 50% and multiclass is 16.7%.

Corpus	Type	A	V	L	A+V	V+L	A+L	A+V+L
Instruction	binary	68.44 ± 1.27	**64.25** ± 0.10	69.30 ± 0.94	71.35 ± 0.93	71.58 ± 1.47	73.40 ± 0.76	75.09 ± 1.23
Negotiation	binary	**70.78** ± 0.46	62.83 ± 0.49	**71.91** ± 0.71	**72.40** ± 0.43	**73.84** ± 1.39	**76.99** ± 1.08	**77.37** ± 0.55
Instruction	multiclass	**46.42** ± 1.74	**34.09** ± 1.85	**48.19** ± 1.26	**49.99** ± 2.93	**52.75** ± 1.94	**54.97** ± 1.64	**59.10** ± 0.89
Negotiation	multiclass	24.00 ± 0.75	20.78 ± 2.77	32.52 ± 1.52	27.63 ± 0.43	33.13 ± 1.93	33.45 ± 2.42	34.43 ± 1.59
INS → NEG	binary	**52.60** ± 1.16	**49.90** ± 0.87	**53.56** ± 2.10	**53.32** ± 1.81	**53.22** ± 1.34	**56.27** ± 1.40	**55.68** ± 1.44
NEG → INS	binary	46.47 ± 1.57	40.72 ± 1.07	52.57 ± 0.96	47.01 ± 1.85	50.72 ± 0.81	52.38 ± 1.42	49.56 ± 1.31

Figure 6: Confusion matrix for multiclass classification.

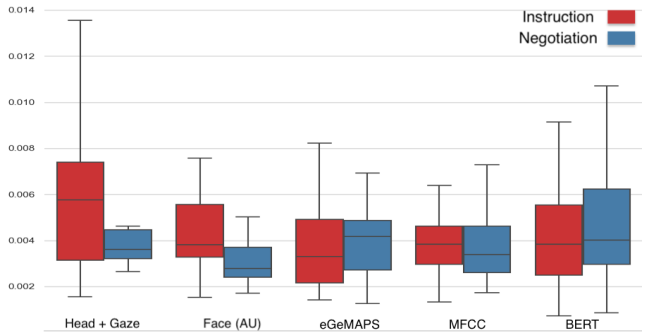

Figure 7: Individual feature F-score (importance) grouped by modality. A higher variance can be observed in visual features in the instruction corpus.

to examine if verbal and non-verbal responses to failures change over time. A Bonferroni correction is applied to adjust significance for multiple comparisons, and only coefficients larger than 0.15 are reported. Very few behaviors seem to be associated with how early or late failures appear in the interaction, and it is unclear whether the changes are due to failures or due to the task. For the Instruction corpus, utterance duration is inversely correlated with the index of spoken utterances (r=-.19, p<.001). Gaze behaviors are also inversely associated with the index of spoken utterances (Head Pitch: r=-.15, p<.001, Head Yaw: r=-.19, p<.001), as users look more towards the robot as the interaction progresses and more failures appear. On the Negotiation corpus, no significant associations are found in reactions to failures over time.

We also examine in a post-hoc analysis what features are activated (importance determined by F-scores) in the construction of gradient boosted trees, trained in early fusion with all features concatenated. We observe that the visual features are more prominent in the Instruction corpus when detecting states of failure, while the lexical and acoustic features perform equally well in both datasets (Figure 7). This variance can be explained by how the task is composed in instructions, with embodied contributions to the ground when performing robot requests ('take some cucumber from the table'). Less variance in head movements may be expected in negotiation grounding sequences ('why did you choose the map'). Additionally, a correlation in feature importance across datasets yields no significant results (r=.03, p=.60), further indicating that the models for each dataset are inherently different.

6 DISCUSSION

In this work, we adopt a predictive approach in modeling miscommunication with robots in the form of failures, and conduct a statistical analysis to observe reactions in different verbal and nonverbal behaviors. We find that certain modes of communication in acoustic and visual signals tend to be common reactions to failures across contexts, while some utterance construction features may be dependent on the collaborative task between humans and robots. We demonstrate that there are distinct multimodal signals in reactions to failures. We present a systematic analysis across datasets, along with analyses in different failure types. In this section, we present the main findings and lessons learned from the challenges working across datasets in human-robot interaction failures.

RQ 1: Human Reactions to Robot Failures

The results suggest that several nonverbal responses to failures may generalize across contexts. However, utterances in response to failures are different among the two tasks. In Instruction tasks, users may not need to talk in cases of no-failure and instead demonstrate positive understanding through actions in response to robot requests. Failures in referential communication may need to be resolved via dialogue, and therefore longer utterances may be spoken. Conversely, in Negotiation tasks, users tend to discuss the decisions taken in cases of no-failure in longer utterances and instead 'freeze' when the robot seems to be non-responsive [18]. These subtle differences in how individuals produce utterances in response to failures may indicate that grounding behaviors in dialogue breakdown also depend on dialogue composition. This is further illustrated in behaviors important for failure detection, as a larger part of variance is explained by visual features at the Instruction dataset, where users contribute to common ground with embodied actions. Smaller variance in visual features in the Negotiation dataset may also be explained by the use of a screen (iPad) to do the task.

In attempts to generalize the findings to other contexts, we find that loudness and spectral features are significantly higher in utterances produced after failures, indicating that users put additional

effort to resolve miscommunication with robots. At the same time, as expected, utterances produced post-failure indicate negative tone and emotion in terms of their linguistic structure. We also encounter similar gaze behavior across datasets, with users gazing at the robot when a failure has occurred. This behavior is also not surprising and in agreement with findings from prior robot failure research [10, 18, 25, 41], which we confirm in two separate datasets, as users may need to attend to any subtle robot signal to resolve miscommunication. We nevertheless find contradictory behaviors in positive expressions such as smiling (AU06 & AU12), further challenging cross-corpus evaluation, which may be explained by the nature of the task or whether users are aware that failures are intentional. Apart from similarities in loudness, we also observe higher activation intensity of facial action units associated with expressions influenced by phonetic content, further indicating that users spend significantly more effort in pronouncing their utterances in failures.

RQ 2: Failure Detection

The results from within-corpus experiments show the feasibility of recognizing a robot's failure by human behavioral signals with approximately 75% and 77% accuracy. The method we employed is fully automated with an utterance-level analysis allowing for near real-time use. However, the model learned in one task is not as effective when tested on a new task, potentially due to the biased linguistic representations of each task. There are several other factors that may have impacted the performance of the cross-corpus classification models. First, both datasets have separate types of failures. 'Wrong answer/ingredient', 'repeat previous utterance/instruction' and 'delayed response/silence' could be considered similar across the datasets, however in different settings. Second, the tasks in the two datasets are also different. As a result, the reactions in these two datasets might not be similar, further indicating challenges of failure detection in real-world settings. One of the challenges is the language used by participants in the tasks (cooking vs. ranking), forming different semantic representations across tasks. Consequently, the classification models might learn different behaviors associated with failure reactions on one dataset and perform poorly on predictions on other datasets. The feature analysis confirms that the models for each task are substantially different, as separate features are activated to identify and detect failures in each dataset.

RQ 3: Robot Failure Types

Recognition of reactions to failures yields an accuracy far better than chance (59.10% on [INS] and 34.43% on [NEG]). In Figure 6, one can see that there is a significant drop in the classification accuracy of the *No failure (NF)* class in comparison with the binary classification task, which is around 75% to 77%. This is mainly an effect of using the SMOTEENN algorithm [6], which combines over-sampling and under-sampling to form balanced training sets. However, the re-sampling algorithm helps to improve the overall accuracy by approximately 6% on the Negotiation dataset and 10% on the Instruction dataset. *Repetition (RP)* is a shared type of failure between the 2 datasets that the classification models perform poorly on both datasets. The models classify *Repetition* failures correctly 31% of the time on the Instruction dataset and 13% of the time on the Negotiation dataset (worse than chance), and worse than the accuracy of around 70% of the other 4 failure types. Inspecting the datasets, we find that participants in both studies do not show varied reactions when *Repetition* occurs. Specifically, similar to [25], the distribution of the reactions on *Repetition* failures are similar to the reactions when *No failures* occur, especially on some anger or surprise related features (*e.g.*, AU04, AU26 and loudness).

Lessons Learned

We found several behaviors to generalize across tasks, namely vowel hyper-articulation, speaking louder and gazing at the robot when failures occur, perhaps showing surprise and frustration when robots fail. This may indicate that robots observing human reactions should mainly focus on non-verbal features that seem to generalize across tasks, while features that focus on semantic representations and utterance construction may be domain dependent. Lexical features show the best performance in failure detection, and representing language that is mostly dependent to the task can have a large impact on how these representations can be adapted to a new domain. Therefore features that represent emotion and affective states may be more useful in generalized failure detection models. In this article, we have not discussed affective dimensions extensively. We found similarities across datasets in speech features such as loudness, tone and emotion and also facial features associated with surprise and vowel pronunciation. Future research in multimodal HRI should investigate failure detection among a larger variety of tasks between humans and robots and also focus on affective dimensions of failure analyses.

7 CONCLUSION

The main purpose of this article is to investigate the differences in people's reactions to failures from varied interaction paradigms. We expected that certain behaviors in reactions to failures would generalize across tasks and that they can be used for failure detection. Our assumption was partially correct, a systematic analysis on human responses demonstrates that acoustic and visual behaviors are consistent in reactions to failures. We also find that failure detection achieves good performance and most responses to failures are independent of when failures happen in the interaction. While cross-corpus validation does not work as expected, these results have implications to HRI research, namely what human responses generalize across domains. Detecting failures is an important robot mechanism for identifying opportunities to repair dialogue breakdowns. Future work should look at how to best combine domain adaptation mechanisms to leverage the learned behaviors from humans in new tasks and interactions with robots.

ACKNOWLEDGMENTS

This research was sponsored by the Swedish Foundation for Strategic Research (FACT: GMT14-0082) and by the U.S. Army Research Office accomplished under Cooperative Agreement Number W911NF-20-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. We thank Ron Artstein and Gale Lucas for sharing the Negotiation corpus.

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 *International Conference on Social Robotics*. Springer, 293–303.

[2] Deepali Aneja, Daniel McDuff, and Mary Czerwinski. 2020. Conversational Error Analysis in Human-Agent Interaction. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–8.

[3] Ron Artstein, Jill Boberg, Alesia Gainer, Jonathan Gratch, Emmanuel Johnson, Anton Leuski, Gale Lucas, and David Traum. 2018. The Niki and Julie Corpus: collaborative multimodal dialogues between humans, robots, and virtual agents. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

[4] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–10.

[5] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48. https://doi.org/10.18637/jss.v067.i01

[6] Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter* 6, 1 (2004), 20–29.

[7] Dan Bohus. 2007. *Error awareness and recovery in conversational spoken language interfaces*. Technical Report. CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE.

[8] Dan Bohus and Alexander Rudnicky. 2005. Sorry and I Didn't Catch That!-An Investigation of Non-understanding Errors and Recovery Strategies. In *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*. 128–143.

[9] Judee K Burgoon and Jerold L Hale. 1988. Nonverbal expectancy violations: Model elaboration and application to immediacy behaviors. *Communications Monographs* 55, 1 (1988), 58–79.

[10] Dito Eka Cahya, Rahul Ramakrishnan, and Manuel Giuliani. 2019. Static and Temporal Differences in Social Signals Between Error-Free and Erroneous Situations in Human-Robot Collaboration. In *International Conference on Social Robotics*. Springer, 189–199.

[11] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[12] Herbert H Clark and Susan E Brennan. 1991. Grounding in communication. (1991).

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[14] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. 1459–1462.

[15] Russell H Fazio and Michael A Olson. 2007. Attitudes: Foundations, functions, and consequences. *The handbook of social psychology* (2007), 123–145.

[16] Rebecca Flook, Anas Shrinah, Luc Wijnen, Kerstin Eder, Chris Melhuish, and Séverin Lemaignan. 2019. On the impact of different types of errors on trust in human-robot interaction: Are laboratory-based HRI experiments trustworthy? *Interaction Studies* 20, 3 (2019), 455–486.

[17] Raphaela Gehle, Karola Pitsch, Timo Dankert, and Sebastian Wrede. 2015. Effects of a robot's unexpected reactions in robot-to-group interactions. (2015).

[18] Manuel Giuliani, Nicole Mirnig, Gerald Stollnberger, Susanne Stadler, Roland Buchner, and Manfred Tscheligi. 2015. Systematic analysis of video data from different human–robot interaction studies: a categorization of social signals during error situations. *Frontiers in psychology* 6 (2015), 931.

[19] Joakim Gustafson and Linda Bell. 2000. Speech technology on trial: Experiences from the August system. *Natural Language Engineering* 6, 3-4 (2000), 273–286.

[20] Cory J Hayes, Maryam Moosaei, and Laurel D Riek. 2016. Exploring implicit human responses to robot mistakes in a learning from demonstration task. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 246–252.

[21] Julia Hirschberg, Diane Litman, and Marc Swerts. 2004. Prosodic and other cues to speech recognition failures. *Speech communication* 43, 1-2 (2004), 155–175.

[22] 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.

[23] Jeesun Kim and Chris Davis. 2016. The Consistency and Stability of Acoustic and Visual Cues for Different Prosodic Attitudes.. In *INTERSPEECH*. 57–61.

[24] Dimosthenis Kontogiorgos, Andre Pereira, and Joakim Gustafson. 2021. Grounding behaviours with conversational interfaces: effects of embodiment and failures. *Journal on Multimodal User Interfaces* (2021), 1–16.

[25] Dimosthenis Kontogiorgos, Andre Pereira, Boran Sahindal, Sanne van Waveren, and Joakim Gustafson. 2020. Behavioural responses to robot conversational failures. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 53–62.

[26] Dimosthenis Kontogiorgos, Sanne van Waveren, Olle Wallberg, Andre Pereira, Iolanda Leite, and Joakim Gustafson. 2020. Embodiment Effects in Interactions with Failing Robots. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[27] Jacqueline Marie Kory-Westlund. 2019. *Relational ai: Creating long-term interpersonal interaction, rapport, and relationships with social robots*. Ph.D. Dissertation. Massachusetts Institute of Technology.

[28] Diane Litman, Marilyn Walker, and Michael S Kearns. 1999. Automatic detection of poor speech recognition at the dialogue level. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. 309–316.

[29] Raveesh Meena, José Lopes, Gabriel Skantze, and Joakim Gustafson. 2015. Automatic detection of miscommunication in spoken dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 354–363.

[30] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 72–78.

[31] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.

[32] R Core Team. 2020. R: A Language and Environment for Statistical Computing. (2020). https://www.R-project.org/

[33] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3973–3983.

[34] Albert Rilliard, Donna Erickson, Takaaki Shochi, and João Antônio de Moraes. 2013. Social face to face communication-American English attitudinal prosody.. In *INTERSPEECH*. 1648–1652.

[35] Emanuel A Schegloff and Harvey Sacks. 1973. Opening up closings. *Semiotica* 8, 4 (1973), 289–327.

[36] Chao Shi, Masahiro Shiomi, Christian Smith, Takayuki Kanda, and Hiroshi Ishiguro. 2013. A Model of Distributional Handing Interaction for a Mobile Robot.. In *Robotics: science and systems*. 24–28.

[37] Elaine Schaertl Short, Mai Lee Chang, and Andrea Thomaz. 2018. Detecting contingency for HRI in open-world environments. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 425–433.

[38] Gabriel Skantze. 2005. Exploring human error recovery strategies: Implications for spoken dialogue systems. *Speech Communication* 45, 3 (2005), 325–341.

[39] Gabriel Skantze and Jens Edlund. 2004. Early error detection on word level. In *COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction*.

[40] David R Traum and Peter A Heeman. 1996. Utterance units in spoken dialogue. In *Workshop on Dialogue Processing in Spoken Language Systems*. Springer, 125–140.

[41] Pauline Trung, Manuel Giuliani, Michael Miksch, Gerald Stollnberger, Susanne Stadler, Nicole Mirnig, and Manfred Tscheligi. 2017. Head and shoulders: automatic error detection in human-robot interaction. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 181–188.

[42] Marilyn Walker, Jerry Wright, and Irene Langkilde. 2000. Using natural language processing and discourse features to identify understanding errors in a spoken dialogue system. In *Proceedings of the 17th international conference on machine learning*. 1111–1118.