Does the Appearance of Autonomous Conversational Robots Affect User Spoken Behaviors in Real-World Conference Interactions?

Zi Haur Pang
Graduate School of Informatics
Kyoto University
Kyoto, Japan
pang.haur.42a@st.kyoto-u.ac.jp

Yahui Fu
Graduate School of Informatics
Kyoto University
Kyoto, Japan
fu.yahuiii@gmail.com

Divesh Lala
Graduate School of Informatics
Kyoto University
Kyoto, Japan
divesh.lala@gmail.com

Mikey Elmers
Graduate School of Informatics
Kyoto University
Kyoto, Japan
elmers@sap.ist.i.kyoto-u.ac.jp

Koji Inoue
Graduate School of Informatics
Kyoto University
Kyoto, Japan
inoue@sap.ist.i.kyoto-u.ac.jp

Tatsuya Kawahara
School of Informatics
Kyoto University
Kyoto, Japan
kawahara@i.kyoto-u.ac.jp

Abstract

We investigate the impact of robot appearance on users' spoken behavior during real-world interactions by comparing a human-like android, ERICA, with a less anthropomorphic humanoid, TELECO. Analyzing data from 42 participants at SIGDIAL 2024, we extracted linguistic features such as disfluencies and syntactic complexity from conversation transcripts. The results showed moderate effect sizes, suggesting that participants produced fewer disfluencies and employed more complex syntax when interacting with ERICA. Further analysis involving training classification models like Naïve Bayes, which achieved an F1-score of 71.60%, and conducting feature importance analysis, highlighted the significant role of disfluencies and syntactic complexity in interactions with robots of varying human-like appearances. Discussing these findings within the frameworks of cognitive load and Communication Accommodation Theory, we conclude that designing robots to elicit more structured and fluent user speech can enhance their communicative alignment with humans.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; **Field studies**.

Keywords

Anthropomorphism, Conversational Robots, Linguistic Markers

ACM Reference Format:
Zi Haur Pang, Yahui Fu, Divesh Lala, Mikey Elmers, Koji Inoue, and Tatsuya Kawahara. 2025. Does the Appearance of Autonomous Conversational Robots Affect User Spoken Behaviors in Real-World Conference Interactions?. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3706599.3720179

CHI EA '25, Yokohama, Japan

ACM ISBN 979-8-4007-1395-8/25/04
https://doi.org/10.1145/3706599.3720179

1 Introduction

One of the key objectives of conversational robots is to achieve human-level interaction. To achieve this, previous research has explored various aspects, including reasoning [10, 60], empathy [14, 40], and personality [30, 61], to enable robots to produce more human-like responses. Additionally, non-verbal interaction features, such as backchanneling [1, 21], head nodding [4, 64], and gestures [2, 36, 56], have been extensively studied to enhance the naturalness of human-robot interactions.

Besides these communication elements, the appearance of robots remains a critical factor in shaping user perceptions and behavior. A robot's appearance strongly influences the user's first impression, particularly regarding its perceived level of human-likeness. Previous studies have shown that robots with more human-like appearances can enhance perceived warmth [26, 32], empathy [62, 66], and social presence [49, 53], demonstrating their effectiveness in various social and cultural settings [29, 41, 59, 63].

However, research on the effect of robot appearance on user behavior faces several limitations. First, many studies involve participants interacting with images or videos of robots rather than real, physical robots, which may overlook the impact of social presence during interactions [5, 23, 48]. Second, experiments often rely on teleoperated or Wizard-of-Oz (WoZ) methodologies, where the robot's behavior is controlled by a human operator [51, 53, 58]. Such setups may not fully reflect the dynamics of interactions in autonomous systems. Third, most studies are conducted in laboratory settings, where participants are recruited for controlled experiments [6, 47, 57]. These conditions may differ significantly from real-world scenarios, potentially affecting user behavior. Finally, prior research often relies on simple metrics (e.g., conversation length, informativeness, etc.) through self-reported scales (e.g., 7-point Likert scale [31]) [24, 37, 65, 66], leaving space for deeper analysis through other perspectives.

Figure 1: Photo of interview dialogue with ERICA by SIGDIAL participant

Figure 2: Photo of interview dialogue with TELECO by SIGDIAL participant

To address these gaps, we investigate how robot appearance affect user spoken behavior during real interactions with physical, fully autonomous conversational robots at an international conference. This setting allows us to capture user responses that closely resemble everyday behavior, providing insights into how appearance influences user speech in genuine social contexts. We leverage various features in natural language processing (NLP) and conversation analysis, derived from a linguistic perspective, to offer a fine-grained analysis of user behavior. Furthermore, we developed a machine learning model capable of predicting the robot's human-likeness based on observed user behavior.

Our contributions are twofold:

- We investigated how the human-likeness of an autonomous conversational robot influenced user spoken behavior in real-world interactions, from linguistic perspective.
- We developed a predictive model that classifies a robot's human-likeness based upon a user's spoken behavior, illuminating important linguistic cues through feature importance analysis.

2 Dataset

2.1 Conversational Robots

In this study, we employed two robots with different levels of human-likeness:

- **ERICA [17, 22, 25]:** An android robot designed to resemble an adult female (see Figure 1).
- **TELECO [20, 27]:** A humanoid robot featuring an OLED display for its face and simplified joint structures (see Figure 2).

To isolate the impact of appearance on user behavior, we implemented the same human-like spoken dialogue system for both robots [39]. They used identical dialogue behaviors, gestures, and facial expressions to ensure any observed differences could be attributed primarily to the differences in appearances. The system architecture can be found in Figure 3.

2.2 Data Collection

We conducted our study at SIGDIAL 2024[1], an international conference attended by over 160 participants. As shown in Figure 1 and Figure 2, attendees took part in brief, one-on-one interviews with either ERICA or TELECO, each lasting two to three minutes. We selected an interview approach because it is both engaging and well-suited to the conference setting, allowing us to observe how human-likeness might influence user behavior under natural interaction conditions.

No formal questionnaires were administered to maintain a casual atmosphere and natural interaction experience. Consent was obtained through informing the attendees about the study at the conference's opening session and through clearly displayed notices in the interview room, that only the transcripted dialogue, captured by Automatic Speech Recognition (ASR), would be recorded. Accordingly, our analysis focuses on user spoken behavior extracted from the interview transcripts to examine how variations in the robots' appearances affect user interactions.

3 User Behavior Analysis

3.1 Behavior Metrics

We investigated multiple dimensions of user spoken behavior, each reflecting different linguistic constructs. We grouped the metrics into four main categories—(1) Linguistic, (2) Dialogue, (3) Emotion, and (4) Behavioral Mimicry. Below, we outline the primary measures:

(1) Linguistic. This category quantifies the structural and lexical complexity of user responses:

- *Number of Words and Utterance Length*: Total word count and the average number of words per utterance, indicating how extensively or succinctly participants responded.
- *Lexical Diversity*: The ratio of unique words to total words, widely considered a hallmark of expressive vocabulary and verbal fluency.

[1] https://2024.sigdial.org/

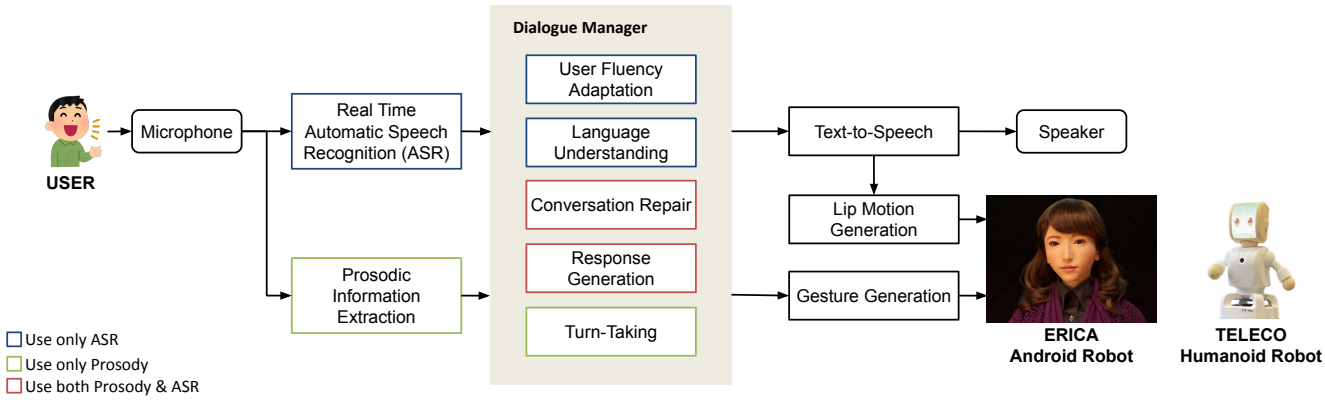

Figure 3: Overall architecture of the interview system implemented in our study. This comprehensive system architecture includes modules for real-time automatic speech recognition (ASR), prosodic information extraction, language understanding, and user fluency adaptation, among others. Central to the system is the dialogue manager, which coordinates turn-taking, response generation, and conversation repair. Also included are the text-to-speech, gesture generation, and lip motion generation components, enhancing the robots' interactive capabilities.

- *Syntactic Complexity*: Measured by dependency-parse depth through SpaCy[2] , reflecting how participants layered or embedded phrases within each sentence (a higher average suggests more complex syntax).

(2) Dialogue. This component examines user interactions and response dynamics:

- *Number of disfluencies and disfluencies ratio*: Counting interjections (e.g., "um," "uh") and consecutive repeated words. Interjections is calculated using the `en_core_web_lg` model[3] in spaCy [4]. Repeated words is calculated through bi-gram models in NLTK library[5].
- *Politeness Score*: We evaluated whether user politeness was influenced by the robots appearance. The score is derived from an XLM-RoBERTa-based classification of polite vs. impolite utterances [52], normalized between 0 and 1.
- *Word Commonness*: We evaluated how common word usage was influenced by the robots appearance. By using the Brown Corpus [7] as a reference corpus, we built a commonness score calculator based on the normalized frequency. Then we computed the final score through the average commonness score for all words in a single utterance, indicating the "commonness" of a user's vocabulary.
- *Personal Pronouns Ratio*: We tracked whether users referred to the robot using personal pronouns (e.g., "you," "he," "she"), or impersonal pronouns (e.g., "it," "the robot"), which can shed light on animacy attribution and self-reference patterns [38, 43].
- *Hedging Word Ratio*: We evaluated user hedging behavior based the robot's appearance. Hedging involves words or phrases expressing uncertainty or non-commitment (e.g., "maybe," "seem," "usually"). We detected such terms using a predefined list[6].

(3) Emotion. Focused on the emotional content of user interactions:

- *Sentiment Score*: To evaluate whether users' utterances conveyed positive or negative sentiment, we used SiEBERT [19]. We computed the sentiment score through polarity calculation (positive sentiment score - negative sentiment score). We then normalized the value to a range between 0 and 1, where higher scores indicate more positive sentiment.
- *Emotion Score (Joy, Sadness, Anger, Fear, Disgust, and Surprise)*: We implemented the likelihood estimates assigned by a DistilRoBERTa-based emotion classifier [18], clarifying which emotions predominated in each user's speech.

(4) Behavioral Mimicry. Investigates the extent of user mimicry of robot behavior, a key indicator of empathy and social bonding [15, 42, 50]:

- *Lexical Mirroring*: Overlap in word choice between the user's and robot's utterances, normalized by the user's total number of words.
- *Semantic Mirroring*: A BERTScore-based measure indicating semantic similarity between user and robot dialogue.
- *Syntactic Mirroring*: Assessed via POS-based cosine similarity, capturing how closely users' grammatical structures mirror the robot's output.

3.2 Result

A total of 42 participants interacted with our robots during the conference, 24 with ERICA and 18 with TELECO. Results are detailed in Table 1. Given our small sample size, we focused on practical significance, assessing effect sizes using Rank-biserial correlations rather than relying solely on statistical significance, which is less sensitive

[2]https://github.com/explosion/spaCy
[3]https://spacy.io/models/en#en_core_web_lg
[4]https://github.com/explosion/spaCy
[5]https://github.com/nltk/nltk
[6]https://github.com/words/hedges

Behavior	ERICA	TELECO	Effect Size
(Linguistic)			
# of Words	143.46 (66.17)	109.39 (50.81)	0.31
Utterance Length	10.32 (5.02)	8.48 (3.92)	0.17
Lexical Diversity	0.59 (0.09)	0.61 (0.09)	0.12
Syntactic Complexity	2.81 (1.01)	2.20 (0.50)	0.41
(Dialogue)			
# of Disfluencies	7.92 (6.51)	12.78 (8.89)	0.42
Disfluencies Ratio	0.61 (0.59)	1.05 (0.78)	0.41
Politeness Score	0.63 (0.15)	0.60 (0.13)	0.10
Word Commonness	6.11 (0.40)	5.91 (0.44)	0.26
Personal Pronouns Ratio	2.74 (3.23)	1.89 (1.64)	0.10
Hedging Word Ratio	1.15 (0.63)	0.94 (0.59)	0.22
(Emotion)			
Sentiment Score	0.67 (0.18)	0.64 (0.15)	0.13
Joy	0.14 (0.08)	0.11 (0.09)	0.28
Sadness	0.03 (0.02)	0.04 (0.03)	0.33
Anger	0.06 (0.02)	0.07 (0.02)	0.21
Fear	0.19 (0.08)	0.20 (0.07)	0.15
Disgust	0.08 (0.03)	0.08 (0.04)	0.05
Surprise	0.03 (0.02)	0.03 (0.02)	0.02
(Behavioral Mimicry)			
Lexical Mirroring	0.36 (0.07)	0.35 (0.11)	0.02
Semantic Mirroring	0.51 (0.02)	0.50 (0.02)	0.19
Syntactic Mirroring	0.29 (0.09)	0.28 (0.08)	0.01

Table 1: Comparative analysis of user behaviors based on robot appearance. Data presented includes the mean and standard deviation (in parentheses) for each behavior metric across the robots, accompanied by the computed effect sizes through Rank-biserial correlations.

to sample size limitations [54]. Additionally, we performed Mann-Whitney U tests; several features such as number of disfluencies (p = 0.022), disfluency ratio (p = 0.027), and syntactic complexity (p = 0.024) showed moderate unadjusted p-values. However, achieving statistical significance after Bonferroni correction (α = 0.002) was challenging due to multiple comparisons (21 in total).

Key findings include notable differences in disfluencies and syntactic complexity between interactions with ERICA and TELECO. Participants exhibited more disfluencies with TELECO both in total count (12.78 vs. 7.92, effect size = 0.417) and ratio (1.05 vs. 0.61, effect size = 0.405). They also used more complex syntax with ERICA (2.81 vs. 2.20, effect size = 0.405). Lesser variations were observed in the number of words (143.46 vs. 109.39, effect size = 0.31), word commonness (6.11 vs. 5.91, effect size = 0.26), and hedging word ratio (1.15 vs. 0.94, effect size = 0.22).

Regarding affective content, both groups reported positive sentiment scores (0.67 vs. 0.64), with slightly higher joy levels observed with ERICA (0.14 vs. 0.11). Emotional expressions such as anger, sadness, fear, surprise, and disgust showed minimal differences between the groups. Behavioral mimicry across lexical, semantic, and syntactic dimensions showed negligible differences (effect sizes = 0.02, 0.19, and 0.01, respectively).

4 Predictive Model

4.1 Experimental Setup

In addition to user behavior analysis, we developed a predictive model to identify whether a participant interacted with the more human-like robot (ERICA) or the less human-like robot (TELECO), based solely on metrics of spoken behavior detailed in Section 3.1. We selected input features for the model based on their differences and effect sizes both exceeding 0.1, which include the number of words, utterance length, syntactic complexity, number and ratio of disfluencies, word commonness, and hedging word ratio, highlighting discernible variations between the two datasets. Results using all user behavior metrics as input features are presented in Table 4 in 6 as part of an ablation study.

We evaluated a range of machine learning classifiers, including Random Forest [9], Gradient Boosting[13], and Naïve Bayes [46], along with a random baseline, employing default hyperparameters for each algorithm. The dataset was divided into an 80% training set and a 20% test set. To reduce bias and validate model robustness, we conducted 3-fold cross-validation on the training set and repeated the training/validation process ten times with different random seeds to ensure reproducibility and assess stability. We averaged the performance metrics—accuracy, macro-average precision, recall, and F1-score—across folds and seeds, compiling the final performance metrics. Detailed results for each seed are documented in Table 5 in Appendix 6.

Model	Accuracy	Precision	Recall	F1-score
Random Baseline	48.57	48.29	47.94	46.76
Random Forest	64.29	64.33	64.20	63.32
Gradient Boosting	63.10	63.56	63.21	61.98
Naïve Bayes	**72.38**	**73.49**	**72.79**	**71.60**

Table 2: Predictive Model Evaluation Result [%]

4.2 Feature Importance Analysis

In addition to model development, we explored how specific features influence our predictive models using two suggested interpretability methods [35]:

Permutation Feature Importance (PFI) [3]. This model-agnostic technique evaluates the impact of individual features by measuring the reduction in model performance (e.g., F1-score) when the values of a feature are randomly shuffled in the test set. Significant performance drops indicate critical feature importance.

SHapley Additive exPlanations (SHAP) [34]. SHAP provides a detailed measure of each feature's contribution to individual predictions, where we use the TreeExplainer [33] to calculate SHAP values in this study.

4.3 Results

Table 2 shows that the Naïve Bayes model outperformed all other evaluated algorithms with the highest F1-score (71.60%), significantly exceeding the random baseline (46.76%) and other methods

such as Random Forest, and Gradient Boosting (61-63%). This underscores that linguistic features effectively differentiate participants' perceptions of robot human-likeness.

Feature impact analysis on the Naïve Bayes model revealed that syntactic complexity was the most influential predictor, contributing 22.51% in SHAP and 0.079 in PFI. Other significant features included the number of words (17.17% in SHAP, 0.047 in PFI) and various disfluency measures, with number and ratio contributing 15.34% and 15.57% in SHAP and 0.031 and 0.026 in PFI, respectively. These insights are visualized in Figures 4 and summarized in Table 3.

Behavior	Permutation Feature Importance (PFI)	SHAP	
		Mean	Percentage
Syntactic Complexity	0.079 (0.050)	0.102	22.51
Number of Words	0.047 (0.035)	0.078	17.17
Disfluencies Ratio	0.026 (0.036)	0.070	15.57
Number of Disfluencies	0.031 (0.035)	0.069	15.34
Word Commonness	0.053 (0.041)	0.056	12.32
Utterance Length	0.003 (0.028)	0.045	10.02
Hedging Word Ratio	0.003 (0.031)	0.032	7.07

Table 3: Feature importance analysis for the Naïve Bayes model using Permutation Feature Importance (PFI) and SHapley Additive exPlanations (SHAP). The table displays the mean SHAP values and their respective contribution percentages, along with the PFI scores (mean and standard deviation in parentheses), sorted by their contribution to the model's predictive performance.

5 General Discussion

This section reflects on key findings from the analysis and modeling results concerning disfluencies and syntactic complexity metrics. Notably, users engaged with ERICA, the more human-like robot, exhibited more complex syntax as detailed in Section 3.2. This aligns with the predictive model results in Section 4.3, where syntactic complexity emerged as a significant predictor. This observation may be attributed to the perceived cognitive capabilities of the robots, influenced by their level of human-likeness. According to Communication Accommodation Theory, individuals tend to adjust their communication style to match their conversational partner's perceived attributes or capabilities [16]. Previous studies also indicate that more human-like robots are often ascribed higher cognitive functions [12]. Thus, ERICA with a human-like appearance will lead users to adopt more complex language, akin to human-human interactions, under the assumption that ERICA can process such communication.

In contrast, disfluencies demonstrated an opposite trend, with users exhibiting more disfluencies when interacting with the less human-like robot, TELECO. This phenomenon may be attributed to increased cognitive load, which refers to the utilization of working memory resources during cognitive tasks. Cognitive psychology suggests that unfamiliar tasks require more information processing, which can escalate cognitive load. Prior research has indicated that heightened cognitive load can manifest through various temporal characteristics such as altered speech rate and increased ratio of pauses, or through interactions with less legible text, leading to a rise in disfluency rates [8, 28]. In our context, TELECO's less human-like appearance might have been perceived as unfamiliar or unnatural compared to typical human interactions, necessitating greater cognitive effort and consequently, higher disfluency rates. This observation aligns with previous findings where interactions with autonomous systems, characterized by less human-like features, elicited more disfluencies than more anthropomorphic Wizard-of-Oz (WoZ) conditions [11].

An intriguing outcome from our study relates to the behavior mimicry in Section 3.2, where no significant differences in mimicry levels were found across human-likeness levels, contrasting with prior research that suggests humans often mimic robots during interactions [45, 55]. This deviation could be explained by the context of the interactions, which were structured as interviews rather than dynamic social engagements. Previous literature indicates that mimicry is more prevalent in emotionally charged real-life interactions, where empathy and social bonding are more critical [15, 42, 44, 50]. The controlled interview setting of our study likely limited emotional engagement, thus reducing the occurrence of mimicry.

6 Conclusion

This study investigated the impact of a robot's human-likeness appearance on user speech patterns during real-world, fully autonomous interactions at an international conference. Data from 42 participants, engaging with the highly anthropomorphic ERICA and the more basic TELECO, were analyzed and modeled. The results align across experiments, showing that linguistic markers such as disfluencies, syntactic complexity, and utterance length vary noticeably between interactions with the two robots, demonstrating moderate effect sizes and substantial contributions in feature importance analyses. These findings suggest that these linguistic features are influential in shaping user perceptions of a robot's human-likeness. We further explore the relationship between these features and robot human-like appearance from a cognitive science perspective, linking increased syntactic complexity with more human-like robots via Communication Accommodation Theory, and higher disfluencies with less human-like robots due to greater cognitive load.

Looking forward, this research could be extended in several ways. Increasing the sample size and employing more controlled manipulations of robot appearances could further validate these initial findings. Moreover, exploring factors beyond appearance—such as enhanced AI capabilities or different interaction methods (Wizard-of-Oz vs. autonomous)—could enrich our understanding of human-likeness in robots. Additionally, expanding our analytical modalities to include non-verbal cues like facial expressions and speech tone could offer a more comprehensive view of how robot appearance affects human interactions. Integrating such non-verbal signals would provide a broader context for understanding the dynamic between robot embodiments and human behavior. Ultimately, the nuanced understanding of syntax and disfluencies revealed by our study could inform the design of robots that more effectively mirror human communicative norms, thereby improving both the perceived humanness and the quality of interactions.

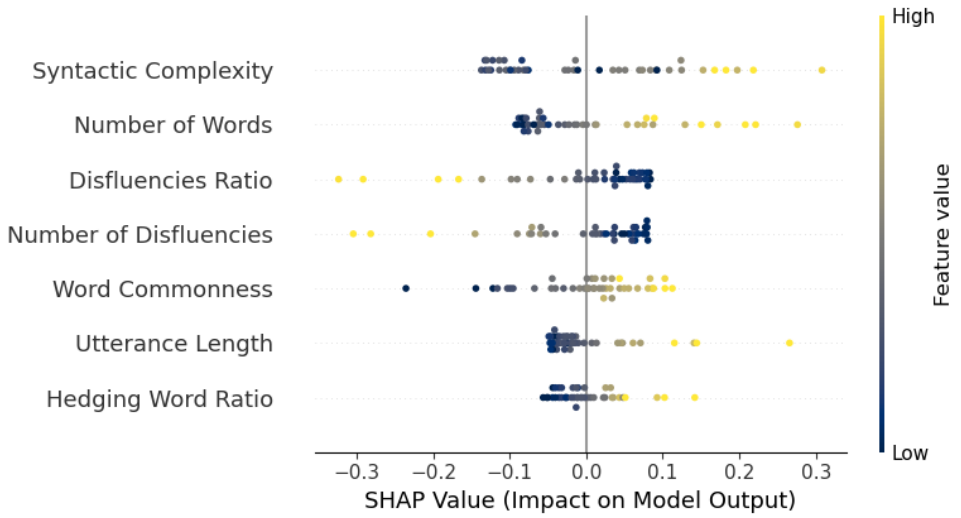

Figure 4: Distribution of SHAP values for each behavioral feature. The SHAP values, which quantify the impact on the model output, are plotted along the x-axis against each feature on the y-axis. Each point represents an individual instance. Points in blue indicate feature values below the average, affecting the model negatively (left of the vertical zero line), while points in yellow denote values above the average, contributing positively to the prediction (right of the zero line).

References

[1] Amalia Istiqlali Adiba, Takeshi Homma, and Toshinori Miyoshi. 2021. Towards immediate backchannel generation using attention-based early prediction model. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7408–7412.

[2] Ghazanfar Ali, Myungho Lee, and Jae-In Hwang. 2020. Automatic text-to-gesture rule generation for embodied conversational agents. *Computer Animation and Virtual Worlds* 31, 4-5 (2020), e1944.

[3] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. 2010. Permutation importance: a corrected feature importance measure. *Bioinformatics* 26, 10 (2010), 1340–1347.

[4] Öykü Zeynep Bayramoğlu, Engin Erzin, Tevfik Metin Sezgin, and Yücel Yemez. 2021. Engagement rewarded actor-critic with conservative Q-learning for speech-driven laughter backchannel generation. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 613–618.

[5] Daniel Belanche, Luis V Casaló, Jeroen Schepers, and Carlos Flavián. 2021. Examining the effects of robots' physical appearance, warmth, and competence in frontline services: The Humanness-Value-Loyalty model. *Psychology & Marketing* 38, 12 (2021), 2357–2376.

[6] Maxwell Bennett, Tom Williams, Daria Thames, and Matthias Scheutz. 2017. Differences in interaction patterns and perception for teleoperated and autonomous humanoid robots. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 6589–6594.

[7] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

[8] Judit Bóna and Mária Bakti. 2020. The effect of cognitive load on temporal and disfluency patterns of speech: evidence from consecutive interpreting and sight translation. *Target* 32, 3 (2020), 482–506.

[9] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.

[10] Xiuying Chen, Zhi Cui, Jiayi Zhang, Chen Wei, Jianwei Cui, Bin Wang, Dongyan Zhao, and Rui Yan. 2021. Reasoning in dialog: Improving response generation by context reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 12683–12691.

[11] Mikey Elmers, Koji Inoue, Divesh Lala, Keiko Ochi, and Tatsuya Kawahara. 2024. Analysis and Detection of Differences in Spoken User Behaviors Between Autonomous and Wizard-of-Oz Systems. In *2024 27th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE, 1–6.

[12] Leopoldina Fortunati, Anna Maria Manganelli, Joachim Höflich, and Giovanni Ferrin. 2023. Exploring the perceptions of cognitive and affective capabilities of four, real, physical robots with a decreasing degree of morphological human likeness. *International Journal of Social Robotics* 15, 3 (2023), 547–561.

[13] Jerome H Friedman. 2002. Stochastic gradient boosting. *Computational statistics & data analysis* 38, 4 (2002), 367–378.

[14] Yahui Fu, Koji Inoue, Divesh Lala, Kenta Yamamoto, Chenhui Chu, and Tatsuya Kawahara. 2023. Dual variational generative model and auxiliary retrieval for empathetic response generation by conversational robot. *Advanced Robotics* 37, 21 (2023), 1406–1418.

[15] Vittorio Gallese. 2006. Embodied simulation: from mirror neuron systems to interpersonal relations. In *Empathy and Fairness: Novartis Foundation Symposium 278*. Wiley Online Library, 3–19.

[16] Howard Giles, Tania Ogay, et al. 2007. Communication accommodation theory. *Explaining communication: Contemporary theories and exemplars* (2007), 293–310.

[17] Dylan F Glas, Takashi Minato, Carlos T Ishi, Tatsuya Kawahara, and Hiroshi Ishiguro. 2016. Erica: The erato intelligent conversational android. In *2016 25th IEEE International symposium on robot and human interactive communication (RO-MAN)*. IEEE, 22–29.

[18] Jochen Hartmann. 2022. Emotion English DistilRoBERTa-base. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/.

[19] Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. More than a Feeling: Accuracy and Application of Sentiment Analysis. *International Journal of Research in Marketing* 40, 1 (2023), 75–87. https://doi.org/10.1016/j.ijresmar.2022.05.005

[20] Yukiko Horikawa, Takahiro Miyashita, Akira Utsumi, Shogo Nishimura, and Satoshi Koizumi. 2023. Cybernetic avatar platform for supporting social activities of all people. In *2023 IEEE/SICE International Symposium on System Integration (SII)*. IEEE, 1–4.

[21] Koji Inoue, Divesh Lala, Gabriel Skantze, and Tatsuya Kawahara. 2024. Yeah, Un, Oh: Continuous and Real-time Backchannel Prediction with Fine-tuning of Voice Activity Projection. *arXiv preprint arXiv:2410.15929* (2024).

[22] Koji Inoue, Pierrick Milhorat, Divesh Lala, Tianyu Zhao, and Tatsuya Kawahara. 2016. Talking with ERICA, an autonomous android. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 212–215.

[23] Yoonwon Jung and Sowon Hahn. 2023. Social Robots As Companions for Lonely Hearts: The Role of Anthropomorphism and Robot Appearance. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2520–2525.

[24] Takayuki Kanda, Takahiro Miyashita, Taku Osada, Yuji Haikawa, and Hiroshi Ishiguro. 2008. Analysis of humanoid appearances in human–robot interaction. *IEEE transactions on robotics* 24, 3 (2008), 725–735.

[25] Tatsuya Kawahara. 2019. Spoken dialogue system for a human-like conversational robot ERICA. In *9th International Workshop on Spoken Dialogue System Technology*. Springer, 65–75.

[26] Seo Young Kim, Bernd H Schmitt, and Nadia M Thalmann. 2019. Eliza in the uncanny valley: Anthropomorphizing consumer robots increases their perceived

warmth but decreases liking. *Marketing letters* 30 (2019), 1–12.
[27] Yoshihisa Kondo, Hiroyuki Yomo, Shogo Nishimura, Akira Utsumi, and Takahiro Miyashita. 2023. A Practical Implementation of Multi-Radio Wi-Fi for Tele-operated Mobile Robots. In *2023 IEEE International Conference on Omni-layer Intelligent Systems (COINS)*. IEEE, 1–6.
[28] Tim Kühl and Alexander Eitel. 2016. Effects of disfluency on cognitive and metacognitive processes and outcomes. *Metacognition and Learning* 11 (2016), 1–13.
[29] Dingjun Li, PL Patrick Rau, and Ye Li. 2010. A cross-cultural study: Effect of robot appearance and task. *International Journal of Social Robotics* 2 (2010), 175–186.
[30] Yunlong Liang, Fandong Meng, Ying Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021. Infusing multi-source knowledge with heterogeneous graph neural network for emotional conversation generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 13343–13352.
[31] Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology* (1932).
[32] Xing Stella Liu, Xiao Shannon Yi, and Lisa C Wan. 2022. Friendly or competent? The effects of perception of robot appearance and service context on usage intention. *Annals of Tourism Research* 92 (2022), 103324.
[33] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2, 1 (2020), 2522–5839.
[34] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf
[35] Christoph Molnar, Gunnar König, Julia Herbinger, Timo Freiesleben, Susanne Dandl, Christian A Scholbeck, Giuseppe Casalicchio, Moritz Grosse-Wentrup, and Bernd Bischl. 2020. General pitfalls of model-agnostic interpretation methods for machine learning models. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*. Springer, 39–68.
[36] Rajmund Nagy, Taras Kucherenko, Birger Moell, André Pereira, Hedvig Kjellström, and Ulysses Bernardet. 2021. A framework for integrating gesture generation models into interactive conversational agents. *arXiv preprint arXiv:2102.12302* (2021).
[37] Andreea Niculescu, Betsy Van Dijk, Anton Nijholt, Haizhou Li, and Swee Lan See. 2013. Making social robots more attractive: the effects of voice pitch, humor and empathy. *International journal of social robotics* 5 (2013), 171–191.
[38] Grant Packard, Sarah G Moore, and Brent McFerran. 2018. (I'm) happy to help (you): The impact of personal pronoun use in customer–firm interactions. *Journal of Marketing Research* 55, 4 (2018), 541–555.
[39] Zi Haur Pang, Yahui Fu, Divesh Lala, Mikey Elmers, Koji Inoue, and Tatsuya Kawahara. 2024. Human-Like Embodied AI Interviewer: Employing Android ERICA in Real International Conference. *arXiv preprint arXiv:2412.09867* (2024).
[40] Zi Haur Pang, Yahui Fu, Divesh Lala, Keiko Ochi, Koji Inoue, and Tatsuya Kawahara. 2024. Acknowledgment of Emotional States: Generating Validating Responses for Empathetic Dialogue. *arXiv preprint arXiv:2402.12770* (2024).
[41] Akanksha Prakash and Wendy A Rogers. 2015. Why some humanoid faces are perceived more positively than others: effects of human-likeness and task. *International journal of social robotics* 7, 2 (2015), 309–331.
[42] Ryszard Praszkier. 2016. Empathy, mirror neurons and SYNC. *Mind & Society* 15 (2016), 1–25.
[43] Jianhong Qu, Ronggang Zhou, and Zhe Chen. 2022. The effect of personal pronouns on users and the social role of conversational agents. *Behaviour & Information Technology* 41, 16 (2022), 3470–3486.
[44] Inbal Ravreby, Mayan Navon, Eliya Pinhas, Jenya Lerer, Yoav Bar-Anan, and Yaara Yeshurun. 2024. The many faces of mimicry depend on the social context. *Emotion* (2024).
[45] Laurel D Riek, Philip C Paul, and Peter Robinson. 2010. When my robot smiles at me: Enabling human-robot rapport via real-time head gesture mimicry. *Journal on Multimodal User Interfaces* 3 (2010), 99–108.
[46] Irina Rish et al. 2001. An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, Vol. 3. Seattle, WA, USA;, 41–46.
[47] Kerem Rızvanoğlu, Özgürol Öztürk, and Öner Adıyaman. 2014. The impact of human likeness on the older adults' perceptions and preferences of humanoid robot appearance. In *Design, User Experience, and Usability. User Experience Design Practice: Third International Conference, DUXU 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014, Proceedings, Part IV 3*. Springer, 164–172.
[48] Waka Saeki and Yoshiyuki Ueda. 2024. Impact of politeness and performance quality of android robots on future interaction decisions: a conversational design perspective. *Frontiers in Robotics and AI* 11 (2024), 1393456.
[49] Jordan A Sasser, Daniel S McConnell, and Janan A Smither. 2024. Investigation of Relationships Between Embodiment Perceptions and Perceived Social Presence in Human–Robot Interactions. *International Journal of Social Robotics* (2024), 1–16.
[50] Martin Schulte-Rüther, Hans J Markowitsch, Gereon R Fink, and Martina Piefke. 2007. Mirror neuron and theory of mind mechanisms involved in face-to-face interactions: a functional magnetic resonance imaging approach to empathy. *Journal of cognitive neuroscience* 19, 8 (2007), 1354–1372.
[51] Sichao Song, Jun Baba, Junya Nakanishi, Yuichiro Yoshikawa, and Hiroshi Ishiguro. 2022. Costume vs. Wizard of Oz vs. Telepresence: how social presence forms of tele-operated robots influence customer behavior. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 521–529.
[52] Anirudh Srinivasan and Eunsol Choi. 2022. TyDiP: A Dataset for Politeness Classification in Nine Typologically Diverse Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. 5723–5738.
[53] Ilona Straub. 2016. 'It looks like a human!'The interrelation of social presence, interaction and agency ascription: a case study about the effects of an android robot on social agency ascription. *AI & society* 31 (2016), 553–571.
[54] Gail M Sullivan and Richard Feinn. 2012. Using effect size—or why the P value is not enough. *Journal of graduate medical education* 4, 3 (2012), 279–282.
[55] Apurv Suman, Rebecca Marvin, Elena Corina Grigore, Henny Admoni, and Brian Scassellati. 2016. Prior behavior impacts human mimicry of robots. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1057–1062.
[56] Gabriele Trovato, Martin Do, Ömer Terlemez, Christian Mandery, Hiroyuki Ishii, Nadia Bianchi-Berthouze, Tamim Asfour, and Atsuo Takanishi. 2016. Is hugging a robot weird? Investigating the influence of robot appearance on users' perception of hugging. In *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*. IEEE, 318–323.
[57] Fang-Wu Tung. 2016. Child perception of humanoid robot appearance and behavior. *International Journal of Human-Computer Interaction* 32, 6 (2016), 493–502.
[58] Alberto Villani, T Lisini Baldi, Nicole D'Aurizio, Giulio Campagna, and Domenico Prattichizzo. 2024. Does Robot Anthropomorphism Improve Performance and User Experience in Teleoperation?. In *2024 IEEE-RAS 23rd International Conference on Humanoid Robots (Humanoids)*. IEEE, 76–83.
[59] Michael L Walters, Dag S Syrdal, Kerstin Dautenhahn, René Te Boekhorst, and Kheng Lee Koay. 2008. Avoiding the uncanny valley: robot appearance, personality and consistency of behavior in an attention-seeking home scenario for a robot companion. *Autonomous Robots* 24 (2008), 159–178.
[60] Jiashuo Wang, Wenjie Li, Peiqin Lin, and Feiteng Mu. 2021. Empathetic response generation through graph-based multi-hop reasoning on emotional causality. *Knowledge-Based Systems* 233 (2021), 107547.
[61] Yuwei Wu, Xuezhe Ma, and Diyi Yang. 2021. Personalized response generation via generative split memory network. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1956–1970.
[62] Wenjing Yang and Yunhui Xie. 2024. Can robots elicit empathy? The effects of social robots' appearance on emotional contagion. *Computers in Human Behavior: Artificial Humans* 2, 1 (2024), 100049.
[63] Shengliang Zhang, Guanyu Tang, Xiaodong Li, and Ai Ren. 2023. The effects of appearance personification of service robots on customer decision-making in the product recommendation context. *Industrial Management & Data Systems* 123, 2 (2023), 578–595.
[64] Mohan Zhou, Yalong Bai, Wei Zhang, Ting Yao, Tiejun Zhao, and Tao Mei. 2022. Responsive listening head generation: a benchmark dataset and baseline. In *European Conference on Computer Vision*. Springer, 124–142.
[65] Xiaoling Zhu, Wenrui Liang, Wenjun Xv, and Yimin Wang. 2023. The Key Strategies for Increasing Users' Intention of Self-Disclosure in Human-Robot Interaction through Robotic Appearance Design. In *SHS Web of Conferences*, Vol. 165. EDP Sciences, 01012.
[66] Jakub Złotowski, Hidenobu Sumioka, Shuichi Nishio, Dylan F Glas, Christoph Bartneck, and Hiroshi Ishiguro. 2016. Appearance of a robot affects the impact of its behaviour on perceived trustworthiness and empathy. *Paladyn, Journal of Behavioral Robotics* 7, 1 (2016), 000010151520160005.

Appendix

Model	Accuracy	Precision	Recall	F1-score
Random Baseline	49.76	50.89	50.33	48.72
Random Forest	55.71	55.25	54.81	53.27
Gradient Boosting	53.57	52.59	52.55	50.73
Naive Bayes	**57.62**	**59.35**	**58.74**	**56.59**

Table 4: Performance evaluation of predictive models [%] without feature selection. Best performing values for each metric are highlighted in bold.

Model	Accuracy	Precision	Recall	F1-score
Seed 1				
Random Baseline	0.33	0.34	0.33	0.33
Random Forest	0.69	0.69	0.69	0.69
Gradient Boosting	0.69	0.69	0.69	0.68
Naive Bayes	**0.76**	**0.79**	**0.78**	**0.76**
Seed 2				
Random Baseline	0.52	0.57	0.54	0.51
Random Forest	0.60	0.61	0.61	0.59
Gradient Boosting	0.62	0.59	0.59	0.59
Naive Bayes	**0.74**	**0.74**	**0.73**	**0.73**
Seed 3				
Random Baseline	0.55	0.57	0.56	0.53
Random Forest	0.60	0.59	0.58	0.57
Gradient Boosting	0.62	0.62	0.61	0.60
Naive Bayes	**0.74**	**0.81**	**0.75**	**0.72**
Seed 4				
Random Baseline	0.48	0.44	0.45	0.44
Random Forest	**0.74**	**0.73**	**0.73**	**0.73**
Gradient Boosting	0.64	0.67	0.66	0.63
Naive Bayes	0.69	0.70	0.70	0.69
Seed 5				
Random Baseline	0.48	0.46	0.47	0.46
Random Forest	0.64	0.65	0.65	0.64
Gradient Boosting	0.62	0.63	0.62	0.62
Naive Bayes	**0.71**	**0.72**	**0.72**	**0.71**
Seed 6				
Random Baseline	0.40	0.40	0.41	0.40
Random Forest	0.67	0.67	0.67	0.66
Gradient Boosting	0.69	0.69	0.69	0.69
Naive Bayes	**0.71**	**0.72**	**0.72**	**0.71**
Seed 7				
Random Baseline	0.40	0.38	0.38	0.38
Random Forest	0.64	0.64	0.64	0.63
Gradient Boosting	0.67	0.68	0.67	0.66
Naive Bayes	**0.71**	**0.72**	**0.73**	**0.71**
Seed 8				
Random Baseline	0.45	0.46	0.46	0.42
Random Forest	0.57	0.56	0.56	0.56
Gradient Boosting	0.50	0.51	0.51	0.50
Naive Bayes	**0.71**	**0.72**	**0.72**	**0.71**
Seed 9				
Random Baseline	0.69	0.68	0.69	0.68
Random Forest	**0.71**	**0.71**	**0.72**	**0.71**
Gradient Boosting	**0.71**	**0.71**	0.71	0.70
Naive Bayes	**0.71**	0.70	0.70	0.70
Seed 10				
Random Baseline	0.55	0.53	0.52	0.52
Random Forest	0.57	0.57	0.57	0.56
Gradient Boosting	0.55	0.56	0.55	0.53
Naive Bayes	**0.74**	**0.72**	**0.72**	**0.72**

Table 5: Predictive model evaluation results [%] for each random seed. The highest performing values for each metric per seed are bolded.