# Prediction of negative user reactions towards system responses during attentive listening

Divesh Lala, Koji Inoue and Tatsuya Kawahara
Graduate School of Informatics, Kyoto University, Japan
E-mail: [lala][inoue][kawahara]@sap.ist.i.kyoto-u.ac.jp

***Abstract*—Detecting user dissatisfaction with system responses is an important task for situated spoken dialogue systems. In this work we train models to predict negative user reactions which display this dissatisfaction. The dialogue scenario used in this work is attentive listening and we use two distinct corpora for analysis. We take a multimodal approach and use audio, visual and linguistic features in our model. Results show that audio features are the most influential and multimodality improves model performance. The importance of visual and linguistic modalities differ according to the type of corpus. We identify the key features of each corpus and find that a model trained on both corpora with seven key features reaches a precision of 0.562, outperforming the baseline for this task.**

## I. Introduction

Spoken dialogue systems have advanced in recent years with more accurate automatic speech recognition (ASR) and large language models (LLMs). However, in a situated environment there is still a need to gauge the reaction of the user towards system utterances, particularly negative reactions. We define negative reactions as user behaviors that indicate some kind of dissatisfaction with the response.

While user reactions to text-based chatbot systems can only be inferred based on text input, situated spoken dialogue systems provide information on the user state through modalities such as the tone of voice and facial expressions. By analyzing this multimodal data, we can potentially create a model which can detect negative reactions in real-time.

During robot conversation, system or dialogue model errors can produce negative reactions, so evaluation models could be used to detect these. However there are other reasons why users may be dissatisfied with the dialogue system, such as if the response is uninteresting as in Fig. 1. In this case there is nothing wrong with the dialogue itself, but the user still appears dissatisfied. Our target phenomena is related to the state of the user, not the performance of the system.

Another contribution of this work is to understand the relative importance of the individual features in our multimodal model. In this work we train models on two separate corpora with different user demographics and analyze feature importance for each of them. We then identify common important features for use in a more generalizable model. We require this type of analysis to understand what makes a reaction "negative" and to act as a basis for discussion and future research directions. The dialogue system and experiments described in this work were carried out in Japanese.

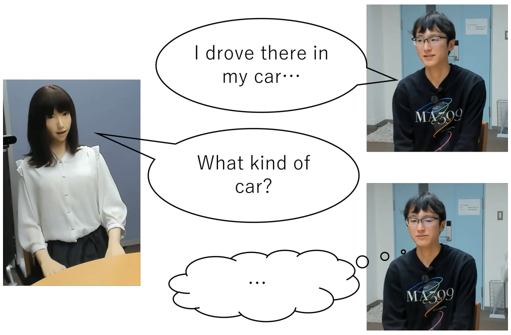

Fig. 1. Target of phenomena of this work. We study the reaction of the user after a system response and predict if they are dissatisfied even if the response is appropriate.

## II. Related work

Detection of robot errors through multimodal social signals has been a focus of much previous work [1]–[4] with several of these being applied to collaborative tasks. Other research attempts to classify confusion during interactions [5]–[7] although some are unimodal or are not designed to be used in a real-time system. For spoken interaction, dialogue breakdowns are also a common research topic [8]–[10]. These works often attempt to use linguistic analysis to detect breakdowns, although other modalities have been tested [11], [12]. Other studies aim to detect disengagement [13]–[15] as a means to understand when the user has lost interest.

We use these previous works as grounding for this research, where user reactions are modeled in free-talk conversation. Our work is differentiated by the target phenomena and the scenario. Firstly, we do not aim to detect system errors in the conversation, only the reactions of users. There may be reasons why the user may be dissatisfied with the response for reasons unrelated to the system, that could be described as social errors [16]. Several works also elicit reactions by deliberately inserting system errors into task or conversation scenarios [5], [17], [18]. In our work the reactions in the corpus arise naturally during conversation and are not a result of deliberate manipulation.

Secondly, the nature of free talk is that interaction between the user and system is through dialogue alone, although in our particular case the user does the majority of talking. Unlike other works [2], [17] there is no set task for participants in the conversation. The success of the conversation is based primarily on the user's perception and their behaviors are varied and in some cases the signals are subtle, arguably making this task more difficult than task-based scenarios.

## III. Dialogue system

In this work we use attentive listening as the scenario for the spoken dialogue system. The goal is for the system to act as an empathetic listening partner by listening to the talk of the user and providing meaningful responses. In this work both casual and more serious talks are analyzed as two distinct corpora.

We use our previous attentive listening dialogue system, details of which can be found in previous work [19], [20]. In the attentive listening scenario, the user does the majority of talk while the system uses several types of responses. The general approach is to extract a focus word which is used to generate these responses. For example, if the user is talking about pasta the dialogue system may produce an elaborating question such as "What type of pasta?".

This type of dialogue can produce negative reactions if the dialogue system makes a mistake. There are also situations where the dialogue system produces a coherent response, but it is not engaging enough for the user. The reaction of the user can indicate whether the response is actually satisfactory.

The nature of attentive listening makes detecting reactions different than for task or collaboration-based agent systems. Users often continue with their talk even if the system makes a mistake, but signal their dissatisfaction through audio and facial cues. These dissatisfaction signals, although subtle, may be obvious for humans to understand but pose a challenge for autonomous models.

## IV. Data collection and analysis

We collected sessions of users engaging with the attentive listening dialogue system. We compare two distinct corpora with differing subjects and topics - elderly and COVID. The first corpus is with elderly subjects and the topic of conversation was casual, such as a recent trip or an interesting episode in their life. The second corpus is comprised of university students, with the topic of conversation being their experiences during the COVID pandemic.

The interface is the android ERICA, a humanoid robot with a custom-built text-to-speech system. Users were instructed that they would be talking with ERICA and were given time to prepare their talk. Each session lasted for approximately 5-10 minutes and there were 56 sessions in total (36 from the elderly corpus and 20 from the COVID corpus).

Two experts viewed video of the corpora and independently annotated each of the adjacency pairs surrounding non-generic system responses as shown in Fig. 2. They then discussed their annotations together to generate the final labels.

We first annotated the appropriateness of the system response in the first adjacency pair to confirm that user reactions are not just correlated with the appropriateness of the actual system utterance. For attentive listening an appropriate utterance is a question or assessment which can stimulate more talk from the user. If an elaborating or repeated question was used, it was only deemed appropriate if the focus word was uttered by the user and that there was no other focus word which would have been more appropriate.

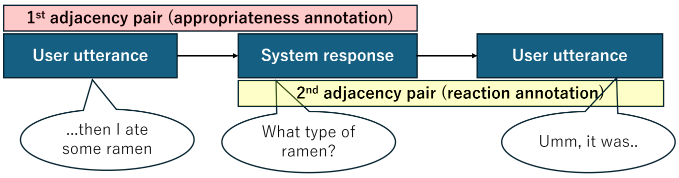

Fig. 2. Overview of annotation targets. Annotators used the 1st adjacency pair to assess appropriateness and the 2nd to evaluate user reaction.

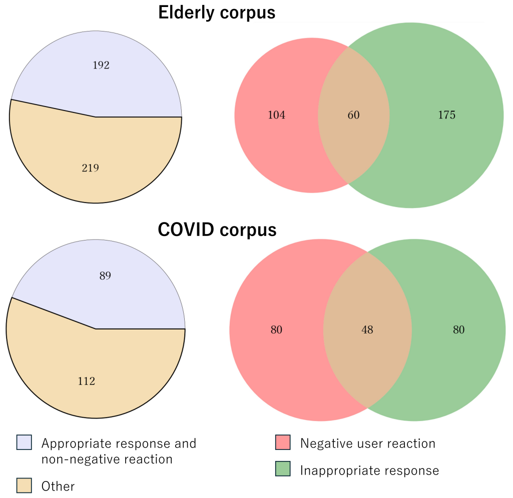

Fig. 3. Annotation results of the two corpora. The left side shows the distribution of system utterances which are both appropriate and followed by a user reaction which is not negative. The right side is the subset of "Other", showing utterances which have either a negative reaction, are inappropriate, or both.

Annotation of user reaction is toward the second adjacency pair, without considering the content of the response. Annotators were asked to observe the user's reaction and speech to determine if they were dissatisfied with the preceding response.

The elderly and COVID corpora consists of 411 and 201 non-generic responses respectively. Fig. 3 summarizes the distributions of reactions and appropriateness for non-generic responses. Our analysis shows that there are still a significant number of utterances which are appropriate but are followed by a negative reaction by subjects. We justify our focus on negative reactions as not completely caused by dialogue mistakes and therefore should be studied as a separate phenomena.

## V. Data modalities

In this section we describe the modalities which will be investigated in this work as predictors of negative reactions.

### A. Audio features

We extract pitch and power information using an online pitch tracker [21]. Audio is divided into system and user inter-pausal units (IPUs) with a 200ms segmentation for silence.

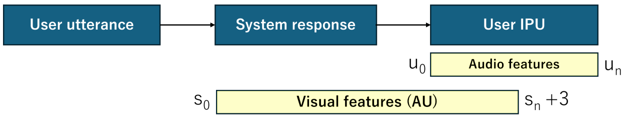

Fig. 4. Target time windows for extracting features.

All IPUs are transcribed and turn-taking was annotated to accurately handle backchannel utterances.

### B. Facial action units and head pose

We also used the well-known facial action units (AUs) [22] which have also been used in previous research for error detection [2]. These are a taxonomy of specific movements or features in the face and are extracted and classified for each frame of the video using the OpenFace toolkit [23]. There are two types of output - the presence of an action unit (AUP) and the intensity of an action unit (AUI). AUPs are a binary value while AUIs are scored on a 5-point scale, higher values having more intensity. We also used OpenFace to extract head position and rotation information for each frame.

### C. Linguistic features

We also used the capabilities of using automatic evaluation as a tool to predict reactions and appropriateness. For this task we used ChatGPT [24], asking the model to evaluate the system response on a scale of 1 to 10, using the following prompt with the previous user turn and system response:

> *"A and B are having a short dialogue in Japanese. Rate the appropriateness of B's response to A's talk on a scale of 1 to 10, with 1 being completely inappropriate and 10 being completely appropriate."*

We use this numerical rating directly as a feature in the model to capture negative reactions which are a result of inappropriate or incoherent responses.

## VI. Models

In this section we describe the construction of several models used in this work and their features sets.

### A. Windows of analysis

First we decide the target time window for analysis for prosodic and visual features, shown in Fig. 4. We denote the beginning and end of the system response as $s_0$ and $s_n$ and the next user IPU (not the entire turn) as $u_0$ and $u_n$. Note that in the case of overlapping turns it is possible that $s_n \geq u_0$.

Prosodic features are extracted from $u_0$ to $u_n$, as they are calculated only for the user's IPU. For the AU model the target window is $s_0$ to $s_n + 3$ to capture the visual features straight after the system begins its response.

TABLE I
Feature set used for models.

| Feature type | Description and no. of features |
|---|---|
| Audio (IPU) | turn switch time, IPU length (2) |
| Audio (prosodic) | mean + median power diff., unvoiced % (3) |
| Visual (AUs) | AUP percentages, AUI averages (35) |
| Visual (pose) | Total head position and rotation change (2) |
| Linguistic | ChatGPT rating (1) |

### B. Feature selection and extraction

For audio features we used turn switch time and the duration of the next user's utterance to capture hesitations and fillers which may indicate confusion. For prosodic features we extracted the difference in mean and median power of the next user IPU compared to the corresponding values over the entire session. The aim was to capture self-talk, which was found in negative reactions in previous work [5]. The percentage of the IPU that was unvoiced (no pitch detected) was also included to find disfluencies in user speech indicating confusion.

For action unit features, we used the the percentage of presence for each AUP and the average value of the AUIs over the target window. We also calculate the total movement of the head over the window (position and rotation) similar to previous work [5]. Table I provides an overview of the features used. In total there are 43 features.

We also attempted to use streamed audio as an input to fine-tuned transformer models, however these were unsuccessful. Therefore, in this work we focus on the relatively simpler statistical classifiers.

## VII. Results

We implemented logistic regression, random forests and XGBoost classifier models using the Python toolkit scikit-learn [25]. Each model was trained using 10-fold cross-validation and 100 trials on the dataset were conducted to account for differences in the folds. We tested different subsets of features: audio, visual and linguistic. Hyper-parameter selection through grid search was used for the random forest and XGBoost models. Default values for random forest were satisfactory but for XGBoost we twe set the hyper-parameters of `learning_rate`= 0.3, `max_depth`= 4, `min_child_weight`= 7 and `colsample_bytree`= 0.5. In this work we focus on improving precision to increase our confidence in a negative reaction prediction.

Models were trained on each corpus. We compare them to a baseline which randomly predicts the user reaction based on the distribution of the corpus. Results for the elderly corpus are shown in Table II. Logistic regression with all features was the best performing in terms of recall and F-score. However random forest and XGBoost models are better for precision. Multimodality improves the audio-only models.

Results for the COVID corpus are displayed in Table III. For this corpus using the ChatGPT evaluation improves the audio-only model, but adding visual features does not.

To understand the contribution of the features, we analyzed the models using Shapley Additive Explanations (SHAPs)[26]

TABLE II
PERFORMANCE FOR MODELS TRAINED ON ELDERLY CORPUS.

| Model | Precision | Recall | F-score |
|---|---|---|---|
| *Baseline* | 0.253 | 0.253 | 0.253 |
| **Logistic Regression** | | | |
| Audio | 0.446 | 0.558 | 0.496 |
| Visual | 0.340 | 0.567 | 0.424 |
| Audio + Visual | 0.42 | 0.606 | 0.496 |
| Audio + Linguistic | 0.433 | 0.558 | 0.487 |
| All | 0.426 | **0.606** | **0.500** |
| **Random Forest** | | | |
| Audio | 0.595 | 0.423 | 0.494 |
| Visual | 0.563 | 0.173 | 0.265 |
| Audio + Visual | 0.630 | 0.327 | 0.430 |
| Audio + Linguistic | 0.568 | 0.404 | 0.472 |
| All | **0.633** | 0.298 | 0.405 |
| **XGBoost** | | | |
| Audio | 0.547 | 0.394 | 0.458 |
| Visual | 0.459 | 0.163 | 0.241 |
| Audio + Visual | 0.569 | 0.394 | 0.466 |
| Audio + Linguistic | 0.512 | 0.423 | 0.463 |
| All | 0.577 | 0.394 | 0.469 |

TABLE III
PERFORMANCE FOR MODELS TRAINED ON COVID CORPUS.

| Model | Precision | Recall | F-score |
|---|---|---|---|
| *Baseline* | 0.398 | 0.398 | 0.398 |
| **Logistic Regression** | | | |
| Audio | 0.554 | 0.575 | 0.564 |
| Visual | 0.361 | 0.438 | 0.395 |
| Audio + Visual | 0.479 | 0.562 | 0.517 |
| Audio + Linguistic | 0.540 | **0.588** | 0.563 |
| All | 0.495 | 0.575 | 0.532 |
| **Random Forest** | | | |
| Audio | 0.533 | 0.500 | 0.516 |
| Visual | 0.500 | 0.250 | 0.333 |
| Audio + Visual | 0.510 | 0.312 | 0.388 |
| Audio + Linguistic | 0.594 | 0.512 | 0.550 |
| All | 0.480 | 0.300 | 0.369 |
| **XGBoost** | | | |
| Audio | 0.590 | 0.488 | 0.534 |
| Visual | 0.490 | 0.312 | 0.381 |
| Audio + Visual | 0.523 | 0.425 | 0.469 |
| Audio + Linguistic | **0.611** | 0.550 | **0.579** |
| All | 0.531 | 0.425 | 0.472 |

on the XGBoost model which included all the features. We calculated the average SHAP values for every sample over 100 trials. We generated a plot which shows the SHAP value of every sample in a corpus and orders it by the contribution towards the model. The SHAP summary for both corpora are shown in Fig. 5, with only the top 15 features displayed.

For both corpora the feature which contributes the most is turn switching time, with a higher time more likely to be a negative reaction. However many other features are important in one corpus but not the other. For example, the average intensity of AU20 (lip stretcher) is a strong contributor to the elderly model, but not for the COVID model. Only in the COVID corpus is the GPT rating score a contributing feature.

For a combined model we decided to only use the features that were present in both corpora's top 15 SHAP features. This resulted in 7 features as shown in Table IV.

There are two audio features (switch time and utterance duration) and five visual features. The AUs themselves are in all different parts of the face: eyebrow, eyelid, cheek, nose and chin. AU intensities over the target window appear to be more influential than just the presence of AUs.

TABLE IV
FEATURES IN THE TOP 15 SHAP VALUES FOR BOTH CORPORA.

| Feature | Rank (elderly) | Rank (covid) |
|---|---|---|
| Turn switch time | 1 | 1 |
| Utterance duration | 3 | 7 |
| AU4 (brow lowerer) average | 6 | 4 |
| AU9 (nose wrinkler) average | 5 | 11 |
| AU5 (upper lid raiser) average | 12 | 8 |
| AU6 (cheek raiser) average | 15 | 12 |
| AU17 (chin raiser) percentage | 13 | 15 |

TABLE V
PERFORMANCE FOR XGBOOST MODELS TRAINED ON BOTH CORPORA.

| Model | Precision | Recall | F-score |
|---|---|---|---|
| *Baseline* | 0.301 | 0.301 | 0.301 |
| Audio | 0.510 | 0.402 | 0.450 |
| Visual | 0.547 | 0.223 | 0.317 |
| Audio + Visual | **0.575** | **0.418** | **0.484** |
| Audio + Linguistic | 0.511 | 0.391 | 0.443 |
| All | 0.545 | 0.397 | 0.459 |
| Minimal | 0.562 | **0.418** | 0.480 |

Longer turn switching times are correlated with negative reactions, which is intuitive as users hesitate. However, for utterance duration the opposite pattern is true. This can be explained by the fact that many users acknowledged the system after it responded, with the first IPU being "*hai*". For negative reactions, it was more likely that this acknowledgment was not given as users hesitated before continuing their talk.

For the visual features, negative reactions were associated with a lower intensity of brow lowering (AU4), upper lid raising (AU5), and nose wrinkling (AU9). However, for the intensity of cheek raising (AU6) and the amount of chin raising (AU17) the patterns were different between corpora.

We trained XGBoost models on both corpora and included our minimal model which was trained using only the seven features shown in Table IV. Results are shown in Table V.

For the combined corpus, the model using all audio and visual features is the best performing. However the minimal model also shows similar performance with a fewer number of features. Fig. 6 also shows the generated SHAP summary plot for the minimal model. The order of feature importance of the previous SHAP analyses is maintained, with audio features being the most influential.

## VIII. DISCUSSION

Our results showed that a multimodal approach is needed for negative reaction prediction and that the two corpora required different types of models for improved performance over an audio-only model. For elderly people visual features had a large effect, while for the university students ChatGPT evaluations were more useful for predicting negative reactions. This discrepancy could be a result of either the demographics of the subjects or the nature of the talk.

Our SHAP analysis identified a subset of features which could best contribute to the model. According to our minimal

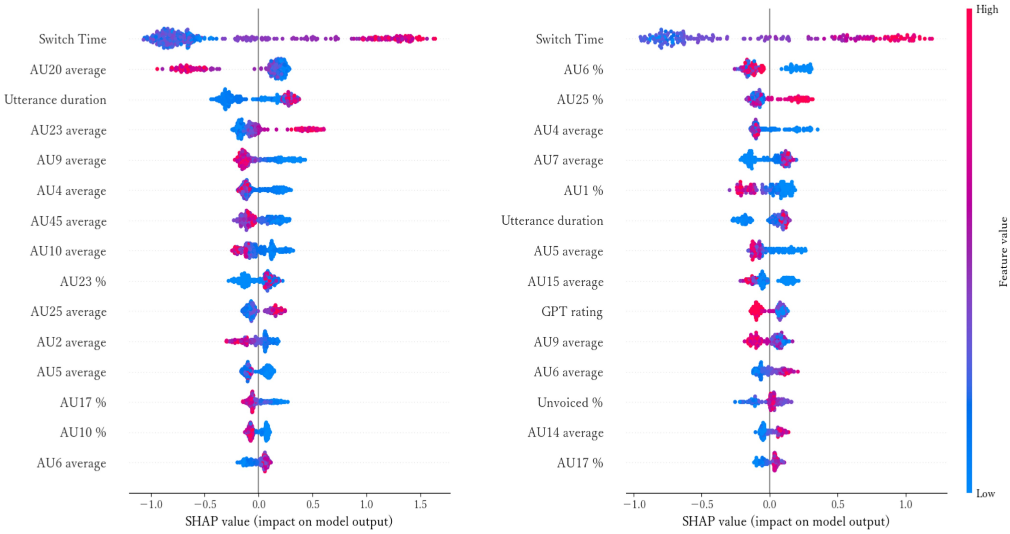

Fig. 5. SHAP value summary plots for elderly (left) and COVID (right) corpora. The top 15 features are listed in order of importance. Each point represents one sample with color signifying the relative value of the feature. Points towards the right side are more likely to be classified as negative reactions.

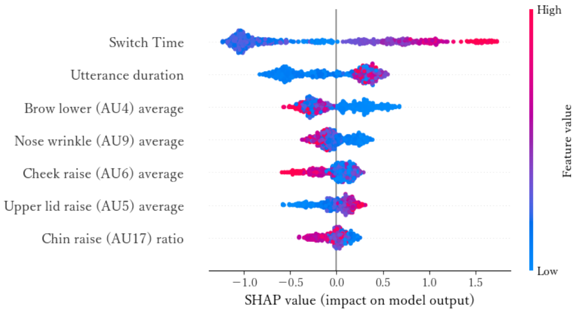

Fig. 6. SHAP summary for minimal model.

model, we can describe an exemplar of a negative reaction as when the user hesitates before responding to the system and continues with their talk without acknowledgment. They tend not to move the middle of their face, except to raise their eyelids. These features should not be seen as highly reliable, but it does allow us to provide explainability to our model and for future work this type of analysis is critical.

Our models were trained only on attentive listening data so we cannot prove that they would generalize to other open-domain dialogue scenarios. Attentive listening is quite different from mixed-initiative free talk where the user can ask questions to the system. It is possible that question-answering errors will be reacted to more strongly than those in our corpora.

There are several limitations in this work. Due to privacy concerns, the annotations of user reactions were conducted by only two experts. An approach with crowd-sourcing could produce different labels, although they may be more inconsistent. We also found that using powerful transformer models with audio streams was not successful even with fine-tuning, perhaps due to the comparatively low number of samples.

## IX. Conclusion

We used multimodal features to detect negative reactions in two attentive listening corpora. We extracted audio, visual and linguistic features and trained models on each corpora, finding that facial action unit features were important in the corpus of elderly subjects, while ChatGPT evaluation was more influential for university students. We then used SHAP values to identify seven key features and trained a combined model which had comparable performance to one which used significantly more features.

## Acknowledgment

This work was supported by JST Moonshot R&D Goal 1 Avatar Symbiotic Society Project (JPMJPS2011).

## References

[1] J. Ravishankar, M. Doering, and T. Kanda, "Zero-shot learning to enable error awareness in data-driven hri," in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 592–601.

[2] M. Stiber, R. H. Taylor, and C.-M. Huang, "On using social signals to enable flexible error-aware hri," in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 2023, pp. 222–230.

[3] D. Kontogiorgos, M. Tran, J. Gustafson, and M. Soleymani, "A systematic cross-corpus analysis of human reactions to robot conversational failures," in *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 112–120.

[4] M. Giuliani, N. Mirnig, G. Stollnberger, S. Stadler, R. Buchner, and M. Tscheligi, "Systematic analysis of video data from different human–robot interaction studies: A categorization of social signals during error situations," *Frontiers in psychology*, vol. 6, p. 931, 2015.

[5] M. Saeki, K. Miyagi, S. Fujie, *et al.*, "Confusion detection for adaptive conversational strategies of an oral proficiency assessment interview agent," English, *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2022-September, pp. 3988–3992, 2022.

[6] N. Li, J. D. Kelleher, and R. Ross, "Detecting interlocutor confusion in situated human-avatar dialogue: A pilot study," in *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, Potsdam, Germany: SEMDIAL, Sep. 2021.

[7] R. Cumbal, J. Lopes, and O. Engwall, "Detection of listener uncertainty in robot-led second language conversation practice," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, ser. ICMI '20, Association for Computing Machinery, 2020, 625–629.

[8] Y. Tsuta, N. Yoshinaga, S. Sato, and M. Toyoda, "Rethinking response evaluation from interlocutor's eye for open-domain dialogue systems," *arXiv preprint arXiv:2401.02256*, 2024.

[9] A. Yamaguchi, K. Iwasa, and K. Fujita, "Dialogue act-based breakdown detection in negotiation dialogues," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, pp. 745–757.

[10] R. Higashinaka, L. F. D'Haro, B. Abu Shawar, *et al.*, "Overview of the dialogue breakdown detection challenge 4," in *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, Springer, 2021, pp. 403–417.

[11] K. Tsubokura, Y. Iribe, and N. Kitaoka, "Dialog breakdown detection using multimodal features for non-task-oriented dialog systems," in *2022 IEEE 11th Global Conference on Consumer Electronics (GCCE)*, IEEE, 2022, pp. 352–356.

[12] T. Mori, K. Jokinen, and Y. Den, "On the use of gestures in dialogue breakdown detection," in *2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, IEEE, 2021, pp. 70–75.

[13] A. Ben-Youssef, C. Clavel, and S. Essid, "Early detection of user engagement breakdown in spontaneous human-humanoid interaction," *IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 776–787, 2021.

[14] N. Rollet and C. Clavel, ""talk to you later" doing social robotics with conversation analysis. towards the development of an automatic system for the prediction of disengagement," *Interaction Studies*, vol. 21, no. 2, pp. 268–292, 2020.

[15] A. Ben-Youssef, G. Varni, S. Essid, and C. Clavel, "On-the-fly detection of user engagement decrease in spontaneous human–robot interaction using recurrent and deep neural networks," *International Journal of Social Robotics*, vol. 11, no. 5, pp. 815–828, 2019.

[16] L. Tian and S. Oviatt, "A taxonomy of social errors in human-robot interaction," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 2, pp. 1–32, 2021.

[17] D. Kontogiorgos, A. Pereira, B. Sahindal, S. van Waveren, and J. Gustafson, "Behavioural responses to robot conversational failures," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 53–62.

[18] D. E. Cahya, R. Ramakrishnan, and M. Giuliani, "Static and temporal differences in social signals between error-free and erroneous situations in human-robot collaboration," in *Social Robotics: 11th International Conference, ICSR 2019, Madrid, Spain, November 26–29, 2019, Proceedings 11*, Springer, 2019, pp. 189–199.

[19] D. Lala, P. Milhorat, K. Inoue, M. Ishida, K. Takanashi, and T. Kawahara, "Attentive listening system with backchanneling, response generation and flexible turn-taking," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, K. Jokinen, M. Stede, D. DeVault, and A. Louis, Eds., Association for Computational Linguistics, Aug. 2017, pp. 127–136.

[20] K. Inoue, D. Lala, K. Yamamoto, S. Nakamura, K. Takanashi, and T. Kawahara, "An attentive listening system with android ERICA: Comparison of autonomous and WOZ interactions," in *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, O. Pietquin, S. Muresan, V. Chen, *et al.*, Eds., Association for Computational Linguistics, 2020, pp. 118–127.

[21] C. T. Ishi, H. Ishiguro, and N. Hagita, "Automatic extraction of paralinguistic information using prosodic features related to f0, duration and voice quality," *Speech Communication*, vol. 50, no. 6, pp. 531–543, 2008.

[22] P. Ekman and W. Friesen, *Facial action coding system: A technique for the measurement of facial movement*. Palo Alto, CA: Consulting Psychologists Press, 1978.

[23] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: An open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1–10.

[24] OpenAI, *Chatgpt (gpt4, march 2024 version)*, 2024.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[26] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., Curran Associates, Inc., 2017, pp. 4765–4774.