I Know Your Feelings Before You Do

I Know Your Feelings Before You Do: Predicting Future Affective Reactions in Human-Computer Dialogue

Yuanchao Li*
yuanchao.li@ed.ac.uk
University of Edinburgh
UK

Koji Inoue†
inoue@sap.ist.i.kyoto-u.ac.jp
Kyoto University
Japan

Leimin Tian†
leimin.tian@monash.edu
Monash University
Australia

Changzeng Fu
changzeng.fu@irl.sys.es.osaka-u.ac.jp
Osaka University & RIKEN
Japan

Carlos Ishi
carlos@atr.jp
ATR & RIKEN
Japan

Hiroshi Ishiguro
ishiguro@irl.sys.es.osaka-u.ac.jp
Osaka University & ATR
Japan

Tatsuya Kawahara
kawahara@i.kyoto-u.ac.jp
Kyoto University
Japan

Catherine Lai
c.lai@ed.ac.uk
University of Edinburgh
UK

ABSTRACT

Current Spoken Dialogue Systems (SDSs) often serve as passive listeners that respond only after receiving user speech. To achieve human-like dialogue, we propose a novel future prediction architecture that allows an SDS to anticipate future affective reactions based on its current behaviors before the user speaks. In this work, we investigate two scenarios: speech and laughter. In speech, we propose to predict the user's future emotion based on its temporal relationship with the system's current emotion and its causal relationship with the system's current Dialogue Act (DA). In laughter, we propose to predict the occurrence and type of the user's laughter using the system's laughter behaviors in the current turn. Preliminary analysis of human-robot dialogue demonstrated synchronicity in the emotions and laughter displayed by the human and robot, as well as DA-emotion causality in their dialogue. This verifies that our architecture can contribute to the development of an anticipatory SDS.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**.

KEYWORDS

emotion, dialogue act, interaction, laughter, spoken dialogue system

*Part of this work was done while working on the JST ERATO ISHIGURO Symbiotic Human-Robot Interaction Project at Kawahara Lab at Kyoto University.
†Authors contributed equally and are listed in alphabetical order.

CHI EA '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3585869

ACM Reference Format:
Yuanchao Li, Koji Inoue, Leimin Tian, Changzeng Fu, Carlos Ishi, Hiroshi Ishiguro, Tatsuya Kawahara, and Catherine Lai. 2023. I Know Your Feelings Before You Do: Predicting Future Affective Reactions in Human-Computer Dialogue. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3544549.3585869

1 INTRODUCTION

Spoken dialogue is a common and natural form of communication in human social interaction. Thus, we are witnessing a growing interest in advancing Spoken Dialogue Systems (SDSs) capable of delivering task-specific services, both in research and in commercial applications. For example, voice assistants, such as Amazon Echo or Google Home, are widely used for information queries in people's daily lives. Meanwhile, embodied SDSs, for instance the humanoid robot Pepper, have been deployed to enhance human workforce in application scenarios such as hospitality and elderly care. The majority of these SDSs, however, converse passively and utter words more as a matter of response than asking questions or leading the conversation on their own initiative. Furthermore, many existing SDSs are built on top of natural language processing and generation models developed with written text data, overlooking the rich conversational and affective phenomena unique to spoken dialogue, such as non-verbal vocalizations and affective bursts. As a consequence, current SDSs are often perceived as stagnant and mechanical. To mitigate this issue, researchers have been investigating various dialogue-specific behaviors, such as turn-taking, backchanneling, and laughter, as they have been found to serve important functions in conversation, including the marking of prominence, syntactic disambiguation, attitudinal reactions, uncertainty, and topic shifts [18, 27, 44, 45].

Studies on these dialogue behaviors usually include both detection and synthesis tasks. The detection task aims at predicting user behaviors from the received signals, while the synthesis task generates system behaviors. For the detection task, acoustic features such

as Mel Frequency Cepstral Coefficients (MFCCs) and prosodic features such as pitch, energy, and duration of pause-bounded phrases are typically used as cues for the detection of these human-like behaviors [14, 34, 40]. With the recent advance of deep learning techniques, such detection has started to become more robust and is being applied in real-world applications [15]. Compared to the detection task, the synthesis task is more challenging for two major reasons: 1) It depends on the accuracy of the detection task. If the user behaviors are classified incorrectly at the very beginning, the process for system behavior synthesis could become totally meaningless or even harmful to user engagement. For example, if the system detects turn-yielding cues but the user is actually holding the turn, then the user's speech will be interrupted by the system. 2) Unlike the detection task, where audio alone can achieve reasonable performance (although lexical cues often help), the synthesis task requires suitable generation of both acoustic and lexical behaviors. When synthesizing fillers and backchannels, the meanings largely depend on their morphological forms [15, 21, 22, 32]. To address this challenge, the majority of the synthesis task is still performed following a rule-based method. Take backchanneling as an example: First, the user's speech is converted into a sequence of words by Automatic Speech Recognition (ASR). Next, the focus word of the sequence is extracted. If the focus word matches any entries in the pre-built query-response database, the system generates a backchannel based on the query-response pair. Otherwise, the system generates a short backchannel, such as "Yeah" to indicate it is listening [21].

Such a detection-rule-synthesis process is a widely adopted operation in SDSs, yet it has several limitations. First, this can lead to delayed responses due to the time taken (often correlates to the duration of the input speech) to process the user's speech and synthesize suitable responses. Such delay can accumulate when there are several detection components (e.g., dialogue act recognition, emotion recognition, and turn-taking detection). Second, previous research in linguistics and communication theories suggests that human listeners have the ability to anticipate the interlocutor's behavior in real time based on the dialogue context and history [8, 37, 42], and such predictive power is core to human brains [2, 31, 35]. Furthermore, human listeners can start planning their responses or even talking before the interlocutor finishes, resulting in cooperative overlaps or appropriate interruptions that are key to establishing rapport and sympathy [10, 43]. Current SDSs, however, are incapable of exhibiting such anticipatory and collaborative dialogue behaviors.

To alleviate this problem, recent SDS research has started to investigate the feasibility of enabling the system to actively lead the conversation instead of behaving as a passive follower. Wu et al. [46] proposed a knowledge graph that sequentially changed the discussion topics following a given conversation goal to keep the dialogue as natural and engaging as possible. Besides, Lala et al. [17] and Li et al. [23] proposed attentive and proactive listening systems. These systems have a proactive initiator that can make the dialogue systems behave somewhat actively to ask a follow-up question related to the most recent topic or start a new topic. Moreover, proactive behaviors have proven helpful in rendering a more competent and reliable system that could ultimately lead to a more trustworthy interaction partner [16]. Nevertheless, these functions dealt with only the linguistic aspect (e.g., spoken content) without considering the paralinguistic aspect (e.g., affective expressions).

Therefore, we are motivated to build an anticipatory SDS by endowing it with the ability to predict the future affective reactions of the user. Inspired by findings in cognitive science that humans can anticipate certain future events, including affective ones [5, 7, 30], we propose an architecture that allows the SDS to mimic this human ability to predict affective reactions in the user's next turn based on its current turn. We consider two scenarios: speech and laughter, which are distinguished as two acoustic events (though co-occurrence also exists) [39, 43]. In the speech scenario, we look at the prediction of future emotions (valence and arousal). In the laughter scenario, we investigate the prediction of future laughter (occurrence and type). Moreover, we propose a self-correction and adaptation function that updates the future prediction model using the outputs of a recognition model on the user's speech. When the future prediction model has low confidence in its outputs, the recognition model generates outputs using the user's speech collected so far as ground truths to correct the prediction model. We conducted a preliminary analysis on human-robot dialogue, which confirms the feasibility of implementing the proposed anticipatory and adaptive architecture.

2 RELATED WORK

A dialogue is made up of a series of utterances, with the previous response determined by the history information [39]. Previous studies have found that the emotions of the interlocutors have a mapping relationship in human-human conversation. In persuasion dialogue, Acosta and Ward [1] discovered that the listener's dimensional emotion (valence, arousal, dominance) can be predicted from the emotion expressed in the immediately preceding speaker's utterance. Majumder et al. [26] demonstrated that in dyadic conversation, the listener's discrete emotion can be predicted by the context given by the preceding utterances and the emotion expressed. Such relationships have been considered when designing SDSs: by having a virtual or embodied agent mirror a user's emotions, i.e., the "affective mirror" [36].

Furthermore, emotion is affected by other aspects of dialogue, such as the Dialogue Act (DA), which represents the communicative function of an utterance. Despite the fact that there is a mutual influence between emotion and DA in conversations [6], understanding of such relationships is limited. However, a recent study has found that there is a clear temporal causal relationship between DAs and emotions, providing specific emotion-DA and DA-emotion pairs. For instance, *Happiness* of the speaker's utterance has a great chance of causing the DA of *appreciation* in the following response. The DA of *signal-non-understanding* and *backchannel-question-form* usually raise *surprise* [4]. Apart from emotions expressed through speech, affective bursts, especially laughter, are paralinguistic events that occur frequently in spontaneous dialogues [29, 43]. Laughter usually shows a contagious phenomenon that hearing laughter from others is known to trigger laughter in ourselves [38]. In dialogues, such a contagion is called "shared laughter" [9, 33]. A recent work has developed a shared laughter system that can decide whether to generate social laughter, mirthful laughter, or not laugh, depending on the detection of user laughter [12].

Based on these novel findings and developments, we can expect to advance SDSs by applying the causal relationships between emotion and DA, as well as the laughter mapping relationship, i.e., predicting the affective reactions of the upcoming user utterance to prepare its next response towards realizing proactive and affective behaviors. To the best of our knowledge, this is the first work to propose such an anticipatory SDS framework, which has the potential to lead to human-like affective dialogue capabilities in SDSs.

3 PROPOSED ARCHITECTURE FOR AN ANTICIPATORY SDS

3.1 The Speech Scenario

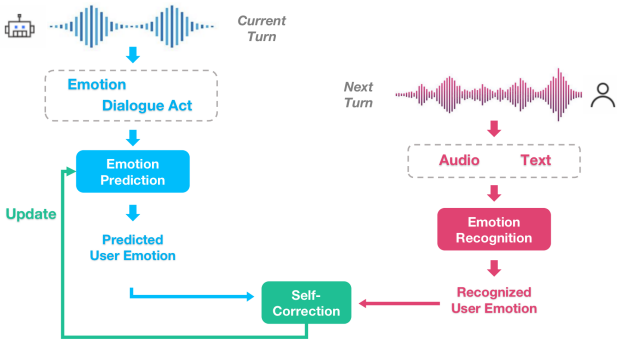
In the speech scenario, as shown in Fig. 1a and Algorithm 1, there are three major components in the proposed architecture: *emotion prediction*, *emotion recognition*, and *self-correction*. The emotion prediction model works as a future prediction function by taking the emotion and DA of the system's current turn as input and outputting an emotion as the prediction of the user's emotion in the next turn. When the prediction probability is high (i.e., the system is confident about its prediction), the system will take it as the user's future emotion and use it to help plan its next turn before the user speaks. Otherwise, the emotion recognition model starts to work by detecting the user's emotion once they finish speaking. The recognition result will be taken as ground truth to fine-tune the emotion prediction model if it has low confidence in its outputs and update the parameters of the prediction model via the self-correction function.

The emotion prediction model can be built by drawing on prior knowledge of emotion-emotion mapping [21] and the DA-emotion causal relationship [4]. We conducted a preliminary analysis demonstrating that in human-robot dialogue, the users indeed mimic the robot's emotion, but the correlations show different patterns in the arousal and valence dimensions (described in Sec. 4.2). Besides, there is limited work on the causal relationship between DAs and emotions. Thus, we aim to expand this research to formulate the emotion prediction problem as a logistic regression model:

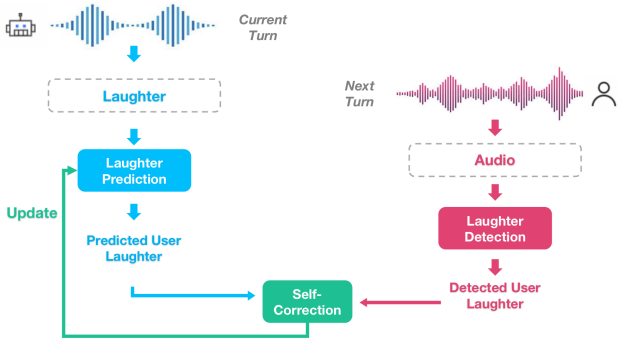
$$EM_{prd} = LR(EM_{cur},\ DA_{cur}) \tag{1}$$

where the EM_{prd} is the predicted user emotion in the next turn, and the EM_{cur} and DA_{cur} are the system's emotion and DA in the current turn, respectively. The regression model can be pre-trained on suitable corpora before being incorporated in the proposed architecture.

Compared to emotion prediction, emotion recognition has been well studied, but the majority of these works did not take into consideration the real-life environment (e.g., noise), which is a long-standing problem for SDSs [19]. Hence, we propose to incorporate text as additional information to tackle this problem. The text features are extracted from transcripts generated by the ASR component in SDSs, so the emotion recognition model needs to be robust to ASR errors. In a recent study, we developed a hierarchical cross-attention fusion model using both audio and text features for ASR error-robust emotion recognition [20], which can be adopted in the proposed architecture. When using ASR transcripts, this model performed similarly to when using ground-truth transcripts.



(a) The Speech Scenario.



(b) The Laughter Scenario.

Figure 1: The proposed architecture for anticipatory dialogue.

Algorithm 1: The speech scenario

Input: System's current emotion EM_{cur} and dialogue act DA_{cur}; Predicted user emotion EM_{prd}; Recognized user emotion EM_{rec}; Pre-defined probability threshold P_{thr1}

```
repeat
for system utterance do
    Generate EM_prd from EM_cur and DA_cur
    if Probability of EM_prd ≥ P_thr1 then
        Action // E.g., planning the system's next turn.
    else
        Generate EM_rec
        Update the emotion prediction model
    end if
end for
until no system utterance
```

The self-correction component follows the rule that when the confidence (i.e., probability) of the emotion prediction result is low, it starts to work by updating (i.e., fine-tuning) the prediction model using the outputs from the emotion recognition model as ground truths. Such a setting allows the emotion prediction component to dynamically adapt to the emotional expression habits of the human participants as the dialogue progresses. Like our human ability to make predictions, the longer the dialogue goes on, the more accurate the prediction becomes.

Algorithm 2: The laughter scenario

Input: System's current laughter LA_{cur}; Predicted user laughter LA_{prd}; Recognized user laughter LA_{rec}; Pre-defined probability threshold P_{thr2}

```
repeat
for system laughter do
    Generate LA_prd from LA_cur
    if Probability of LA_prd ≥ P_thr2 then
        Action // E.g., planning the system's next turn.
    else
        Generate LA_rec
        Update the laughter prediction model
    end if
end for
until no system laughter
```

3.2 The Laughter Scenario

Similar to the speech scenario, there are three major components in the laughter scenario, as shown in Fig. 1b and Algorithm 2: *laughter prediction*, *laughter detection*, and *self-correction*. Based on the system's laughter behavior in the current turn and the shared laughter relationship [12], the laughter prediction model predicts the occurrence and type of the user's laughter in the next turn. If the prediction probability is low, the laughter detection will work by detecting the type of laughter from the user's response and updating the laughter prediction model via the self-correction function.

The laughter prediction model can be built by drawing on recent work that detects the occurrence and type of the user's laughter to generate the system's laughter [12]. We can use this finding reversely by predicting the occurrence and type of the user's laughter based on the system's laughter, which is easy to manipulate in SDSs. The laughter detection model takes acoustic features (e.g., MFCCs) and prosodic features (e.g., pitch and power) as input and feeds them to a stacked recurrent neural network. The recurrent neural network will be implemented using the bi-directional gated recurrent unit, whose feed-forward processing can work in real time, which is essential for SDSs. The self-correction follows the same idea as the speech scenario by updating (fine-tuning) the prediction model when its prediction probability is low.

4 PRELIMINARY ANALYSIS

4.1 Corpora Description

Although there are existing emotional dialogue corpora, most of them are not suitable for the purpose of this work. For example, IEMOCAP [3] contains spontaneous dialogue sessions, yet the improvisation is limited to a set of hypothetical scenarios, which is different from open domain natural dialogue. SEMAINE [28] collected human-agent emotional dialogues, but the human users were not permitted to ask questions. The persuasive dialogue corpus used in [1] is not publicly available. Besides, none of the existing corpora contain sufficient occurrences and variations of laughter.

Therefore, we used two corpora from the JST ERATO ISHIGURO Symbiotic Human-Robot Interaction Project[1] that contain spontaneous dialogue and rich laughter. The corpus for the speech scenario consists of spontaneous dialogue between human participants and a teleoperated humanoid robot ERICA [11]. During data collection, ERICA was teleoperated by a human operator in a Wizard of Oz (WoZ) manner. The participants were students ranging from 18 to 22 years old. Each dialogue session contained two phases and lasted around 15 minutes, and six sessions were conducted. In the first phase, ERICA introduced herself and talked with the participants about their lives, hobbies, and future plans. In the second phase, they talked about robots, especially about ERICA herself. During the dialogue, the robot led the dialogue, and the participant acted as a "follower", which is the scenario we hope to apply our proposed architecture to. The emotions are annotated as *Valence*: -3 (extremely negative) to +3 (extremely positive), and *Arousal*: -3 (extremely passive) to +3 (extremely active).

The corpus for the laughter scenario was collected under almost identical conditions, except that the aim was to get the teleoperators and the participants to know each other quickly [13]. As a result, they behaved friendly by laughing frequently during such speed dating. Each dialogue lasts 10 to 15 minutes, and 82 dialogue sessions were conducted. The laughter was annotated as *Social laughter*, *Mirthful laughter*, and *No laughter*.

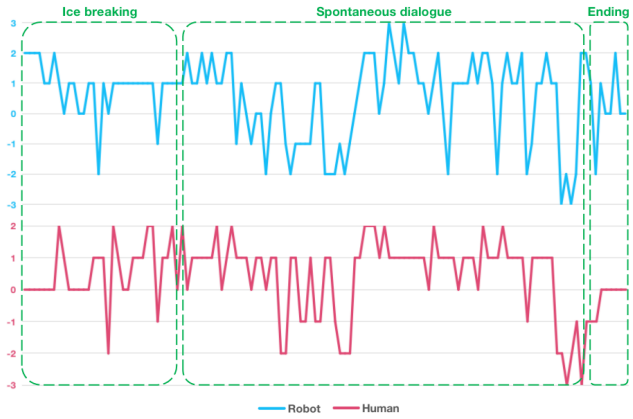


Figure 2: Valence of the robot's and human's speech during an example dialogue session shows a mimicry pattern.

4.2 Exploring the Feasibility of Emotion Prediction

To explore the relationship between the human participant's and the robot's emotions in dialogue, we analyzed the human-robot dialogue sessions from the first corpus. Our preliminary analysis found similar patterns in all six sessions. Because different sessions have different durations and different numbers of utterances, we could not average the emotion labels over the six sessions. Thus, we report one session containing 123 utterance pairs as an illustrative example to discuss our findings.

The valence patterns of the robot and participant are shown in Fig. 2. The dialogue can be roughly divided into three phases. We can see that in the spontaneous dialogue phase, the human participant's valence does mimic the robot's to a large extent, especially when the robot changes its valence significantly (e.g., from -2 to +2 and from

[1] https://www.jst.go.jp/erato/ishiguro/en/index.html

Table 1: Annotated excerpt for analyzing the influence of dialogue acts on emotions in dialogue.

Speaker	Transcript	Valence	Arousal	DA
Robot	Where are you from?	+1	+1	Wh-question
Participant	Tokushima, in Shikoku.	+1	+2	Statement
Robot	Fukushima?	+2	+2	Signal-non-understanding
Participant	No, Tokushima.	0	+1	Reject
Robot	Oh, Tokushima. I'm sorry.	-1	0	Apology

Table 2: How valence is related to dialogue context – the robot

When is the robot positive	When is the robot negative
1. In the initial greetings	1. The participant talking excessively
2. Introducing itself	2. Talking about the limitations of robots
3. The participant saying something amusing	3. Hearing about the participant's limitations
4. The participant feeling positive	4. The participant showing negative feelings
5. Introducing a new topic	
6. Praising the participant	
7. The participant answering questions correctly	
8. Asking a question and expecting a positive answer	

Table 3: How valence is related to dialogue context – the participant

When is the participant positive	When is the participant negative
1. In the initial greetings	1. Talking about their research
2. Talking about their background	2. Failing to explain something clearly to the robot
3. Starting a new topic	3. Feeling bored with a topic
4. Being praised by the robot	4. Describing vague topics (e.g., the future, job plans)
5. Saying something funny	5. Admitting they don't understand something technical
6. Explaining something vividly	6. Being asked a difficult question by the robot
7. Knowing a lot about something (e.g., robots)	7. Having to say something negative about the robot
8. Praising the robot	8. Being embarrassed or feeling bad for the robot
9. Being told their answers are correct by the robot	9. Talking about their own limitations

+1 to -3). Also, during the majority of the time in the spontaneous dialogue phase, the human valence is very close to its previous robot valence, showing a mimicry relationship. Note that, the human participant did not express extremely high valence (i.e., +3). This could be due to individual differences such as cultural background, personality, or expectation of the robot as a novel stimulus. The Pearson's Correlation Coefficient (PCC) of the valence pairs is 0.54 in the spontaneous dialogue phase, i.e., there is a moderate positive relationship between the human and robot valences. This suggests that it is possible to implement a mapping function between the human's valence and the valence expressed by the SDS. Note that, what we need to investigate is the correlation between emotions for a mapping pattern, not the exact values for classification, so we do not report accuracy or F1 scores. Even if the human participant's emotion values are completely different from the robot's, but highly correlated, e.g., [-3, -2, 1, 0, 1] and [-2, -1, 0, 1, 2], it still shows that the participant's emotion follows the robot, which could be used as a basis for implementing our proposed anticipatory SDS.

Interestingly, during the ice-breaking and ending phases, the human's valence hardly resembles the robot's valence with a PCC of 0.09 in the ice-breaking phase and 0.07 in the ending phase. After examining the video recording, we found that both parties were performing the greeting and leave-taking dialogue acts that they consider to be socially appropriate, instead of mimicking their dialogue partner's expressions. That is, emotions in the dialogue were influenced by DAs. Thus, we annotated DAs following the categorization by Stolcke et al. [41] to understand how they impact emotion mimicry.

An annotation excerpt from the ice-breaking phase is shown in Table 1. When the robot asked a "Wh-question" and the participant responded with a "statement", both valence and arousal display mimicry. However, when the robot expressed "signal-non-understanding" and the participant responded "reject", the valence has an obvious drop, but the arousal barely changes. This shows that unlike valence, arousal is less influenced by DAs.

In terms of arousal, we found significant mimicry in the participant's arousal toward the robot's (figure omitted for brevity). The PCC of the arousal pairs is 0.78 over the whole dialogue session, including the ice-breaking and ending phases. This demonstrates that the arousal behind the future response may be relatively easy to predict based on the current utterance.

Our preliminary analysis indicates that it is feasible to predict a future emotion from the current emotion and DA for implementing the prediction component of our proposed architecture. We also found that valence is related to contextual information, such as DAs and personal factors of the interlocutor. We summarized a set of observations on the relationship between dialogue context and the valence of the robot and of the human participant in Table 2 and

Table 4: Annotation for analyzing the laughter prediction.

Speaker	Transcript	Laughter acoustics	Laughter type
Participant	Although I studied only one night, I passed the exam. **Haha**.	Flat pitch, moderate power	
Robot	**Hehe**. I see		Social laughter
Participant	I was told the exam would be held the following week when I arrived. I had the wrong date. **Haha**.	Long duration, jittery, shimmery	
Robot	**Ufufufu**. I see.		Mirthful laughter
Participant	I studied hard for the exam but got a zero. I was very sad. **Haha**.	low pitch and power	
Robot	That is bad.		No laughter

Table 3, respectively. We expect that these observations can contribute to the future implementation of the proposed anticipatory SDS and to the broader research community.

4.3 Exploring the Feasibility of Laughter Prediction

We analyzed a publicized dialogue demo that was built upon the second corpus with the robot's dialogue system replacing the teleoperator[2]. The results are presented in Table 4. As shown here, the robot generated suitable laughter behaviors based on the acoustic features of the previous user laughter. When the user's laughter had a flat pitch and moderate power, the robot responded with social laughter. When the user's laughter had a long duration and was jittery and shimmery, the robot responded with mirthful laughter. The robot also "understood" not to laugh when the user laughed only to relieve embarrassment. Based on previous research on the contagious phenomenon of laughter [38], we aim to expand on the shared laughter research by adjusting the acoustic features of the laughter generated by the SDS, allowing it to predict the future laughter behaviors of the user in response to the system's laughter.

5 DISCUSSION

Our preliminary analysis of human-robot dialogue suggests that it is feasible to implement an anticipatory SDS that predicts the emotion and laughter of the user in the next turn using the current turn of the system. However, there remain open challenges in implementing the proposed architecture, which we will discuss in this section. Further, we will discuss potential application scenarios for the proposed architecture.

5.1 Implementation Challenges

The proposed architecture relies on accurate and robust recognition of emotions from speech, which remains an open challenge due to the variability in speech and emotion expression. The prediction component in the proposed architecture may not be applicable to all user turns, as humans can express emotions and laughter arbitrarily without considering the system's expression. In this case, the architecture can be "downgraded" with only emotion recognition and laughter detection working and the prediction and self-correction components frozen. Further, noise in real-world applications of SDSs can reduce the reliability of both the emotion recognition and laughter detection models. Therefore, we plan to incorporate other communicative modalities (e.g., text and vision) to further improve the proposed architecture [24].

[2]https://www.youtube.com/watch?v=6tMiWog4l00

5.2 Potential Application Scenarios

In social and open-domain dialogue, an SDS with our proposed architecture can generate appropriate conversational and affective behaviors, such as a backchannel "Yeah" or laughter, in real time, instead of having delayed turn-taking that may interrupt the user's next turn. Further, it is especially useful in scenarios where the SDS is expected to take initiatives and lead the conversation, such as in healthcare and education. For example, the SDS can adjust its generated emotions and DAs to support a user's emotion regulation process by eliciting certain emotions in people with depression or autism [25]. In education, the SDS can express emotions during collaborative problem solving with children to increase their participation in the learning activities and resulting learning outcomes, as Zhou and Tian [47] found that when the robots exhibited emotional expressions, participants were more likely to collaborate with them and achieve task success faster.

6 CONCLUSIONS

In this work, we propose an anticipatory SDS architecture that predicts the affective reactions of the user in a future turn using its behaviors in the current turn. We investigate its viability in both speech and laughter scenarios. Based on preliminary analysis of human-robot dialogue, we demonstrated that: 1) The emotion of a future turn can be predicted from the current turn. The arousal dimension has a significant mimicry relationship, in which the human user's arousal follows the robot's arousal during dialogue. The valence, however, is also related to the previous DA. 2) The laughter behavior of the human user in a future turn has a mapping pattern with the laughter behavior of the robot in the current turn. The preliminary analysis paves the way for our future research. In particular, we aim to identify the relationship between current DA and future emotion, as well as current laughter and future laughter to implement the emotion prediction model and the laughter prediction model of the proposed architecture. Moreover, we plan to include history information and dialogue context beyond the current turn to improve the prediction accuracy. Achieving anticipatory SDSs requires every individual component to be accurate and robust, as well as a seamless collaboration between the components. Thus, in the future, we plan to implement the complete architecture and evaluate its outcomes in user studies.

REFERENCES

[1] Jaime C Acosta and Nigel G Ward. 2011. Achieving rapport with turn-by-turn, user-responsive emotional coloring. *Speech Communication* 53, 9-10 (2011), 1137–1148.

[2] Luc H Arnal and Anne-Lise Giraud. 2012. Cortical oscillations and sensory predictions. *Trends in cognitive sciences* 16, 7 (2012), 390–398.

[3] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335–359.

[4] Shuyi Cao, Lizhen Qu, and Leimin Tian. 2021. Causal Relationships Between Emotions and Dialog Acts. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–8.

[5] Cristiano Castelfranchi and Maria Miceli. 2011. Anticipation and emotion. In *Emotion-oriented systems*. Springer, 483–500.

[6] Richard Craggs and Mary McGee Wood. 2003. Annotating emotion in dialogue. In *Proceedings of the Fourth SIGdial Workshop of Discourse and Dialogue*. 218–225.

[7] Richard J Davidson. 2001. Toward a biology of personality and emotion. *Annals of the New York academy of sciences* 935, 1 (2001), 191–207.

[8] Daiga Deksne and Raivis Skadiņš. 2021. Predicting Next Dialogue Action in Emotionally Loaded Conversation. In *Proceedings of the Future Technologies Conference*. Springer, 264–274.

[9] Sarah Estow, Jeremy P Jamieson, and Jennifer R Yates. 2007. Self-monitoring and mimicry of positive and negative social behaviors. *Journal of Research in Personality* 41, 2 (2007), 425–433.

[10] Ceenu George, Philipp Janssen, David Heuss, and Florian Alt. 2019. Should I interrupt or not? Understanding interruptions in head-mounted display settings. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. 497–510.

[11] 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.

[12] Koji Inoue, Divesh Lala, and Tatsuya Kawahara. 2022. Can a robot laugh with you?: Shared laughter generation for empathetic spoken dialogue. *Frontiers in Robotics and AI* (2022), 234.

[13] 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.

[14] Lakshmish Kaushik, Abhijeet Sangwan, and John HL Hansen. 2015. Laughter and filler detection in naturalistic audio. (2015).

[15] Tatsuya Kawahara, Takashi Yamaguchi, Koji Inoue, Katsuya Takanashi, and Nigel G Ward. 2016. Prediction and Generation of Backchannel Form for Attentive Listening Systems.. In *Interspeech*. 2890–2894.

[16] Matthias Kraus, Nicolas Wagner, and Wolfgang Minker. 2020. Effects of proactive dialogue strategies on human-computer trust. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 107–116.

[17] Divesh Lala, Pierrick Milhorat, Koji Inoue, Masanari Ishida, Katsuya Takanashi, and Tatsuya Kawahara. 2017. Attentive listening system with backchanneling, response generation and flexible turn-taking. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. 127–136.

[18] Divesh Lala, Shizuka Nakamura, and Tatsuya Kawahara. 2019. Analysis of Effect and Timing of Fillers in Natural Turn-Taking. In *Interspeech*. 4175–4179.

[19] Yuanchao Li. 2018. Towards Improving Speech Emotion Recognition for In-Vehicle Agents: Preliminary Results of Incorporating Sentiment Analysis by Using Early and Late Fusion Methods. In *Proceedings of the 6th International Conference on Human-Agent Interaction*. 365–367.

[20] Yuanchao Li, Peter Bell, and Catherine Lai. 2022. Fusing ASR outputs in joint training for speech emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7362–7366.

[21] Yuanchao Li, Carlos Toshinori Ishi, Koji Inoue, Shizuka Nakamura, and Tatsuya Kawahara. 2019. Expressing reactive emotion based on multimodal emotion recognition for natural conversation in human–robot interaction. *Advanced Robotics* 33, 20 (2019), 1030–1041.

[22] Yuanchao Li and Catherine Lai. 2022. Robotic Speech Synthesis: Perspectives on Interactions, Scenarios, and Ethics. *arXiv preprint arXiv:2203.09599* (2022).

[23] Yuanchao Li, Catherine Lai, Divesh Lala, Koji Inoue, and Tatsuya Kawahara. 2022. Alzheimer's Dementia Detection through Spontaneous Dialogue with Proactive Robotic Listeners. In *HRI*. 875–879.

[24] Yuanchao Li, Tianyu Zhao, and Xun Shen. 2020. Attention-based multimodal fusion for estimating human emotion in real-world HRI. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 340–342.

[25] Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2018. Eliciting positive emotion through affect-sensitive dialogue response generation: A neural network approach. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[26] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive RNN for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 6818–6825.

[27] Chiara Mazzocconi, Ye Tian, and Jonathan Ginzburg. 2020. What's your laughter doing there? A taxonomy of the pragmatic functions of laughter. *IEEE Transactions on Affective Computing* (2020).

[28] Gary McKeown, Michel F Valstar, Roderick Cowie, and Maja Pantic. 2010. The SEMAINE corpus of emotionally coloured character interactions. In *2010 IEEE International Conference on Multimedia and Expo*. IEEE, 1079–1084.

[29] Willem A Melder, Khiet P Truong, Marten Den Uyl, David A Van Leeuwen, Mark A Neerincx, Lodewijk R Loos, and B Stock Plum. 2007. Affective multimodal mirror: sensing and eliciting laughter. In *Proceedings of the international workshop on Human-centered multimedia*. 31–40.

[30] Maria Miceli and Cristiano Castelfranchi. 2014. *Expectancy and emotion*. OUP Oxford.

[31] Yukie Nagai and Minoru Asada. 2015. Predictive learning of sensorimotor information as a key for cognitive development. In *Proc. of the IROS 2015 Workshop on Sensorimotor Contingencies for Robotics*, Vol. 25.

[32] Ryosuke Nakanishi, Koji Inoue, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. 2019. Generating fillers based on dialog act pairs for smooth turn-taking by humanoid robot. In *9th International Workshop on Spoken Dialogue System Technology*. Springer, 91–101.

[33] Costanza Navarretta. 2016. Mirroring facial expressions and emotions in dyadic conversations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 469–474.

[34] Hiroaki Noguchi and Yasuharu Den. 1998. Prosody-based detection of the context of backchannel responses. In *Fifth International Conference on Spoken Language Processing*.

[35] Anja Philippsen and Yukie Nagai. 2018. Understanding the cognitive mechanisms underlying autistic behavior: a recurrent neural network study. In *2018 Joint IEEE 8th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. IEEE, 84–90.

[36] Rosalind W Picard. 2000. *Affective computing*. MIT press.

[37] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access* 7 (2019), 100943–100953.

[38] Robert R Provine. 1992. Contagious laughter: Laughter is a sufficient stimulus for laughs and smiles. *Bulletin of the Psychonomic Society* 30, 1 (1992), 1–4.

[39] Norbert Reithinger, Ralf Engel, Michael Kipp, and Martin Klesen. 1996. Predicting dialogue acts for a speech-to-speech translation system. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, Vol. 2. IEEE, 654–657.

[40] Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language* 67 (2021), 101178.

[41] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics* 26, 3 (2000), 339–373.

[42] Koji Tanaka, Junya Takayama, and Yuki Arase. 2019. Dialogue-act prediction of future responses based on conversation history. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. 197–202.

[43] Khiet P Truong and David A Van Leeuwen. 2007. Automatic discrimination between laughter and speech. *Speech Communication* 49, 2 (2007), 144–158.

[44] Nigel Ward. 1996. Using prosodic clues to decide when to produce back-channel utterances. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, Vol. 3. IEEE, 1728–1731.

[45] Nigel G Ward. 2019. *Prosodic patterns in English conversation*. Cambridge University Press.

[46] Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goals. *arXiv preprint arXiv:1906.05572* (2019).

[47] Shujie Zhou and Leimin Tian. 2020. Would you help a sad robot? Influence of robots' emotional expressions on human-multi-robot collaboration. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1243–1250.