The Acoustically Emotion-aware Conversational Agent with Speech Emotion Recognition and Empathetic Responses

Jiaxiong Hu,Yun Huang, Xiaozhu Hu, and Yingqing Xu, *Senior Member, IEEE*

Abstract—Emotion is important for the conversational user interface. In prior research, conversational agents (CAs) employ natural language process techniques to create affective interaction based on text. However, the use of acoustic features of speech for voice-based CAs is under exploration. This work presents an acoustically emotion-aware CA that enables speech emotion recognition and stylizes responses with empathetic feedback and interjections. We conducted a user study in which 75 participants interacted with the CA under emotion stimulating to evaluate their perceived emotional intelligence (PEI). Our results show that the acoustically emotion-awareness increased the participants' PEI of the CA, and empathetic responses from the CA helped alleviate some participants' negative emotions. Our work provides implications for designing future CAs with better PEI.

Index Terms—Human-centered computing, emotion in human-computer interaction, influencing human emotional states, intelligent agents

1 Introduction

Conversational agents (CAs) are recently popular in various applications such as personal fashion counseling [1], survey [2], [3], education [4], [5] and mental health [6], [7], [8], [9]. Thus, the emotional intelligence of CAs has become a popular research topic [10], [11], [12]. The original psychological definition of emotional intelligence (EI) involves human's ability to appraise and express emotions, regulate emotions, and utilize emotions [13]. In the context of Human-Computer Interaction, the perceived emotional intelligence (PEI) of a CA is evaluated based on the CA's ability to perceive user emotions (e.g., detecting and deciphering emotions from words and voice), use user emotions (e.g., leveraging the emotions to support cognitive tasks), understand emotions (e.g., comprehending emotions and knowing their triggers), and manage the emotions (e.g., regulating emotions) [11], [14]. It is also suggested that improving the PEI of CAs has manifold benefits, including but not limited to: enriching interpersonal relationships, increasing engagement, and enhancing user experience [12].

Conversational text is available for both text-based and voice-based CAs to enable PEI. Sentiment analysis of text is often applied to detect and understand users' emotions from their text input [15], [16] and to help users regulate emotions [17]. Text sequence-to-sequence models [18], [19] are also applied to help CAs generate emotionally appropriate responses by taking tone information into account [18]. Different chatting styles are further designed to improve users' PEI, e.g., a self-disclosing chatbot was found to improve people's perceived intimacy [20]. Besides textual content in speech, acoustic signals contain rich emotional information [21], [22]. Thus, compared to text-based CAs, voice-based CAs can explore more possibilities of emotional interactions via acoustic signals. For example, several speech emotion recognition (SER) algorithms are developed to detect users' emotions from acoustical cues, by labeling a voice input as a specific emotional category, e.g., happy, sad, or angry [23], [24], [25], or by predicting the valence and arousal of a voice input [26]. To respond to users' emotions, voice-based CAs have been relying on users' self-reported emotions [27] or conversational states such as "standby" or "listening" [28]. In human-human conversations, people show empathy–the ability to comprehend others' feelings and to re-experience them oneself–using emotive interjections (e.g., "WoW!") [29], [30]. Similarly, by inserting interjections and fillers in part of its utterances for conversational contexts like "to change the topic", the CAs are found to receive higher user ratings [31]. However, prior research has not systematically evaluated the effectiveness of leveraging SER and interjections to generate empathetic responses in improving the PEI of CAs.

Therefore, in this paper, we proposed the acoustically emotion-aware CA that combines the speech emotion recognition and the empathetic responses, i.e., *the SER-enabled empathetic responses*. Specifically, we embedded speech emotion recognition (SER) [23] in the design to better perceive user emotions, and we emotionalized the CA's responses by using interjections and empathetic feedback including praising, distracting, and reappraising to create a sense of empathy. Our previous work [32] evaluated people's PEI of the emotionally aware CA as bystanders by observing the video clips. Focusing on evaluating the proposed acoustically emotion-aware CA when the participant actually converses with the CA, we designed a new experiment.

- *Jiaxiong Hu and Yingqing Xu are with the Academy of Arts & Design, Tsinghua University, Beijing, China, 100084.*
E-mail: hujx19@mails.tsinghua.edu.cn and yqxu@tsinghua.edu.cn
- *Yun Huang is with the School of Information Sciences at the University of Illinois at Urbana-Champaign, Champaign, IL, 61820.*
E-mail: yunhuang@illinois.edu
- *Xiaozhu Hu is with the Department of Computer Science and Technology, Tsinghua University, Beijing, China, 100084.*
E-mail: huxz19@mails.tsinghua.edu.cn

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2022.3205919

We conducted a user study with 75 participants to evaluate the PEI of the CA by providing different chatting styles. During the experiment, the participants interacted with the CA when playing a computer game, which was designed to adjust the game's level of difficulty over time such that the CA could respond accordingly, e.g., speaking with empathy when detecting the participants' negative emotions and celebrating the participants' accomplishment when they were happy for finishing the game successfully.

This work made the following contributions. First, we explored the possibility of embedding speech emotion recognition (SER) into the design of a voice-based conversational agent (CA) and evaluated it. Our work provided empirical evidence that the PEI of a CA was influenced by how it responded to different emotional contexts. For example, when induced a positive emotion, the participants reported a higher PEI when the CA inserted empathetic interjections into their responses; however, when having a negative emotion, the participants reported a higher PEI when the CA supported reappraisal and distraction. Second, we examined the effects of people interacting with a CA with emotional intelligence on their emotion regulation. More specifically, chatting with the acoustically emotion-aware CA helped participants alleviate their negative feelings when playing a computer game. Third, our work contributed both design and practical implications for improving interaction with voice-based CAs.

2 Related Work

This section will first provide an overview of emotion-aware conversational agents (CAs). We highlight the importance of improving the perceived emotional intelligence (PEI) of CAs and briefly address the relevant works for both text-based and audio-based CAs. We then focus on presenting two major types of research: 1) speech emotion recognition (SER) for identifying emotions, and 2) emotion response generation for CAs to reply to users with emotional intelligence. We also present two research methods for evaluating people's perceived emotional intelligence of CAs. Finally, we propose our research questions.

2.1 Enabling Conversational Agents with Emotional Intelligence

R. W. Picard drew on affective computing in 1997, which appealed to the importance of the ability for computers to "have emotions" [33]. The Media Equation suggests that we tend to treat computers like real people and interact with computers socially, which means emotional communication matters in human and computer interaction [34]. The forms of computers have long changed, in that ubiquitous computing devices, especially mobile phones with conversational agents embedded, are becoming a trend. Conversational agents today have more social characteristics than computers in the 1990s [12], [35]. The importance of affective experience with CAs has been argued [36]. Ma et al. adapted the emotional intelligence model of psychology to human-agent interaction as perceived emotional intelligence, which entails perceiving, using, understanding, and managing emotions, providing a reference for evaluating how the user perceives the emotional intelligence of conversational agents [14], [37].

CAs with emotional intelligence play an important role in various scenarios such as education [4], [5] and mental health [6], [7], [8], [9]. For example, previous conversational agents for mental health [6], [7], [8], [9], either arranged emotion self-reporting or used sentiment analysis of the conversation text to address users' emotions. For text-based CAs, sentiment analysis is often applied. For example, based on a deep learning model, a conversational agent [18] can identify eight primary tones from the input text of customers, including empathetic, passionate, satisfied, polite, impolite, sad, frustrated, and anxious tones. Agents are usually embedded with machine learning models to detect users' emotions from text input [9], [17], [18]. Acoustic signals contain rich emotional information [21], [22]. Thus, to build a voice-based CA with emotional intelligence, SER is considered for emotion perceiving. In the following section, we will focus on reviewing SER technologies.

2.2 Recognizing Emotion in Speech

Databases, features, and classifiers are essential for SER [38]. Two major types of databases are classified by the type of emotion generation: simulated and natural emotion [38]. Actors and actresses act out the emotions in scripts or improvise to create utterances for simulated databases [39], [40], [41], [42]. The other type, such as Voxceleb, collected the emotional utterances from subjects in natural situations [43], [44]. Spectral and prosodic features such as MFCCs (mel-frequency cepstral coefficients) are frequently used [26], [45], [46]. As for classifiers, SVM was the choice in some early works [46], [47]. Neural networks, including the convolutional neural network (CNN) and the long short-term memory recurrent neural network (LSTM-RNN), are more effective for emotion recognition [23], [24], [26]. Only with SER, the CA cannot perform emotional intelligence, because the user receives no signal that the CA can perceive or express any emotions nor affect others' emotions. Therefore, CAs need to generate appropriate responses to address users' emotions to improve the PEI. In the next section, we will review technologies developed to respond to users with emotional sensitiveness.

2.3 CAs' Emotional Responses and Empathy

Generating response text with emotions to the user is a direct and effective approach. To generate a response with an appropriate emotion, Zhou, et al. and Song, et al. built emotional chatting machines with sequence-to-sequence models [19], [48]. By learning many pairs of requests and responses from natural corpus databases, the emotional chatting machines can generate a response with a possible emotion. The CA used polite responses to deal with the recognized user emotions [49]. Ma, et al. implemented a CA with two alternative response styles, dominant or submissive, to deal with verbal abuse from the user [14]. Empathy is considered the key point to connecting emotional appraisal and expression [13]. CAs that respond with empathetic utterances achieve better user satisfaction and engagement [50]. Participants in a previous study regarded the perceived empathy of the therapist chatbots as the best thing of the experience [7]. In learning scenarios, previous works indicated that virtual agents performing empathy help students relieve emotions of fear and persist in learning, and get more perceptions of presence [27], [51]. The empathic CA also has potential in mental health applications [52], [53]. To better explore the effects of CA's empathic responses to emotions on PEI in a user study, we implemented our proposed emotion-aware CA with practical approaches to response generation as follows. First, computer praise can be used to deal with positive emotions since it effectively improves user motivation and engagement [54], [55]. Then, to deal with negative emotions, two typical emotion regulation strategies of human are referenced: distraction and reappraisal [56], [57], [58], [59]. Consequently, guiding the user to distract from the negative feelings or reappraise negative stimuli can also help create empathetic CA conversations. Meanwhile, nonverbal cues are also considered in this work. No matter in which language, people express feelings and empathy with emotive interjections [29]. For conversational CAs, using interjections in responses to express emotions can provoke empathy [31], [60]. Therefore, we choose interjections as the nonverbal cues for emotional expression in this work. Moreover, we utilize SER to select appropriate interjections. Here, we define interjections as short nonverbal words or expressions that express a spontaneous feeling, e.g., an exclamation (wow!) or hesitation (um). Since open resources of mature emotional speech synthesis in Chinese are still unavailable, we exclude other manipulations of speech synthesis, such as pitch, prosody, and speed. This work focuses on acoustic cues, so we exclude emotional expression approaches in other modalities, e.g., facial expression and gestures [11], [61].

2.4 Evaluating the Perceived Emotional Intelligence (PEI) of CAs

In prior research, people have been asked to evaluate their perceived emotional intelligence of CAs as bystanders [14], [32]. For example, in a pilot study [14], researchers found it difficult to ask people to interact with a CA in a particular way. Especially if it is a new kind of interaction, people feel artificial and have difficulty immersing themselves in the scenarios. Therefore, they conducted a one-on-one video study with online questionnaires for the actual experiment. The subjects, as bystanders, perceived the emotional intelligence of the CAs in the video. However, an earlier experiment on socialbots [31] suggests that the role of the rater (as a bystander or as an active participant) is an important factor in assessing social dialogue. When people are immersed in a specific emotional state, their perception, attention, memory, and executive functions would be affected [62]. For instance, a previous study demonstrated the influence of a sad mood on memory in terms of emotional words and facial emotion recognition [62]. To better understand the effect of our design, we designed the experiment where the participant actively interacted with the acoustically emotion-aware CA.

The above literature review suggests that SER technologies for identifying emotions and technologies for responding with empathy are promising to improve the PEI of a CA. However, to the best of our knowledge, they have been investigated separately. Little is known about the effect of integrating them on people's perceived emotional intelligence of a CA, especially when people directly interact with the CA. In this work, we investigate the effectiveness of the CA integrating SER and empathetic response generation based on a lab study in which participants interact with the CA actually. To gain a deeper understanding, we also conduct a semi-constructed interview based on the user's experiences interacting with the acoustically emotion-aware CA. In particular, we propose the following RQs:

RQ1:*How will SER-enabled empathetic responses affect participants perceiving a voice-based conversational agent's emotional intelligence?*

RQ2:*What is the effect of interacting with a voice-based conversational agent that provides SER-enabled empathetic responses?*

3 The Design of the Acoustically Emotion-Aware Conversational Agent

In this section, we present the design of the conversational agent that can perceive and empathetically respond to a user's emotion in speech. We combined speech emotion recognition and strategy-based response generation, namely, the SER-enabled empathetic responses.

3.1 The SER-enabled Empathetic Responses

In the user study of this work, the user emotion was natural instead of acted out. Therefore, we employed the convolutional neural network [23] for SER and trained the model with Voxceleb [44] in this work. The model was widely applied and proved effective in classifying a speech segment as a discrete emotion label, such as happy. The SER component in this study had the same structure as the student network (a 12-layer CNN) proposed in [23], the input of which was the vectorized short-term amplitude spectrogram extracted from four-second segments of raw speech audio.

The original SER model output a vector with five dimensions representing the confidence (from 0 to 1) of five emotion types. We reconfigured the SER to classify the speech as positive, negative, or neutral. Thus, the response generation script would not be too complicated. To specify, we firstly defined the mapping from discrete emotion label to a two-dimension valence-arousal coordinate, such as happy=(1,1), neutral=(0,0), sad=(-1,-1), surprise=(0,1) and angry=(-1,1). Then, the five coordinates were averaged as one with the previously output confidence of five emotion types as weights. Based on the averaged coordinate, the SER component applied the rule shown in Fig. 1 to output a final result of positive, negative, or neutral.

As Fig. 2 shows, the overall CA system in this study was rule-based. A dialogue manager retrieved responses from the database according to the context and the SER output. In this way, we could easily control the conversations and better investigate the effectiveness of the SER-enabled empathetic responses. After recognizing the user's emotional state with the SER component, the CA applied the empathetic strategy to react to the recognized emotion, replying with positive emotion when recognizing positive or with

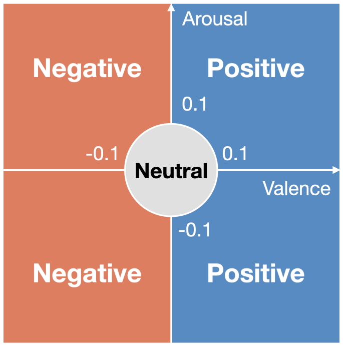


Fig. 1. The SER component applied the rule to convert the two-dimension valence-arousal coordinate into positive, negative, or neutral. The area of neutral was set small to make the SER more sensitive.

TABLE 1
One CA responses neutrally and three CAs use different empathetic response strategies: With Interjections (*WI*), Empathetic Feedback (*EF*), and With Interjections + Empathetic Feedback (*WI+EF*).

CA Conditions	SER	Response Strategy
Control	Without SER	Default Response
WI	With SER	Default Response + Interjections (e.g., "wow", "ha-ha", "um...")
EF	With SER	Default Response + Empathetic Feedback (praising, distraction and reappraisal)
WI+EF	With SER	Default Response + Interjections + Empathetic Feedback

negative emotion when recognizing negative. We developed two response approaches: (1) inserting a short interjection in the original response content and (2) replacing the original neutral response content with empathetic content.

The two approaches can be applied simultaneously or individually, as the example shown in Table 2. All the responses were stored in a database, and the CA retrieved responses from it according to the recognized emotion. For example, when assisting the user in math tasks and receiving the user's speech "Next one...", the CA is supposed to respond neutrally with "OK" and continue the tasks. However, the speech emotion recognition outputs a negative result. Thus, the CA inserts the interjection "Hmm..." in the beginning to show its awareness of the user's negative feelings or/and gives empathetic feedback "You're doing great!" to help the user regulate emotion.

3.1.1 With Interjections

According to the semantic meaning of the user's speech, the CA retrieves an initial neutral response from the database. If the SER outputs positive, the CA inserts a positive interjection at the beginning, e.g., "Ha-ha"; if negative, an interjection of hesitation, e.g., "Hmm...". The CA reciprocates the user's emotion by expressing the same emotion because mimicry is a way for humans to express empathy [63], [64]. In this study, we pre-wrote responses with appropriate interjections in the database for the CA to retrieve.

3.1.2 Empathetic Feedback

For the empathetic feedback, three strategies are considered. When the user performs well and feels positive, praising improves engagement and helps the user persist in tasks [54], [55]. Sometimes, when the user is frustrated, two strategies are available to help the user regulate emotion: distracting from the current task and reappraising the negative stimuli to guide the user to a more positive emotional state [56], [57], [58], [59]. Same as above, we pre-wrote responses according to the strategies in the database for the CA to retrieve.

4 Method

Prior work conducted video studies where participants observed human-agent conversations [14], [32]. However, the emotional state of the active participants affects cognition [62] and is suggested to influence their evaluation of the CA [31]. To further explore the experience of conversing with the CA as active participants, we designed this experiment.

We install the CA in the lab for participants to interact with (**RQ1**). It takes approximately 25 minutes to complete the user study. Meanwhile, to discover the effect of interacting with a CA with SER-enabled empathetic responses (**RQ2**), we focus on examining two aspects: perceived emotional intelligence improvement and emotion regulation support. A semi-constructed interview based on the experiences of interacting with the CA in the lab is conducted to understand people's preferences for the emotionally intelligent CA.

Since we need to track how active participation in real conversations and emotional states affected PEI. Therefore, instead of asking participants to act out a certain emotion, we invited them to play a game that elicits various negative and positive emotions. Similar to the emotional inducement in early HCI studies [65], [66], [67], the configurable difficulty of this game was the emotion stimulation. Specifically, increasing the difficulty induces negative emotions, and lowering it induces positive emotions. The CA intervened in the participant's emotions by chitchatting each time mistakes were made in the game. It was motivated by real-world scenarios where conversational agents are designed to provide real-time support for people when doing online learning or taking exams. The participant rated the PEI of the CA according to the interaction experience. Participants were grouped into different CA conditions. Thus, the study was a between-subject design. To avoid practice effects, each participant completes the game only once.

4.1 Conversational Agent Settings

To compare different CA designs in PEI, we implemented CAs in four conditions (see Table 1). Three CAs use different empathetic response strategies: With Interjections (*WI*), Empathetic Feedback (*EF*), and With Interjections + Empathetic Feedback (*WI+EF*), while the control CA only provides neutral responses. For emotion perceiving, *WI*, *EF*, and *WI+EF* enable the SER to recognize user emotion in speech while the control condition doesn't detect any emotional signals. For emotion expressing, the control condition only generates default neutral responses; *WI* inserts emotional interjections at the beginning of the responses; *EF* applies

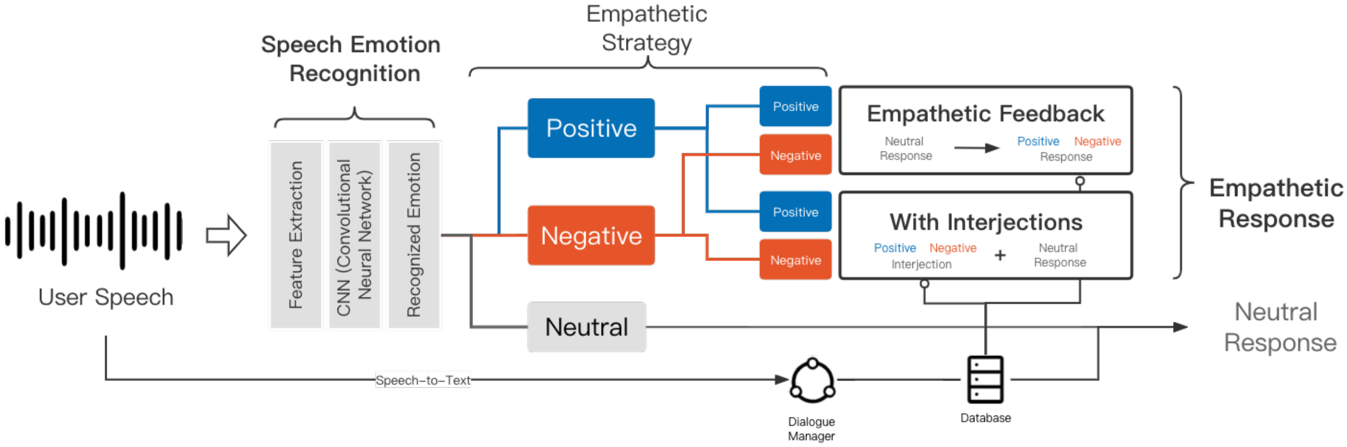


Fig. 2. The framework of the acoustically emotion-aware CA. The CA recognizes the user's emotional state (positive, negative, or neutral) from the speech. Based on the empathetic strategy, the CA uses empathetic feedback and interjections that express similar emotions to generate empathetic responses.

empathetic feedback; *WI+EF* combines both strategies of *WI* and *EF*. During the study, the group information was hidden from participants so that participants could only infer the characteristics of the CAs based on the presented interaction.

4.2 The Game Of Emotion Elicitation: The Keypad of Errors

Figure 3 (a) shows the user interface of the game. Participants are asked to accomplish calculation tasks presented in the task field. The calculation is so difficult that the participant has to ask the CA for answers. No matter which CA is assigned, participants are informed it is named "Anna", and the interface is shown in Figure 3 (b). Given the answer by Anna, the participant entered it through a numeric keypad with 10 number buttons, a delete button and a submit button "Go!". However, the numeric keypad was configured as full of errors. Sometimes it entered a random number into the answer field instead of the expected one, and sometimes the button did not work. It made the game difficult to play. This method was inspired by the deliberately slow computer-game-interface [65] and the Pacman game with randomly missed key presses [67]. In order to increase participants' mental load, a time limitation was set, and the counting down timer was under the keypad.

Besides providing answers to calculation tasks, Anna helped alleviate the participant's negative emotions by chitchats. Each time the participant submitted a wrong answer (caused by the "broken" keypad) or consumed all the given time, Anna intervened and asked questions about the participant's feelings about the game. Anna gave different feedback based on the participant's answers and emotional states. For example, "Did you find the task difficult?" Anna asked. "Yes, it was a little bit hard. (Positive)" the user said. "Oh, but you sound very confident!" Anna replied. Please note that the chitchats were implemented based on a script with 12 questions (see supplementary materials). We started with simple questions as we wanted to focus on evaluating participants' responses under different CA conditions. Without understanding the simple scenarios and starting with complex dialogue scenarios might introduce other variables (e.g., understanding complex semantic meaning) that are hard to control. In addition, Anna in all conditions also responded to the participant's unintentional utterances during the tasks, such as complaints or laughter. It was in Anna's manual mode, which will be introduced in Section 4.4. For example, during the task, the CA in *WI+EF* would respond "Um...Don't worry" when the participant complains or sighs; respond "Hey! You can do it!" when the participant laughs. While the control CA only responded "You can do it". Figure 2 provides examples of Anna's ways of responding in different conditions.

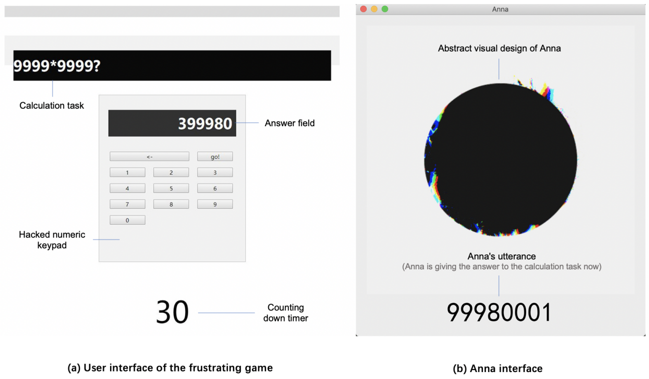


Fig. 3. (a) is the game's interface, which consists of the task field, a keypad with a display, and a timer at the bottom. (b) is the abstract visual design of Anna, and the number in the text field is the answer to the task in the game. The text field also displays other utterances from Anna.

When the correct answer was submitted in time, the next task started. The participant must accomplish five tasks continuously. The opportunity of errors increased with the game progress. If the participant failed any task, the progress returned to zero. The game consisted of two parts (see Figure 4). The first part lasted about 10 minutes.

The tough tasks were impossible to complete because

TABLE 2
An example of Anna's chitchat script. Anna in different conditions would respond differently to the opinion and emotions of participants. The control condition maintains neutral responses. *WI* inserts an interjection that expresses the parallel emotion as the SER result. *EF* praises the user when the SER result is positive. When it is negative, *EF* distracts the user's attention from the negative emotion or helps the user reappraise it.

Question	User's Opinion	Condition	Anna's Response When The SER Result Is	
			Positive	Negative
Did you find the tasks difficult?	Yes.	Control	OK, thanks for the reply.	OK, thanks for the reply.
		WI	Oh, thanks for the reply.	Hmm... Thanks for the reply.
		EF	But you sound confident.	No worries, many people found it difficult.
		WI+EF	Oh, but you sound confident.	Hmm... No worries, many people found it difficult.
	No.	Control	OK, response recorded.	OK, response recorded.
		WI	Wow, response recorded.	Hmm... response recorded.
		EF	You sound confident.	Is that so? You did a good job, though.
		WI+EF	Wow, you sound confident.	Hmm... You did a good job, though.

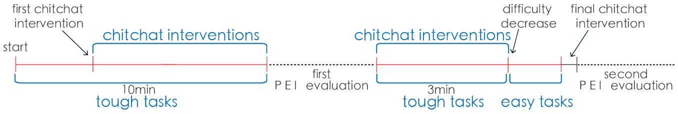


Fig. 4. The game consisted of two parts. Including two PEI evaluations, the total game time was around 20 min, and Anna asked the participant 12-13 questions in total for the chitchat.

we deliberately adjusted the difficulty. And the participant could only finish four tasks at most, which helped elicit negative emotion. During the first part of the game, each participant must fail several times and chat with Anna. After 10 min, participants were asked to evaluate the PEI of Anna. Then the game went to the second part. In the first three min of the second part, tasks were identically difficult. Participants might still fail and chitchat with Anna. After three min, the game restarted and became much easier (error opportunity=0). This was designed to stimulate participants' positive emotions so that we could check the PEI ratings in both positive and negative emotion scenarios. Please note that positive emotion was always stimulated after negative emotion, as we meant to evaluate the effect of the CA designs on the same situation, e.g., helping people transition from potential negative emotions to positive ones. In the second part, participants finished all five tasks in a short time. Then Anna influenced the participant's emotions with the final chitchat. At last, participants were asked to rate the PEI of Anna again with the questionnaire as described in the following section.

4.3 Measurement

The mouse click pressure during the game was tracked to confirm the effectiveness of emotion stimulation. One reason for choosing the mouse click pressure is that participants clicked the mouse through the whole game so that the variation could be recorded without distraction. The other reason is that the mouse is usually utilized to measure the emotional state of the user [68], [69], and Kirsch's work found that the user reacted to a negative valence stimulus with a harder press on the index button of the mouse [70]. Therefore, we set up a mouse with a pressure sensor under the mouse's index button to track participants' dynamic finger pressure.

Aside from the mouse click pressure, we also provided each participant with a questionnaire to evaluate the PEI of the CA in four aspects: perceiving emotions, using emotions, understanding emotions, and managing emotions. We adapted the Perceived Emotional Intelligence Questionnaire [14] as follows. Each question was answered with a 5-point Likert scale (1=totally disagree). According to the collected data, this questionnaire demonstrated good internal consistency with an overall Cronbach's coefficient alpha of 0.92. The corrected-item total correlation coefficients of each item with the total of the remaining items ranged from 0.88-0.92.

Questions of the perceived emotional intelligence questionnaire: (The agent can...)

- Convey a sense that it listens openly to participants' emotions. (Perceiving Emotion)
- Convey a sense that the agent can feel what the user is feeling. (Using Emotions)
- Respond empathetically to the user. (Understanding Emotions)
- Help the user regulate emotions, reduce negative emotions or keep posit

4.4 Apparatus

Figure 5 shows the experimental setting in the lab. Two laptops were installed in this study. A 15″ ASUS ran the game, a 15″ MacBook presented Anna, and a local GPU server ran Anna's SER and the control system. The cameras and speakers of the two laptops were enabled during the study. A Blue Yeti[1] microphone received the voice input. Participants used their right hands to control a mouse with a pressure sensor inside. An Arduino Leonardo[2] read the data from the pressure sensor and sent it to the laptop. The text-to-speech API of iFLY TEC[3] was used to generate the voice of Anna.

Anna had two modes: automatic and manual. Most of the conversations were the chitchat questions and answers, which were in automatic mode. This part was initiated by

[1] https://www.bluedesigns.com/products/yeti/
[2] https://www.arduino.cc/en/Main/arduinoBoardLeonardo
[3] https://www.iflytek.com

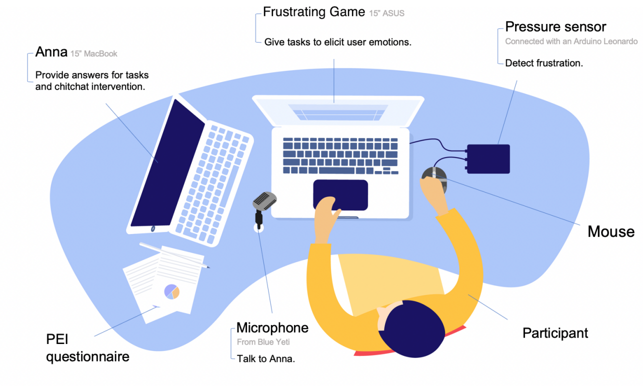

Frustrating Game 15" ASUS
Give tasks to elicit user emotions.
Anna 15" MacBook
Provide answers for tasks and chitchat intervention.
Pressure sensor
Connected with an Arduino Leonardo
Detect frustration.
Mouse
PEI questionnaire
Microphone
From Blue Yeti
Talk to Anna.
Participant

Fig. 5. The experimental settings.

the task failures and automatically done with the SER component and voice activity detection. Each time the participant failed the task, Anna started the chitchat intervention. The voice input was first analyzed by the SER component and speech recognition. Then gensim-based sentence similarity algorithm [71] was used to predict the participant's opinion. According to the SER result, the participant's opinion and the CA condition setting, Anna retrieved a preset response. Anna's basic speech recognition and speech synthesis were supported by iFlytek [72].

The other conversations were in manual mode, including the participant requiring the calculation task results and Anna responding to the participant's unintentional utterances. For the calculation task, we manually controlled Anna since speech recognition's inaccuracy might cause task failures and affect the PEI rating. Thus, an experimenter monitored the participant and sent the correct answers for the tasks to Anna via a socket. A few responses to the unintended utterances of the participant were also manually retrieved. All the responses were prepared based on the strategies mentioned in Section 3.

4.5 Participants

A total of 75 participants (aged 19-73 years, M=24.5, SD=7.78), including 45 females, were recruited for the study. We posted recruitment on social media, and all the participants registered voluntarily. They are all Chinese native speakers, so the CAs used in the study produce spoken Chinese. Please note that the examples of conversations and interviews are all translated into English throughout this paper (the original Chinese contents are in the supplementary material). We surveyed participants on their usage of conversational agents. 69% of participants have used a conversational agent (a voice assistant), with 52% of them using their CA at least once a week. Since it was a between-subject experiment, participants were randomly divided into four experimental conditions, as shown in Table 3.

Experimenters were provided with instructions and set up the devices for the participant. During the experiment, each participant was isolated in a room. Each participant evaluated the PEI of Anna. In the end, participants were invited to an interview.

TABLE 3
The tested conversational agents of four conditions were randomly assigned to participants.

Condition	N	Male	Female	Average Age	S.D. of Age	Interviewed
Control	19	8	11	24.1	6.32	15
WI	15	5	13	24.8	7.54	11
EF	22	9	10	24.6	11.0	19
WI+EF	19	8	11	24.4	4.59	13

Ethics. Because this work studied users' interaction with a conversational agent with emotion-related content, ethical issues have been carefully addressed. Prior work [73] outlined the minimum ethical standards for using conversational agents in mental health support. We set up our study according to the safety standards as follows. First, participants were informed that they were talking to a robot. Before the study, each participant read the instructions and consented to participate in the experiment voluntarily. Second, even though the risk of experiencing strong emotions during the study was kept minimum, one experimenter monitored participants via the webcam in case of any emergency, and participants could withdraw their consent and quit anytime. Third, the total time duration of the experiment was limited to prevent over-reliance. Though little private information was involved in the study, participants were also anonymized. Most importantly, the elicited negative emotions were like those ordinarily encountered in daily lives (e.g., when people play online games or take exams).

4.6 Interview

After the experiment, participants were eligible to be invited to a 20-minute semi-structured interview. The interview was designed to collect feedback about the CA and the participant's preferences for interacting with the emotionally intelligent CA. Moreover, it helped us understand which parts of the CA affected participants' PEI evaluation (**RQ1**) and if/how the CA helped them navigate different affective states when playing the game (**RQ2**). We also surveyed participants' expectations of CAs with SER-enabled empathetic responses to refine the design implication.

4.6.1 Interview Protocol

During the interview, we confirmed with participants Anna's (the assigned CA) helpfulness for the game and her acoustical emotion sensitivity. Specifically, we asked why they thought the CA could or could not perceive and respond to their emotions. The answers from participants should be affected by the settings of the CA in different conditions. Based on participants' own experience of interacting with the CA and the experiment experience, we asked them if the CA can perceive emotions effectively, whether they preferred the conversational agent to perceive their emotions, and why. To inform future designs of CAs with SER-enabled empathetic responses, we asked what emotional topics participants would like to talk about to the CA and when they would express emotions to the CA. A total of 58 participants accepted the invitation and finished the interview, as shown in the last column of Table 3).

4.6.2 Interview Data Analysis

We used thematic analysis for the qualitative data [74]. First, all the interviews were audio recorded and transcribed for analysis with the interviewees' consent. Four researchers revised the raw text transcribed with the speech recognition by iFlytek [72]. Then researchers coded the data in a code book. The code book contained 170 unique codes, e.g., "hearing my interjections", "expecting comforts when getting upset", "hoping she would laugh", "talking to her when alone", and "sharing personal topics". All the codes were grouped into six themes: impression of Anna, PEI of Anna, individual use of voice assistants, expected emotional intelligence, motivations of emotional interaction, and preferred topics of emotional interjection. Finally, we read the interview excerpts from each theme to make sure they were coherent with the theme.

5 Results

The results showed that SER-enabled empathetic responses increased participants' PEI of the CA (**RQ1**). Aside from the PEI improvements, interacting with the CA also alleviated people's negative emotions (**RQ2**). Through the interview, we found the results consistent with the quantitative analysis. Further, the interview results revealed people's preference for using the emotionally intelligent CA and sharing feelings with it.

For the general usability, we asked participants to describe the helpfulness of the CA. And 43 (74% of 58) participants reported that the CA effectively supported their tasks in the interview. "It helped me calculate at least." P11 said, since the CA that served P11 was in the control condition, i.e., without SER-enabled empathetic responses. However, P41, served by the CA in *EF* condition, said, "It was helpful. The responding speed was good. And it could recognize my emotions". Since we used the sentence similarity algorithm to classify participants' intentions, several error words did not cause wrong intention prediction. Therefore, no participant reported mistakes in speech recognition. Overall, the performance of SER was good, with an accuracy of 71.12%. The SER evaluation was based on human raters annotating the participants' speech from the experiment recordings. We also asked participants to appraise the CA's overall emotional intelligence according to the CA's responses since the individual emotional experience was subjective. Only three participants (P51, P58, P68) in CA conditions reported emotion recognition error occurrence. E.g., P51 (*EF*) reported a mistake of SER: when she expressed a little bit of anger, her CA still said "You did great!" to praise her. Though noticing the inaccuracy, P51 thought the CA was "better than many other voice assistants" since the CA's function of emotion perceiving.

5.1 Improving Perceived Emotional Intelligence (RQ1)

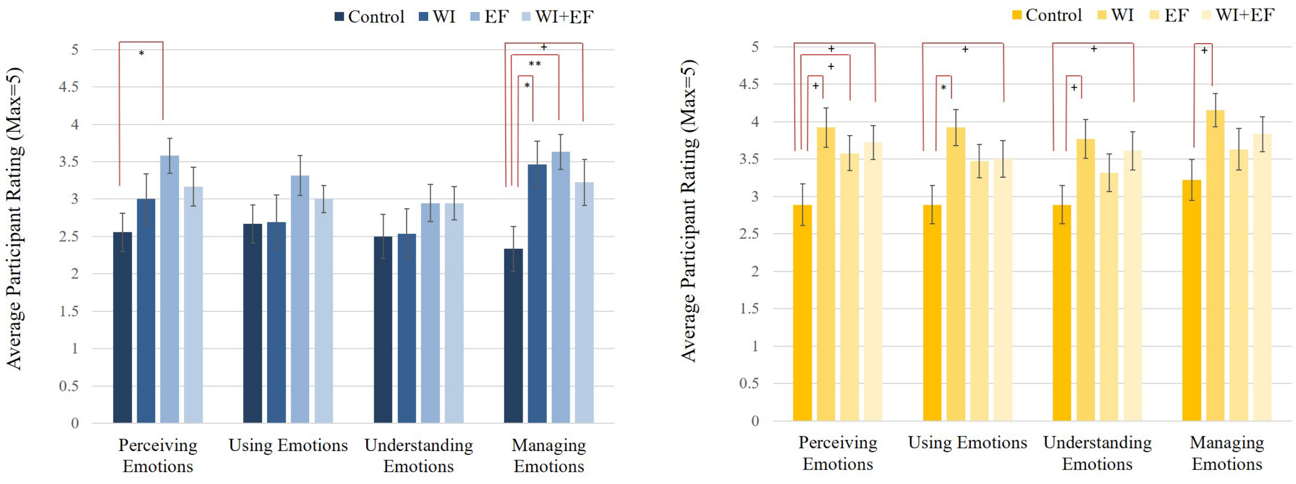
For the PEI data, we checked the normality with the Shapiro-Wilk test. The results did not confirm the normality of the data (p<0.05), which rejected the use of ANOVA. Thus, we ran a non-parametric test. It was a between-subject design, so the Kruskal-Wallis test was employed for the main effect analysis. The Mann-Whitney test with Holm Bonferroni correction was for the post-hoc analysis.

Before the data analysis, the PEI data of six participants were excluded because of software errors during the experiment (labeled in the metadata). Then, we checked participants' emotional states to validate the emotion stimulation of the game. We compared participants' emotional states between the first and the second PEI evaluation by comparing their self-reported emotional states on a scale from 1 to 5, with one being very unhappy and five being very happy. The average rating after tough tasks, was 2.78 (SD = 1.11) and the average rating after easy tasks was 3.68 (SD = 1.08). We checked the normality of the emotion rating data and found the normality was violated. Thus, we conducted a Wilcoxon-signed rank test and found that participants' emotion ratings were significantly lower when they first filled the questionnaire (p <0.05). This suggested that participants' emotional states changed from negative to positive significantly. We also validated the emotion ratings in each condition separately and found the same pattern. As P72 mentioned in the interview, *"The second round was much better, but the first round did annoy and frustrate me."* Thus, we concluded that our emotion stimulation for participants was efficient.

The Kruskal-Wallis test results revealed the main effect of the CA's SER-enabled empathetic responses on the PEI (p<0.05) in managing emotions of the first evaluation and using emotions of the second evaluation. Likewise, the post-hoc test results showed that the CAs significantly improved only in parts of the PEI ratings compared to the control CA (see Figure 6). The average PEI ratings of condition *WI*, *EF*, and *WI+EF* were higher than the control condition. In addition, each participant went through 12.23 (SD = 1.06) conversation turns with Anna. One conversation turn was defined as 1) Anna asked one question, 2) the participant answered, and 3) Anna responded. Comparing among *WI*, *EF*, and *WI+EF*, no significant difference was found.

Quantitative PEI rating results showed that the CA conveyed a sense that it listened to participants' emotions and responded empathetically. Feedback from interviews was consistent with quantitative results. 33 (77% of 43) participants in CA conditions (*WI*, *EF*, and *WI+EF*) said they did regard the CA as emotion perceivable. Their feedback verified the effects of our empathetic response generation design with SER. For instance, P31 (*WI*) said, *"I think she could feel and respond to some of my emotional changes during the tasks"* And P53 (*EF*) also considered the CA's responses were *"consistent with my understanding of emotion."* The CA also contributed to human-likeness as P72 (*WI+EF*) mentioned, *"It felt like Anna (the CA) was more human-like. It felt like she could sense my emotions, which affected my mood."*

However, the other 10 (23% of 43) participants said they overlooked the CA's ability to perceive emotion. From their feedback, they explained three reasons. 1) The participant did not express their emotion in the experiment. P67 (*WI+EF*) said, *"The agent might not be able to perceive my emotion because I didn't express much of my emotion in the experiment."* 2) The response pattern was monotonous. P39 (*EF*) said, *"I felt it didn't understand me. It just replied with some*



(a) The First PEI Evaluation

(b) The Second PEI Evaluation

Fig. 6. The average participant PEI ratings of the three CAs and the control CA (in real experimental settings) in terms of perceiving, using, understanding and managing emotions. +p<.1, *p<.05, **p<.01, ***p<.001; with standard error. (a) was measured when the game was difficult. (b) was measured when the game was easy.

universally applicable sentences." 3) The speech synthesis was toneless. P46 (*EF*) mentioned, *"I don't think it has emotional intelligence because its tone is too stiff."* Additionally, eight (60% of 15) participants in the control condition denied the CA's emotional intelligence as expected, but the other six (40% of 15) participants affirmed it to some extent. P1 (control) mentioned, *"On some level, chatting with the agent helped me calm down. Though its stiff responses, I thought it could perceive my emotion."* A possible reason was that chatting with Anna prevented participants from accumulating more negative emotions.

Feedback on the interjection use was also collected in the interview. Many participants liked the idea. For example, P23 (*WI*) said, *"It did perceive my emotion. For example, it told me not to worry, so I knew it was perceiving my emotions. I heard some interjections like 'oh, aha'."* P60 (*WI+EF*) noticed the CA saying "Hee hee", and it made her feel the task more like a game and alleviated her stress. Using interjections also contributed to human-likeness as P58 (*WI+EF*) said, *"It acted as a guide during the experiment. It replied with interjections like 'haha', so I felt it was more than a cold machine."* P54 (*EF*) even thought it hardly perceived emotions because of the CA (*EF*) not using interjections to express. P15 (control), however, thought some interjections sounded *"silly"*.

5.2 Alleviating Negative Emotions (RQ2)

During the experiment, participants conversed with the CA under emotion stimulation. The chatting styles of the CA helped participants allay the stimulated negative emotions. Instead of ignoring user emotion like the control condition, the CA in other conditions actively recognized emotions in utterances of the participant and generated responses that were related to the emotion. We analyzed the average mouse click pressure differences with and without chitchat intervention to check the CA's effect on negative emotion alleviation. During the first round of tasks, Anna did not chitchat but only provided answers. From Anna starting the first chitchat intervention to the first part ending, the participant was intervened by Anna with chitchat.

Distracted by the chitchat from the negative emotion, the participant clicked the mouse gentler. The average click pressure decrease of control, *WI*, *EF*, and *WI+EF* are respectively 1.51 (SD = 8.26), -38.59 (SD = 11.31), -6.78 (SD = 11.06) and -10.28 (SD = 8.40). The click pressure of participants in the control condition went higher while all those in the experiment condition kept going down. In particular, the decrease in WI condition was the most significant, which means SER-enabled empathetic responses had a main effect on the mouse click pressure decrease (p<0.05). Moreover, the post-hoc analysis revealed that *WI* weakened the mouse click pressure with the best efficiency when the participant was emotionally negative during the game.

The mouse click data proved the CA's effect in alleviating negative emotions, especially *WI*. The mouse click pressure decrease of *WI* differs from other conditions. One possible reason is that interjections are not commonly heard when interacting with CAs in the real world. So, the CA speaking with empathetic interjections attracted people's attention. Eight participants reported in the interview that they noticed the CA speaking with empathetic interjections. E.g., P30 (*WI*) noticed the CA laughed and showed compassion. However, only three of the eight were in condition *WI+EF*. Compared to *WI*, the effect of *WI+EF* was less significant, perhaps because the responses provided in the *WI+EF* condition was longer, which made interjections less noticeable. In addition, Participants' negative emotions were relieved when their unintended utterances were responded to. E.g., when P62 (*WI+EF*) *"heaved a sigh, the agent comforted me with some words."* Though the mouse click data did not reveal, a participant in the control condition reported the CA's support of emotion regulation in the interview. P1 (control) mentioned, *"Chatting with the agent helped calm down."* We attributed it to the unintentional distraction from the negative emotional stimuli. Because whenever participants had chitchat with Anna or their unintended utterances were responded to, they were momentarily distracted from the frustrating tasks.

According to the interview results, 43 (74% of 58) participants appreciated a future CA that can perceive and respond to their emotions. The reasons were as follows. 1) The ability to perceive perceiving and respond to emotion improved the overall performance of the CA. P2 said, *"As a voice assistant, the better it understands my emotions, the more it can improve its performance."* P3 said, *"For example, when you need some smooth music to deal with your anxiety, the agent should perceive your anxiety and play the right music for you without you speaking out the command. This saves time."* 2) Emotionally intelligent CAs offered a better user experience. P23 said, *"When I'm feeling down, I'd like to hear some words of comfort rather than a cold machine-like response."* P4 said, *"I hope the CA could perceive my emotions because that makes it more like a real person."* P58 and P72 also mentioned *"human-likeness"*. 3) Emotionally intelligent CAs helped *"to smooth my mood"*, as P20 said. Likewise, P50 also said, *"The agent should be able to perceive and respond to emotions. For example, it should relieve my anxiety and frustration."*

This proved that there would be great opportunities for CAs with SER-enabled empathetic responses. Several participants (P24, P3, P42) expressly indicated the tendency to express emotions to a CA primarily when they felt *"lonely"*, *"bored"*, or *"sad"*. Additionally, they expected that the CA would understand and respond to their emotions. They prefer the CA that acknowledges and accepts their emotions, especially the negative ones they feel uncomfortable sharing with friends and families because of potential judgment or criticism from other humans (P37, P38, P39, P49). Compared to humans, CAs sometimes appear more neutral and tolerant listeners. Previous research has suggested that the chatbot's understanding of one's emotions may create a feeling of acceptance and belonging in users [12]. This feeling is vital in establishing rapport between CAs and users. Furthermore, appropriate emotional detection and response generation contribute to that feeling. The strategies we adopted for the CA were efficient in dealing with people's negative emotions. Further investigation into these strategies will help improve the PEI of CAs and create better human-agent interactions.

However, 15 (26% of 58) participants expressed their concerns about emotionally intelligent CAs. 1) The privacy concerns. P5 mentioned, *"I will use it only if the agent can guarantee to protect my privacy."* 2) Worrying about the CA being too sensitive. P66 mentioned, *"I would be uncomfortable if it makes me feel like I'm being manipulated."* 3) Worrying about mirroring emotions. P75 said, *"I don't want feedback that mirrors my emotional state. Sometimes, people don't want feedback after expressing their emotions. Getting sympathy from others when I'm feeling down can be frustrating."*

In summary, the CA with SER-enabled empathetic responses was perceived to have more emotional intelligence than the CA without SER-enabled empathetic responses. The two PEI evaluations suggested that the effects of empathetic feedback and interjections varied in different contexts. Participants had more negative emotions when evaluating PEI for the first time as the game became difficult, so the empathetic feedback outperformed in perceiving and managing emotions. On the other hand, the second PEI evaluation was taken after the game became easier to play. Using empathetic interjections was more effective in the four aspects of PEI. Further, the mouse pressure data suggested that SER-enabled empathetic responses of the CA alleviated the user's negative emotions. Last, the interview result revealed people's preferences for the more emotionally intelligent CA, the tendency to express emotions to the CA, and concerns about the emotionally intelligent CA.

6 Discussion

This section discusses our findings, insights into participants' interaction with the CA, and design implications for future voice-based CAs.

6.1 The Effect of the SER-Enabled Empathetic Responses

To maximize customer acceptance of the voice-based CA, manufacturers often use a neutral-to-slightly-positive tone without considering speech emotion. Our results showed that the SER-enabled empathetic responses improved participants' overall PEI and perceived positive impacts on helping them ease negative feelings.

First, using interjections in the CA's responses helped improve the PEI. From the qualitative feedback, we found two possible explanations. 1) interjections in the CA's responses were quite noticeable. Few speech synthesis systems used interjections [31]. Thus, participants changed their impression of the CA as more human-like when hearing interjections from a CA. 2) the interjection was considered as a confirmation that the CA correctly comprehended the user's emotion. Since interjections were more effective in improving PEI in positive emotional contexts than negative, we addressed two possible explanations. 1) some interjections like "wow" are beneficial for expressing enthusiasm and intense emotion in human speech [75]. 2) according to the broaden-and-build theory of positive emotions [76], under the influence of positive emotions, people have wider perceptual access and semantic reach [76]. Nonverbal information like interjections were more minor details compared with the verbal contents. This suggested that positive emotions helped the interjections more likely to access people's minds.

Second, empathetic feedback was found effective in improving PEI. In our work, there were three effective emotion regulation strategies for empathetic feedback: praising when positive and distracting or reappraising when negative. Our findings showed that enabling these strategies at the appropriate moment can enhance PEI. For example, when the next task was about to start but the CA detected negative emotions from acoustic cues, the CA would recommend a short timeout to distract the participant from the negative stimuli. Then the CA helped the user reappraisal negative stimuli. When mentioning the task difficulty, the CA helped the participant reappraise the task as an unimportant test. The CA of *EF* praised the user when a positive emotion was recognized. In the interview, P42 (*EF*) said that praising was expected when she felt happy, e.g., P49 (*EF*) was praised by the CA as self-confident because of her positive tone. Distracting the user from negative emotional stimuli is also effective. As for human, attention altering (distraction) and cognitive change (reappraisal) are typical ways to regulate negative emotions [56], [57], [58], [59].

6.2 Bystander vs Active Participant

Studying bystanders is "safer" since it involves less emotion than directly engaging in the study. Nevertheless, if the "patterns" and claims are the same, then the method can be used in other settings. Many new technologies are evaluated from direct stakeholders–users' perspectives, but studying bystanders is also important. Recent HCI works studied voice-based agents used by multiple users at different locations, e.g., home [77], when people can switch their roles between bystanders and active participants. In our study, the emotional interaction was triggered by their own emotional expression. Participants were elicited emotions from negative to positive by the experiment tasks. Five participants reported in the interview that they had not expressed obvious emotions during the experiment. Two participants reported that they had focused too much on the task and ignored the emotional feedback. This suggests that the stimulated emotion is natural compared to the acted emotion if we instruct participants to act. Thus, our result can be extended to many scenarios where the user speaks to the CA with a specific tone or in a certain emotional state. To induce more emotions from the user, it is important to clearly state the emotion perceiving and responding of the CA, just like other capabilities of the CA. When people get used to expressing emotions to the CA, there would be much more possibilities for human and agent interaction.

6.3 Design Implications

Our findings made several design implications for improving user interaction with voice-based CA.

First, a few participants mentioned that SER was not always accurate when predicting their emotions, though their final PEI was increased due to the effective response strategies. Prior works verified the user tolerance for errors of recognition techniques [78], [79]. Our participants might have a similar tolerance for SER errors, though we did not explicitly track individual predictions. In this work, the CA did not display the SER results in front of participants. Instead, it used euphemistic terms, including emotional interjections and emotion regulation strategies to respond. To improve user experience, future design may consider exploring different approaches and evaluate their impacts on user experience: 1) showing the SER results to the users, such that users may help correct the prediction results; and 2) avoiding presenting the SER results, so that users may focus on the conversation instead of being distracted by a less accurate emotion prediction result.

Second, more response strategies need to be explored. For example, generate responses that distract or reappraise negative stimuli. Reappraisal and distraction are typical human emotion regulation strategies [56], [57], [58], [59]. These were proved to be helpful for participants in regulating their emotional states. However, these strategies were evaluated in the game contexts and needed to be evaluated in more contexts. Meanwhile, an emotionally intelligent CA needs to use flexible response strategies to deal with users' different emotional states. Simultaneously using multiple strategies might lead to wordy responses, as mentioned in the results. Our findings showed that using interjections outperformed in positive emotional contexts, while sentiment adjustment worked better in negative emotional contexts. Future work may explore the effective strategies under diverse emotional contexts and prioritize the strategies in order of effectiveness.

Third, future CAs may consider building speech synthesis systems with natural interjections. Inspired by our results in Figure 6, we found that using interjections is comparatively more effective in dealing with positive emotions. However, drawn from the interview, we found that the toneless synthesized speech was blamed. Therefore, future voice-based CAs may prepare natural interjection records with the professional voice actor/actress.

7 Limitations and Future work

Some limitations of this study are discussed in this section. This paper does not address long-term use. To verify the long-term use effects of the proposed system, a longitudinal study is needed in future work [20]. In this study, the SER algorithm we implemented in this experiment only analyzed the acoustic features of the user's speech. Although the algorithm's ability to detect emotions might be more robust with sentiment analysis of the text, we decided not to include it to avoid the effects of the error words by automatic speech recognition. Aside from using interjections and emotion regulation support, more strategies to generate empathetic responses for CAs can be explored in future work. For example, there is a large amount of research on improving the emotional representations in speech synthesis technology as reviewed in the survey [80]. With more open resources for emotional speech synthesis, prosody will be used to convey emotions in the CA's response generation in the future. Additionally, multi-modal feedback design, including audio, facial, physical touch and gesture, is potentially beneficial to our framework [11], [81]. At last, as the response generation mentioned in this study is rule-based, we consider the end-to-end model for the future CA design, e.g., the sequence-to-sequence model trained with post-response corpus [19].

8 Conclusion

To improve the perceived emotional intelligence of conversational agents, we proposed a CA with acoustically emotion awareness and responding. We presented evidence to support whether the CA design can enhance PEI from the perspective of active participants. Then we discussed the factors affecting PEI. Next, we provided the design implication of combining speech emotion recognition and empathetic response generation. In particular, we offered evidence from the user study that empathetic response generation with SER, a) using empathetic and emotional interjections in responses and b) using empathetic feedback such as praising the user, distracting or reappraising negative stimuli, can both effectively contribute to the emotional

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2022.3205919

intelligence of the agents. As such, we advocate considering speech emotion recognition and empathetic response generation as an acoustically emotion-aware design for conversational agents.

REFERENCES

[1] K. Vaccaro, T. Agarwalla, S. Shivakumar, and R. Kumar, *Designing the Future of Personal Fashion*. New York, NY, USA: Association for Computing Machinery, 2018, p. 1–11. [Online]. Available: https://doi.org/10.1145/3173574.3174201

[2] S. Kim, J. Lee, and G. Gweon, "Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.

[3] Z. Xiao, M. X. Zhou, Q. V. Liao, G. Mark, C. Chi, W. Chen, and H. Yang, "Tell me about yourself: Using an ai-powered chatbot to conduct conversational surveys with open-ended questions," *ACM Trans. Comput.-Hum. Interact.*, vol. 27, no. 3, Jun. 2020. [Online]. Available: https://doi.org/10.1145/3381804

[4] E. Ayedoun, Y. Hayashi, and K. Seta, "Communication strategies and affective backchannels for conversational agents to enhance learners' willingness to communicate in a second language," in *Artificial Intelligence in Education*, E. André, R. Baker, X. Hu, M. M. T. Rodrigo, and B. du Boulay, Eds. Cham: Springer International Publishing, 2017, pp. 459–462.

[5] R. Kumar, H. Ai, J. L. Beuth, and C. P. Rosé, "Socially capable conversational tutors can be effective in collaborative learning situations," in *Intelligent Tutoring Systems*, V. Aleven, J. Kay, and J. Mostow, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 156–164.

[6] A. Ghandeharioun, D. McDuff, M. Czerwinski, and K. Rowan, "Emma: An emotion-aware wellbeing chatbot," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 1–7.

[7] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial," *JMIR mental health*, vol. 4, no. 2, p. e19, 2017.

[8] M. Lee, S. Ackermans, N. van As, H. Chang, E. Lucas, and W. IJsselsteijn, "Caring for vincent: A chatbot for self-compassion," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–13.

[9] J. Huang, Q. Li, Y. Xue, T. Cheng, S. Xu, J. Jia, and L. Feng, "Teenchat: a chatterbot system for sensing and releasing adolescents' stress," in *International Conference on Health Information Science*. Springer, 2015, pp. 133–145.

[10] M. Ivanović, M. Radovanović, Z. Budimac, D. Mitrović, V. Kurbalija, W. Dai, and W. Zhao, "Emotional intelligence and agents: Survey and possible applications," in *International Conference on Web Intelligence*, 2014.

[11] Y. Yang, X. Ma, and P. Fung, "Perceived emotional intelligence in virtual agents," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2017, pp. 2255–2262.

[12] A. P. Chaves and M. A. Gerosa, "How should my chatbot interact? a survey on human-chatbot interaction design," *arXiv preprint arXiv:1904.02743*, 2019.

[13] P. Salovey and J. D. Mayer, "Emotional intelligence," *Imagination, cognition and personality*, vol. 9, no. 3, pp. 185–211, 1990.

[14] X. Ma, E. Yang, and P. Fung, "Exploring perceived emotional intelligence of personality-driven virtual agents in handling user challenges," in *The World Wide Web Conference*. ACM, 2019, pp. 1222–1233.

[15] H. Chen, M. Sun, C. Tu, Y. Lin, and Z. Liu, "Neural sentiment classification with user and product attention," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1650–1659.

[16] Q. Qian, M. Huang, J. Lei, and X. Zhu, "Linguistically regularized LSTM for sentiment classification," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1679–1689. [Online]. Available: https://www.aclweb.org/anthology/P17-1154

[17] H.-Y. Shum, X.-d. He, and D. Li, "From eliza to xiaoice: challenges and opportunities with social chatbots," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 10–26, 2018.

[18] T. Hu, A. Xu, Z. Liu, Q. You, Y. Guo, V. Sinha, J. Luo, and R. Akkiraju, "Touch your heart: A tone-aware chatbot for customer care on social media," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, p. 415.

[19] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[20] Y.-C. Lee, N. Yamashita, Y. Huang, and W. Fu, ""i hear you, i feel you": Encouraging deep self-disclosure through a chatbot," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–12. [Online]. Available: https://doi.org/10.1145/3313831.3376175

[21] C. E. Williams and K. N. Stevens, "Emotions and speech: Some acoustical correlates," *The Journal of the Acoustical Society of America*, vol. 52, no. 4B, pp. 1238–1250, 1972.

[22] T. Johnstone and K. R. Scherer, "The effects of emotions on voice quality," in *Proceedings of the XIVth international congress of phonetic sciences*. Citeseer, 1999, pp. 2029–2032.

[23] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," p. 292–301, 2018. [Online]. Available: https://doi.org/10.1145/3240508.3240578

[24] S. Zhou, J. Jia, Q. Wang, Y. Dong, Y. Yin, and K. Lei, "Inferring emotion from conversational voice data: A semi-supervised multi-path generative neural network approach," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[25] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2227–2231.

[26] S. Chen, Q. Jin, J. Zhao, and S. Wang, "Multimodal multi-task learning for dimensional and continuous emotion recognition," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 19–26.

[27] C. N. Moridis and A. A. Economides, "Affective learning: Empathetic agents with emotional facial and tone of voice expressions," *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 260–272, 2012.

[28] Y. Shi, X. Yan, X. Ma, Y. Lou, and N. Cao, "Designing emotional expressions of conversational states for voice assistants: Modality and engagement," in *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–6.

[29] A. Wierzbicka, *Emotions across Languages and Cultures: Diversity and Universals*, ser. Studies in Emotion and Social Interaction. Cambridge University Press, 1999.

[30] M. Drescher, "French interjections and their use in discourse," *The Language of Emotions*, pp. 233–246, 1997.

[31] M. Cohn, C.-Y. Chen, and Z. Yu, "A large-scale user study of an alexa prize chatbot: Effect of tts dynamism on perceived quality of social dialog," in *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, 2019, pp. 293–306.

[32] J. Hu, Y. Huang, X. Hu, and Y. Xu, "Enhancing the perceived emotional intelligence of conversational agents through acoustic cues," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '21. New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: https://doi.org/10.1145/3411763.3451660

[33] R. W. Picard, *Affective computing*. MIT press, 2000.

[34] B. Reeves and C. I. Nass, *The media equation: How people treat computers, television, and new media like real people and places*. Center for the Study of Language and Information Cambridge University Press, Chicago New York, IL NY, 1996.

[35] C. Nass, J. Steuer, and E. R. Tauber, "Computers are social actors," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1994, pp. 72–78.

[36] X. Yang, M. Aurisicchio, and W. Baxter, "Understanding affective experiences with conversational agents," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–12. [Online]. Available: https://doi.org/10.1145/3290605.3300772

[37] P. E. Salovey and D. J. Sluyter, *Emotional development and emotional intelligence: Educational implications.* Basic Books, 1997.

[38] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: A review," *Int. J. Speech Technol.*, vol. 21, no. 1, p. 93–120, Mar. 2018. [Online]. Available: https://doi.org/10.1007/s10772-018-9491-z

[39] C. S. Ooi, K. P. Seng, L.-M. Ang, and L. W. Chew, "A new approach of audio emotion recognition," *Expert systems with applications*, vol. 41, no. 13, pp. 5858–5869, 2014.

[40] C. Brester, E. Semenkin, and M. Sidorov, "Multi-objective heuristic feature selection for speech-based multilingual emotion recognition," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 6, no. 4, pp. 243–253, 2016.

[41] T. Wu, Y. Yang, Z. Wu, and D. Li, "Masc: a speech corpus in mandarin for emotion analysis and affective speaker recognition," in *2006 IEEE Odyssey-the speaker and language recognition workshop*. IEEE, 2006, pp. 1–5.

[42] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[43] D.-n. Jiang, W. Zhang, L.-q. Shen, and L.-h. Cai, "Prosody analysis and modeling for emotional speech synthesis," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1. IEEE, 2005, pp. I–281.

[44] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-950

[45] E. Bozkurt, E. Erzin, Ç. E. Erdem, and A. T. Erdem, "Improving automatic emotion recognition from speech signals," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[46] J. H. Jeon, D. Le, R. Xia, and Y. Liu, "A preliminary study of cross-lingual emotion recognition from speech: automatic classification versus human perception." in *Interspeech*, 2013, pp. 2837–2840.

[47] L.-Y. Yeh and T.-S. Chi, "Spectro-temporal modulations for robust speech emotion recognition," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[48] Z. Song, X. Zheng, L. Liu, M. Xu, and X.-J. Huang, "Generating responses with a specific emotion in dialog," in *Proceedings of the 57th Conference of the Association for Computational Linguistics*, 2019, pp. 3685–3695.

[49] E. Andre, M. Rehm, W. Minker, and D. Bühler, "Endowing spoken language dialogue systems with emotional intelligence," in *Tutorial and Research Workshop on Affective Dialogue Systems*. Springer, 2004, pp. 178–187.

[50] K. Dohsaka, R. Asai, R. Higashinaka, Y. Minami, and E. Maeda, "Effects of conversational agents on human communication in thought-evoking multi-party dialogues," in *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, ser. SIGDIAL '09. USA: Association for Computational Linguistics, 2009, p. 217–224.

[51] S. W. McQuiggan, J. P. Rowe, and J. C. Lester, "The effects of empathetic virtual characters on presence in narrative-centered learning environments," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 1511–1520. [Online]. Available: https://doi.org/10.1145/1357054.1357291

[52] R. Morris, K. Kouddous, R. Kshirsagar, and S. Schueller, "Towards an artificially empathic conversational agent for mental health applications: System design and user perceptions," *Journal of Medical Internet Research*, vol. 20, 02 2018.

[53] M. Gennaro, E. Krumhuber, and G. Lucas, "Effectiveness of an empathic chatbot in combating adverse effects of social exclusion on mood," *Frontiers in Psychology*, vol. 10, p. 3061, 01 2020.

[54] J. Mumm and B. Mutlu, "Designing motivational agents: The role of praise, social comparison, and embodiment in computer feedback," *Computers in Human Behavior*, vol. 27, no. 5, pp. 1643–1650, 2011.

[55] J.-Y. Tzeng and C.-T. Chen, "Computer praise, attributional orientations, and games: A reexamination of the casa theory relative to children," *Computers in Human Behavior*, vol. 28, no. 6, pp. 2420–2430, 2012.

[56] M. G. Craske, L. Street, and D. H. Barlow, "Instructions to focus upon or distract from internal cues during exposure treatment of agoraphobic avoidance," *Behaviour research and therapy*, vol. 27, no. 6, pp. 663–672, 1989.

[57] K. N. Ochsner and J. J. Gross, "The cognitive control of emotion," *Trends in cognitive sciences*, vol. 9, no. 5, pp. 242–249, 2005.

[58] ——, "Cognitive emotion regulation: Insights from social cognitive and affective neuroscience," *Current directions in psychological science*, vol. 17, no. 2, pp. 153–158, 2008.

[59] P. Kanske, J. Heissler, S. Schönfelder, A. Bongers, and M. Wessa, "How to regulate emotion? neural networks for reappraisal and distraction," *Cerebral Cortex*, vol. 21, no. 6, pp. 1379–1388, 2011.

[60] A. Niculescu, S. Ge, E. van Dijk, A. Nijholt, H. Li, and S. See, "Making social robots more attractive: the effects of voice pitch, humor and empathy," *International journal of social robotics*, vol. 5, no. 2, pp. 171–191, 4 2013, eemcs-eprint-22397.

[61] Özge Nilay Yalçın, "Empathy framework for embodied conversational agents," *Cognitive Systems Research*, vol. 59, pp. 123 – 132, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389041719304826

[62] L. Chepenik, L. Cornew, and M. Farah, "The influence of sad mood on cognition," *Emotion*, vol. 7, no. 4, pp. 802–811, 2007, cited By 104.

[63] J. B. Bavelas, A. Black, C. R. Lemery, and J. Mullett, "Motor mimicry as primitive empathy." 1987.

[64] B. S. Hasler, G. Hirschberger, T. Shani-Sherman, and D. A. Friedman, "Virtual peacemakers: Mimicry increases empathy in simulated contact with virtual outgroup members," *Cyberpsychology, Behavior, and Social Networking*, vol. 17, no. 12, pp. 766–771, 2014.

[65] J. Scheirer, R. Fernandez, and J. Klein, "Frustrating the user on purpose: a step toward building an affective computer," *Interacting with Computers*, vol. 14, no. 2, pp. p.93–118, 2002.

[66] J. Klein, Y. Moon, and P. R.W., "This computer responds to user frustration: Theory, design, and results," *Interacting with Computers*, no. 2, p. 2, 2002.

[67] B. Reuderink, A. Nijholt, and M. Poel, "Affective pacman: A frustrating game for brain-computer interface experiments," in *Intelligent Technologies for Interactive Entertainment*, A. Nijholt, D. Reidsma, and H. Hondorp, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 221–227.

[68] W. S. Ark, D. C. Dryer, and D. J. Lu, "The emotion mouse," in *Hci International*, 1999.

[69] M. Macaulay, "The speed of mouse-click as a measure of anxiety during human-computer interaction," *Behaviour &amp; Information Technology*, vol. 23, no. 6, pp. 427–433, 2004.

[70] D. Kirsch, "The sentic mouse: Developing a tool for measuring emotional valence," *Mit Media Laboratory Perceptual Computing*, 1997.

[71] P. Neculoiu, M. Versteegh, and M. Rotaru, "Learning text similarity with Siamese recurrent networks," in *Proceedings of the 1st Workshop on Representation Learning for NLP*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 148–157. [Online]. Available: https://www.aclweb.org/anthology/W16-1617

[72] J. Du, Y.-H. Tu, L. Sun, F. Ma, H.-K. Wang, J. Pan, C. Liu, J.-D. Chen, and C.-H. Lee, "The ustc-iflytek system for chime-4 challenge," *Proc. CHiME*, pp. 36–38, 2016.

[73] K. Kretzschmar, H. Tyroll, G. Pavarini, A. Manzini, I. Singh, and N. Y. P. A. Group, "Can your phone be your therapist? young people's ethical perspectives on the use of fully automated conversational agents (chatbots) in mental health support," *Biomedical Informatics Insights*, vol. 11, p. 1178222619829083, 2019, pMID: 30858710. [Online]. Available: https://doi.org/10.1177/1178222619829083

[74] G. Guest, K. M. MacQueen, and E. E. Namey, *Applied thematic analysis*. Sage Publications, 2011.

[75] Wierzbicka and Anna, "The semantics of interjection," *Journal of Pragmatics*, vol. 18, no. 2-3, pp. 159–192, 1992.

[76] B. L. Fredrickson, "Positive emotions broaden and build," ser. Advances in Experimental Social Psychology, P. Devine and A. Plant, Eds. Academic Press, 2013, vol. 47, pp. 1 – 53. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780124072367000012

[77] M. Luria, R. Zheng, B. Huffman, S. Huang, J. Zimmerman, and J. Forlizzi, "Social boundaries for personal agents in the interpersonal space of the home," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI '20. New York, NY, USA: Association

for Computing Machinery, 2020, p. 1–12. [Online]. Available: https://doi.org/10.1145/3313831.3376311

[78] R. Van Buskirk and M. LaLomia, "The just noticeable difference of speech recognition accuracy," in *Conference Companion on Human Factors in Computing Systems*, ser. CHI '95. New York, NY, USA: Association for Computing Machinery, 1995, p. 95. [Online]. Available: https://doi.org/10.1145/223355.223446

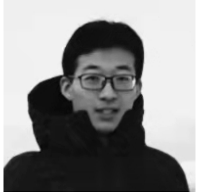
[79] M. Karam and m. c. schraefel, "Investigating user tolerance for errors in vision-enabled gesture-based interactions," in *Proceedings of the Working Conference on Advanced Visual Interfaces*, ser. AVI '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 225–232. [Online]. Available: https://doi.org/10.1145/1133265.1133309

[80] J. Crumpton and C. L. Bethel, "A survey of using vocal prosody to convey emotion in robot speech," *International Journal of Social Robotics*, vol. 8, no. 2, pp. 271–285, Apr 2016. [Online]. Available: https://doi.org/10.1007/s12369-015-0329-4

[81] T. W. Bickmore, R. Fernando, L. Ring, and D. Schulman, "Empathic touch by relational agents," *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 60–71, 2010.

Jiaxiong Hu received the BS degree in computer science from Beijing University of Post and Telecommunication, Beijing, China, in 2015, the MDes degree in information art & design from Tsinghua University, Beijing, China, in 2019. He is working towards a Ph.D. degree in Design at Tsinghua University, Beijing, China. His research interests include human-computer interaction, conversational user interface, and affective computing.

Yun Huang is an assistant professor in the School of Information Sciences at the University of Illinois at Urbana-Champaign. She co-directs the Social Computing Systems (SALT) Lab. Before joining Illinois, she was a faculty member in the School of Information Studies at Syracuse University and a postdoc fellow at Carnegie Mellon University. Her work focuses on social computing systems research, in which she examines context-driven approaches of designing crowdsourcing systems. For example, her research examines how different contexts impact user contributions to crowdsourcing systems; how to leverage these contextual effects to design innovative social computing systems that can better engage users; and how crowdsourced user behavioral data may help better understand users and the community. She received her PhD from the Donald Bren School of Information and Computer Sciences at the University of California, Irvine. She earned her bachelor's degree from the Department of Computer Science and Technology at Tsinghua University, in Beijing, China.



Xiaozhu Hu is a master student in Pervasive HCI Lab, Department of Computer Science and Technology, Tsinghua University (THU), supervised by Prof. Chun Yu. He received the bachelor degree in Industrial Design from Beijing University of Posts and Telecommunications (BUPT). He is going to start the Ph.D. program in the Computational Media and Arts Thrust, Hong Kong University of Science and Technology (HKUST) under the supervision of Prof. Mingming Fan. His research interests are in Human-Computer Interaction (HCI), currently about 1) inclusive interaction technique in the field of Accessibility and Aging, 2) automatic and semi-automatic media content generation and 3) AR/VR interaction techniques.

Yingqing Xu a professor of Academy of Arts & Design, Tsinghua University, Beijing China. At Tsinghua University, he serves as Director of the Future Lab, Director of the Lab for Lifelong Learning (TULLL), Director of Tsinghua University-Alibaba Joint Research Laboratory for Natural Human Computer Interaction, and Director of the Center for Cultural Creative Design Research of The Future Laboratory, Tsinghua University. His teaching and researching are user experience design, natural user interface, immersive perception & interaction, tangible perception & interaction, and e-heritage. Before joined Tsinghua University, he was a Lead Researcher of Microsoft Research Asia (MSRA) where he had worked for 12 years since January 1999. Dr. Xu has published over 100 research papers and granted patents, and served as the conference chair, co-chair or program committee members for the academic conferences. He is a fellow of CCF (China Computer Federation), a member of CAA (China Artists Association), CDIA (China Industrial Design Association), ACM (Association for Computing Machinery); and a senior member of IEEE (Institute of Electrical and Electronics Engineers). He holds the B.Sc degree from Department of Mathematics of Jilin University, and Ph.D degree from Institute of Computing Technology, Chinese Academy of Sciences (CAS).