

Article

The Development of an Emotional Embodied Conversational Agent and the Evaluation of the Effect of Response Delay on User Impression

Simon Christophe Jolibois ^{1,2}, Akinori Ito ^{1,*}  and Takashi Nose ¹ ¹ Graduate School of Engineering, Tohoku University, Sendai 980-8579, Japan;

jolibois.simon.christophe.t1@alumni.tohoku.ac.jp (S.C.J.); nose@tohoku.ac.jp (T.N.)

² École Centrale de Lyon, Av. Guy de Collongue, 69130 Écully, France

* Correspondence: aito.spc@tohoku.ac.jp

Abstract: Embodied conversational agents (ECAs) are autonomous interaction interfaces designed to communicate with humans. This study investigates the impact of response delays and emotional facial expressions of ECAs on user perception and engagement. The motivation for this study stems from the growing integration of ECAs in various sectors, where their ability to mimic human-like interactions significantly enhances user experience. To this end, we developed an ECA with multimodal emotion recognition, both with voice and facial feature recognition and emotional facial expressions of the agent avatar. The system generates answers in real time based on media content. The development was supported by a case study of artwork images with the agent playing the role of a museum curator, where the user asks the agent for information on the artwork. We evaluated the developed system in two aspects. First, we investigated how the delay in an agent's responses influences user satisfaction and perception. Secondly, we explored the role of emotion in an ECA's face in shaping the user's perception of responsiveness. The results showed that the longer response delay negatively impacted the user's perception of responsiveness when the ECA did not express emotion, while the emotional expression improved the responsiveness perception.



Academic Editor: João M. F. Rodrigues

Received: 10 February 2025

Revised: 30 March 2025

Accepted: 3 April 2025

Published: 11 April 2025

Citation: Jolibois, S.C.; Ito, A.; Nose, T. The Development of an Emotional Embodied Conversational Agent and the Evaluation of the Effect of Response Delay on User Impression. *Appl. Sci.* **2025**, *15*, 4256. <https://doi.org/10.3390/app15084256>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: embodied conversational agent; facial expression; emotion recognition; emotional expression; response delay

1. Introduction

An embodied conversational agent (ECA) [1,2] is a kind of user interface that simulates human face-to-face conversation. While conversational agents, such as smart speakers, do not necessarily have human-like bodies [3], ECAs have human-like figures, such as computer-generated figures or humanoid robots. ECAs are used as an intelligent interface for various applications, such as home automation [4–6], education [7–9], healthcare [9–11], and tourism [12].

Since ECAs have human-like figures, users expect these ECAs to behave like humans, including nonverbal behaviors accompanying speech, such as facial expressions, gestures, and gaze [13]. However, because of the limitation of computational models, such as speech recognizers, the response delay of ECAs inevitably becomes longer than that of humans. In human–machine conversation, a larger response delay affects the cognitive performance of humans [14]. Therefore, a spoken dialogue system should respond to a user as quickly as a human would.

In human–human dialogue, an utterance often starts before the interlocutor’s utterance ends [15]. Humans can do this because of the ability to predict the transition relevance place (TRP) [16]. Several systems have tried to predict the timing of turn-taking before the previous utterance ends [17–19]. However, it is still difficult for spoken dialogue systems to predict the TRP and take turns properly in real time.

In human–human dialogue, it is known that pauses tend to be longer in cooperative (friendly) conversations than in arguments [20]. Devillars et al. found that in-utterance pauses in emotional utterances are longer than those in neutral utterances [21]. From these facts, we devised an idea to control the user’s impression induced by response delay by introducing emotional behavior into the ECA.

This paper is organized as follows: In Section 2, we describe the related work of ECAs. In Section 3, we describe the details of the development of the ECA with emotion recognition and facial expressions. Section 4 describes the experiment, including the experimental conditions, procedures, and results. We conclude the paper in Section 5.

2. Related Work

The term “embodied conversational agent” was coined by J. Cassell [1,2]. In [2], he listed the following properties an ECA should have:

- Recognizing and responding to verbal and nonverbal input;
- Generating verbal and nonverbal output;
- Dealing with conversational functions, such as turn-taking, feedback, and repair mechanisms;
- Giving signals that indicate the state of the conversation, as well as contributing new propositions to the discourse.

Nonverbal behaviors such as nodding or changing facial expressions are the special features of an ECA compared with conversational agents without bodies. Cassell et al. developed an ECA named Rea [22], which behaved as a real estate agent. Rea had most of the basic functionalities an ECA should have; however, most functions are realized using handwritten rules. Since then, many ECAs have been developed [23,24].

The emotional behavior of an ECA, including perception of the user’s emotion and expressing emotion using facial expression, is an important aspect when designing an ECA. The effect of introducing emotion in an ECA was first discussed in the late 1990s [25].

The main objective of introducing emotional behavior into ECAs is to enhance the user’s impression of the ECAs. Expressing positive emotion is especially expected to improve this impression [26]. Since then, many studies have been conducted to develop general models of emotional dynamics [27–29]. Not only are facial expressions and gestures effective but so is the paralinguistic property of speech in presenting specific emotions of ECAs [30,31]. Conventional models use rules to determine the ECAs’ emotional behavior, while recent studies use more sophisticated models based on deep learning. For example, Firdaus et al. proposed a model based on the multimodal-attention-based conditional variational autoencoder (M-CVAE) [32].

Since the aim of introducing emotional behavior into ECAs is to improve the user’s impression of the system, such behaviors are often used in ECAs for clinical or healthcare purposes [33–36].

As listed above, most studies have focused on how emotional behavior makes ECAs more human-like and improves the user’s impression. On the other hand, emotional behavior could interact with other factors of dialogue, such as response delay. This is the duration from the end of the user’s utterance to the beginning of the ECA’s utterance. As mentioned before, the response delay affects the user’s cognition performance [14]. Funk et al. reported that 0 to 250 ms of delay gave the best satisfaction and usability for

in-vehicle conversational agents [37]. Gnewuch et al. reported an interesting result on the user's preference for the response delay [38]; they proved that taking a long time to answer complicated questions improved the user's impression of the system and perception of humanness. This result is considered a realization of response delay distribution according to different question types [15]. Heeman et al. showed that the response delay varies according to the dialogue act of the previous utterance [39].

As described, there are studies on how linguistic content affects the response delay; however, no research has been conducted on how emotional behavior and the response delay jointly affect the ECA's impression. If we find any relationship between emotional behavior and response delay, it can be used to make the dialogue behavior of ECAs more natural.

3. Development of an Embodied Emotional Conversational Agent

To conduct the experiment, we first developed an ECA that converses with a user in real time (The system is available at https://github.com/SimonJolibois/Multimodal_Embodied_Conversational_Agent, accessed on 19 June 2024). Note that the technology used in the ECA is not novel; however, we tuned the system to work as quickly as possible so that we can control the response delay. The details of the system are described in our previous paper [40].

3.1. Overview

Figure 1 shows an overview of the system. The system observes the user's face with a camera and captures the user's voice with a microphone. The observed face is used to estimate the user's emotion. The recorded voice is transcribed using the speech recognizer, and the transcription is used to select the system's response and estimate the user's emotion. The agent's verbal response is selected from a predefined query–response database. The facial expression of the agent is controlled using both the emotion of the response sentence and the user's emotion. When playing the response voice, lip movement is generated from the response voice.

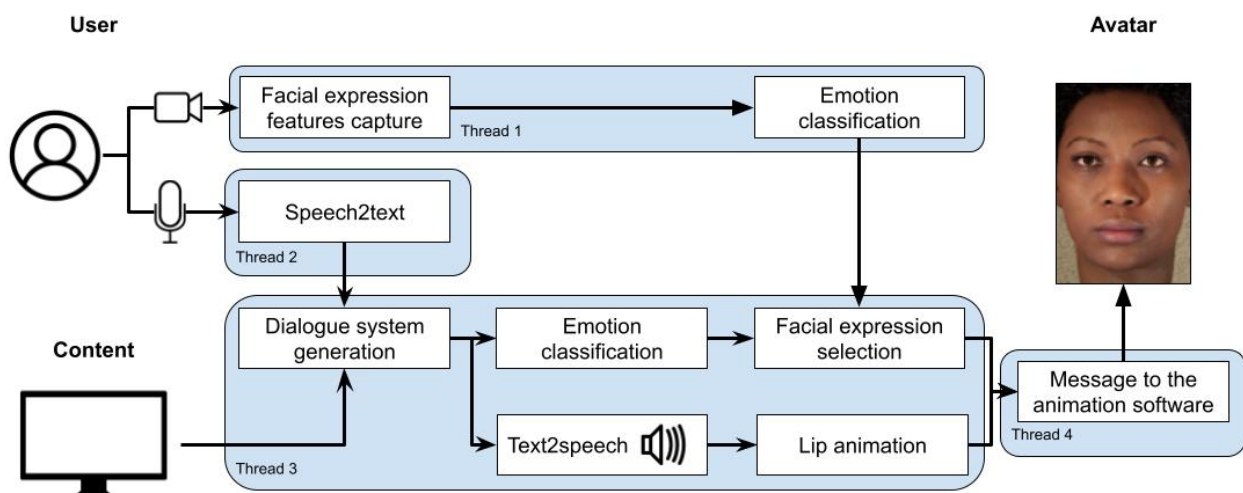


Figure 1. The system architecture. The system captures the user's face using a camera to recognize emotion. At the same time, it records the user's voice using a microphone and transcribes it. The transcribed utterance is used to choose the system's response and estimate emotion from the text. The facial expression of the agent is generated according to the user's estimated emotion. The lip movement of the agent is controlled according to the voice of the system. The process of the system has four threads to process the data in parallel, as explained in Section 3.8.

3.2. Emotion Recognition from Facial Image

Using the OpenFace software toolkit (version 2.2.0) [41], the facial landmarks are collected as a CSV file. The facial expression is expressed as the Facial Action Coding System (FACS) [42]. It breaks down facial poses into distinct muscle actions, known as action units (AUs), allowing researchers to decode and classify emotions and expressions objectively.

We used the 17 AUs shown in Table 1 for emotion classification. Then, we used a four-layer perceptron for the classification; the input dimension was 17; there were 15, 12, and 10 hidden units; and the output dimension was 8 for classifying eight emotions (neutral, anger, contempt, disgust, fear, happiness, sadness, surprise).

We utilized the CK+ dataset [43] for model training, which includes 593 images from 123 subjects, with 327 sequences explicitly labeled for emotions fitting stereotypical definitions. These sequences range from neutral to high-intensity expressions, classified into the same eight emotional categories. The model achieved an average accuracy of 85%.

Table 1. The 17 AUs used for emotion recognition from a facial image.

No.	Description	No.	Description
01	Inner Brow Raiser	14	Dimpler
02	Outer Brow Raiser	15	Lip Corner Depressor
04	Brow Lowerer	17	Chin Raiser
05	Upper Lid Raiser	20	Lip Stretcher
06	Cheek Raiser	23	Lip Tightener
07	Lid Tightener	25	Lips Part
09	Nose Wrinkler	26	Jaw Drop
10	Upper Lip Raiser	45	Blink
12	Lip Corner Puller		

3.3. Speech Recognition and Emotion Estimation

We transcribed the user's speech using Microsoft Azure CognitiveService Speech-to-Text. Then, the user's emotion was estimated from the transcription. We used the GoEmotions dataset [44], which comprises approximately 58,000 Reddit comments categorized among 27 emotions as neutral.

Because some of the emotions in the database are not useful when animating facial expressions (for example, *love* or *remorse*), a subset of the GoEmotions dataset containing 14 emotions was processed in the selection phase: *anger, annoyance, confusion, curiosity, disgust, embarrassment, excitement, fear, grief, joy, nervousness, pride, sadness, surprise*.

We developed the text emotion classification model based on the 'base-uncased' BERT model. Table 2 shows the specification of the model. After fine-tuning the model, we obtained an accuracy of 45% in emotion classification. Detailed results are shown in Table 3. The model attained a weighted macro-average precision of 45%.

Table 2. Specifications of the BERT-based emotion classification model.

Number of layers/transformer blocks	12
Number of units in the hidden layers	768
Total number of parameters	110 M
Number of self-attention heads	12
Number of epochs	4
Batch size	16
Learning rate	5×10^{-5}
Validation split	85/15

Table 3. Text emotion classification results.

Emotion	Accuracy (%)
Anger	46
Annoyance	41
Confusion	35
Curiosity	45
Disgust	37
Embarrassment	37
Excitement	21
Fear	65
Grief	6
Joy	71
Nervousness	25
Pride	0
Sadness	59
Surprise	61
Weighted macro-average	45

3.4. Dialogue Management

Dialogue management is based on a simple Q&A database [45]. In this dialogue management architecture, we prepare a database containing examples of users' utterances and the system's corresponding answers. When a user inputs an utterance, the dialogue system transcribes it using a speech recognizer. Then, the system chooses the example most similar to the input sentence from the database. Finally, the system synthesizes the speech using the corresponding answer text. Because of its simplicity and robustness, many spoken dialogue systems employ Q&A database architecture [5,45,46].

The domain of the dialogue was the question and answer of four artworks, shown in Table 4. The system shows the image of the painting on the system, and then the user starts a conversation with the system by asking questions about the artwork.

To prepare the Q&A database, we first developed one that included information on the artworks and painters, as shown in Table 5. All the information was gathered from Wikipedia and WikiArt. Next, we developed a set of questions and answer templates. For example, we prepare a question "When was the artwork painted?" and an answer template "It was painted on <DATE>". Then, the Q&A database was generated from the template and items in the artwork database by substituting the tag such as <DATE> with the information in the artwork database.

When the transcription of the user's utterance is given, we choose the nearest example sentence in the Q&A database by choosing the sentence with the largest cosine similarity.

Table 4. Artworks used in the dialogue system.

No.	Name of the Artwork	Painter
1	A Huguenot on St. Bartholomew's Day	John Everett Millais
2	Liberty Leading the People	Eugène Delacroix
3	The Roses of Heliogabalus	Lawrence Alma-Tadema
4	The Starry Night	Vincent van Gogh

Table 5. Items in the artwork database.

No.	Item
1	Title in English and original language
2	Place and date of creation
3	Dimensions
4	Genre
5	Format
6	Subject matter
7	Cultural and historical context in which the artwork was born
8	The intent of the author for creating the piece
9	The current conservation place and methods
10	Previous exhibitions the artwork was part of
11	The ownership and sale history
12	Technical analysis done on the artwork
13	The painter's name
14	The painter's biography
15	The painter's artistic movement(s)
16	The painter's famous artworks
17	Inspirations/references used for the artwork

3.5. Generation of Face Image and Facial Expressions

To select an appropriate facial expression, we consider both the emotion predicted from the user's facial expression and the emotion prediction of the agent's generated answer. The procedure is as follows:

- **Prioritizing the answer's emotion:** The primary factor is the emotion predicted from the generated answer. If the prediction emotion is not neutral, the ECA's face is rendered using that emotion.
- **Maintaining Current Expression:** In instances where the predicted emotion of the answer is neutral, the agent keeps the same facial expression.
- **Reflecting the user's emotion:** If the agent's facial expression is neutral twice, the agent mimics the user's facial expression. It mimics the behavior of entrainment, where individuals can adjust their speaking rate, pitch, and gesture to match those of their conversation partner [47]. In the context of an ECA, entrainment can foster more engagement for the human user [48].

3.6. Text-to-Speech and Lip Syncing

Microsoft Azure's Text-to-Speech service generates a response to the user's speech, utilizing Speech Synthesis Markup Language (SSML) to define the language, voice, and pronunciation features, including emotional tones. The service outputs both the audio file and associated viseme data. While phonemes are crucial for audio speech synthesis, visemes are essential for creating accurate lip-syncing in visual representations, such as in animations or virtual avatars. In the context of an ECA, focusing on visemes is important because they directly impact the visual synchronization of the lips with the generated speech, enhancing the realism and coherence of the virtual character.

3.7. Facial Animation Rendering

The facial animation is rendered in the Unreal Engine 5 game engine. The animation is subdivided into four blueprints, each taking a specific role. One is the receiver/emitter, which communicates with the Python 3.10 script that controls the other parts of the system. Another takes the user's input on the screen through the UI. A third blueprint stores all variables of the agent, and the last one handles the animation depending on these variables. The animation is carried out in layers, one after another, as shown in Figure 2.

We first obtained the facial expression, the one we selected before. We added some looping animation for blinking, as it made the agent more aware-like and less robotic. We can also control the direction of the gaze and the head. Lastly, we added the lip-syncing animation.

The character is a Metahuman, a high-quality 3D character developed by Epic Games. It is available for free in Unreal Engine. Metahumans are notable for their high-fidelity textures and procedural skeleton system, which allows for the creation of a diverse range of realistic 3D characters suitable for various projects. These characters come equipped with pre-defined facial expressions and viseme poses, simplifying the process of animating them.

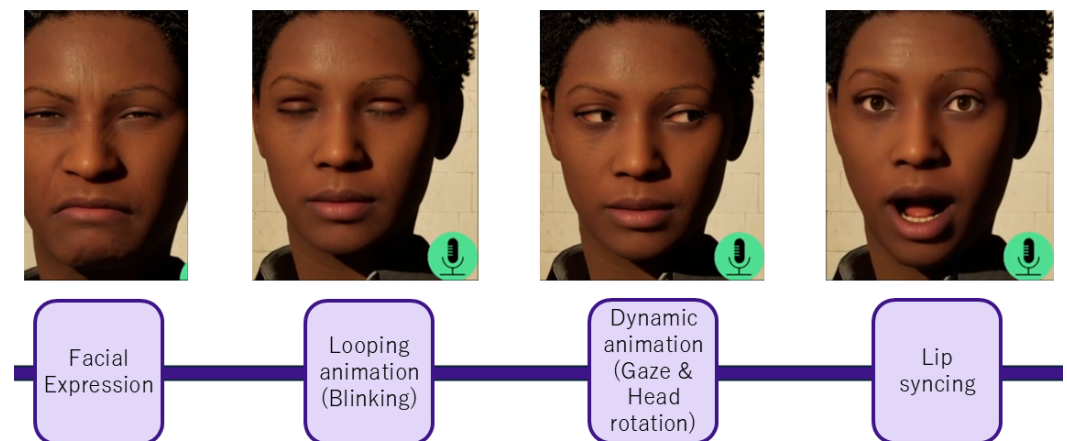


Figure 2. Rendering of the face animation.

3.8. Controlling the Entire System

We divided the entire system into four threads. The first one captures the user's face and extracts AUs. The second observes and recognizes the user's speech using Azure CognitiveService Speech-to-Text. The third performs almost all tasks except animation rendering, including estimating emotion from face and text, dialogue management, and choosing the agent's emotion. The final one renders the facial animation.

Figure 3 shows the user interface of the dialogue system. The ECA is displayed on the left side, and the artwork is on the right. The user talks with the ECA while looking at the artwork. An example conversation video is provided as Supplementary Materials (Video S1).

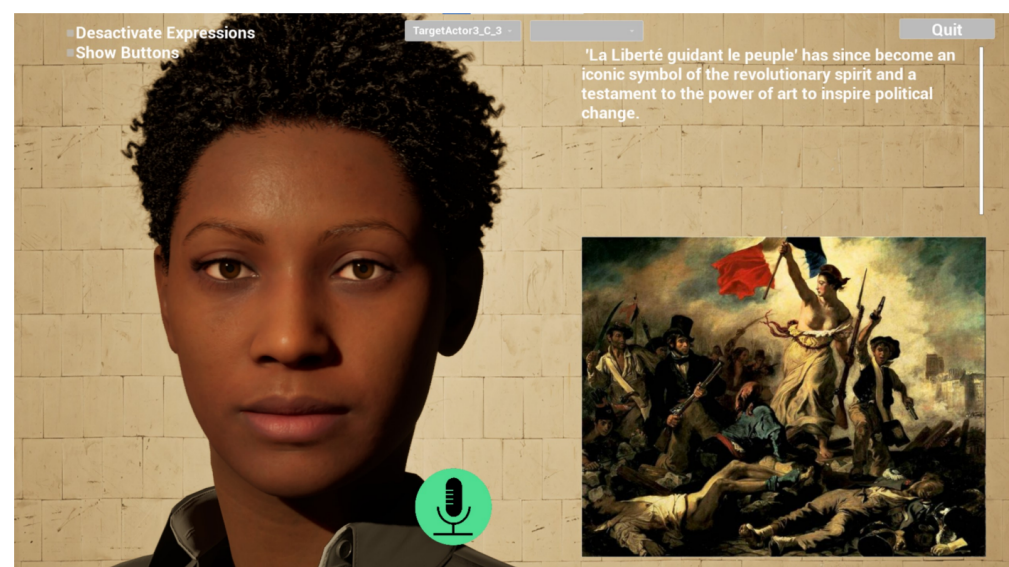


Figure 3. The user interface of the dialogue system.

4. Experiment

4.1. Overview

As an evaluation of the developed system, we evaluated the users' impressions of it related to emotional expression and response delay. The purpose of this experiment is to clarify how the presence or absence of emotional expression and the length of response delay in the ECAs developed here affect users' impressions of the ECAs (e.g., satisfaction, responsiveness, interactivity, and information accuracy). As noted in Section 2, it is not clear how the response delay, which inevitably occurs in ECAs, interacts with the emotional expression of ECAs to alter various user impressions of ECAs. Therefore, in this experiment, we modified the emotional expression and response delay of ECAs to determine how they affect user impressions.

Based on existing research that highlights the response timing in human–human interaction, this research seeks to extend the research to ECAs [20,49,50]. In the experiment proposed by Funk et al. for automotive user interfaces [37], Satisfaction, Usability, and Context Awareness were evaluated depending on the response time of the user interface. This research aims to determine whether the agent response delay affects the user's evaluation of the agent. Their work showed that the user's evaluation declined with response delays exceeding 2 s or preemptive responses under 0 ms. Additionally, we aim to answer the research question of whether displaying emotions on the ECA's face changes the user's perception of the response delay.

4.2. Experimental Conditions

Participants in the experiment were international students enrolled at Tohoku University and were recruited through a bulletin board on campus. There were no restrictions on age range or English ability, but we explained that the dialogue would be conducted in English and asked only those participants who self-evaluated themselves as having sufficient English conversation skills to participate. For the definition of experimental parameters, we needed 24 participants to obtain a statistical power of 0.8 and an effect size of 0.8. Thus, we invited 25 participants (18 males and 7 females) to the experiment. Of these, 13 participants had experience using chatbots. Not all of the participants were native speakers of English; we asked the participants to choose their self-evaluation of English fluency among "beginner", "intermediate", "fluent", and "advanced". As a result, no participants were beginners, five were intermediate, five were fluent, and fifteen were advanced. The procedure of the experiment was as follows:

- We explained the purpose and procedure of the experiment to the participant.
- We asked the participant for a pre-dialogue questionnaire (Table 6).
- The participant held four rounds of dialogue with the agent. In each round, a participant conversed with the ECA about one particular artwork.
- After each conversation, the participant answered the questionnaire (Table 7) evaluating the current dialogue.
- After four rounds of dialogue, the participant answered the post-dialogue questionnaire (Table 8).

The order of the response delay and combination with the artwork are shown in Table 9. The order was determined so that the different participants experienced different orders and combinations of artworks, response delays, and the presence of emotion. Overall, each parameter was balanced.

During the conversation, they asked the agent questions about the displayed artwork for as long as they wanted. The agent's response delay was controlled with a minimum threshold so that the agent waited until the timer had finished. The response delay was set to 500, 1000, 1500, and 2000 ms because most of the response delay in a conversation is

between 500 and 2000 ms [50]. The specific response delay was achieved by inserting an artificial delay after preparing the ECA's response speech.

Table 6. Questionnaire before the conversation.

Q1	Gender
Q2	Are you familiar with the concept of a chatbot? (No, I don't know what a chatbot is./I don't have much knowledge about it./Yes, I am familiar with the concept.)
Q3	Have you already used a chatbot? (Never used before, Seldom use, Frequently use)
Q4	How would you rate your English level? (Beginner, Intermediate, Advanced, Fluent)

Table 7. Questionnaire after each round of conversation (answered on 5-point Likert scale).

Q1	Animation
Q2	Conversation skill of the chatbot (Flow of the conversation, phrasing)
Q3	Information accuracy (Does the agent's answer correspond to your question?)
Q4	Interactivity (Is the agent responsive to your inputs?)
Q5	Interface (Layout, Visibility)
Q6	Naturalness of the conversation (Does the agent look human-like when conversing?)
Q7	Responsiveness (How fast does the agent answer back?)
Q8	Overall satisfaction
Q9	Usability (Ease of use)
Q10	Visual aspect (Graphics)

Table 8. Questionnaire after four rounds of conversation.

Q1	Among the following parameters, what do you think hindered your conversation with the chatbot the most? (Visual aspect, Animation, Conversation skill of the agent, Conversation topic, Interactivity, Responsiveness, Interface)
Q2	Please elaborate on your choice
Q3	Which emotions/expressions were overrepresented? Underrepresented?
Q4	Were you knowledgeable on the topic of artworks? Did the conversation topic hinder your ability to interact with the agent?
Q5	Do you think one of the aspects of the system is not exploited enough?
Q6	If you have any comments, please write them here

When the ECA was in the “with-emotion” condition, the ECA changed its facial expression according to the method explained in the previous section. In the “without-emotion” condition, the ECA's face expressed a neutral emotion.

Table 9. The experimental design. In this table, “No.” is the participant number, and “Art”, “RD”, and “Emo” denote the artwork number, response delay, and existence of emotional expression, respectively.

No.	Round 1			Round 2			Round 3			Round 4		
	Art	RD	Emo	Art	RD	Emo	Art	RD	Emo	Art	RD	Emo
1	1	500	Yes	2	1000	No	3	1500	Yes	4	2000	No
2	2	500	No	3	1000	Yes	4	2000	No	1	1500	Yes
3	3	500	Yes	4	1500	No	1	1000	Yes	2	2000	No
4	2	500	No	1	1500	Yes	4	2000	No	3	1000	Yes

Table 9. Cont.

No.	Round 1			Round 2			Round 3			Round 4		
	Art	RD	Emo	Art	RD	Emo	Art	RD	Emo	Art	RD	Emo
5	1	500	Yes	2	2000	No	3	1000	Yes	4	1500	No
6	2	500	No	3	2000	Yes	4	1500	No	1	1000	Yes
7	3	1000	Yes	4	500	No	1	1500	Yes	2	2000	No
8	4	1000	No	1	500	Yes	2	2000	No	3	1500	Yes
9	1	1000	Yes	2	1500	No	3	2000	Yes	4	500	No
10	2	1000	No	3	1500	Yes	4	500	No	1	2000	Yes
11	3	1000	Yes	4	2000	No	1	500	Yes	2	1500	No
12	4	1000	No	1	2000	Yes	2	1500	No	3	500	Yes
13	1	1500	Yes	2	500	No	3	1000	Yes	4	2000	No
14	2	1500	No	3	500	Yes	4	2000	No	1	1000	Yes
15	3	1500	Yes	4	1000	No	1	500	Yes	2	2000	No
16	4	1500	No	1	1000	Yes	2	2000	No	3	500	Yes
17	1	1500	Yes	2	2000	No	3	1000	Yes	4	500	No
18	2	1500	No	3	2000	Yes	4	500	No	1	1000	Yes
19	3	2000	Yes	4	500	No	1	1000	Yes	2	1500	No
20	4	2000	No	1	500	Yes	2	1500	No	3	1000	Yes
21	1	2000	Yes	2	1000	No	3	1500	Yes	4	500	No
22	2	2000	No	3	1000	Yes	4	500	No	1	1500	Yes
23	3	1500	Yes	4	1000	No	1	500	Yes	2	2000	No
24	4	2000	No	1	1500	Yes	2	1000	No	3	500	Yes
25	2	2000	No	3	1000	Yes	4	500	No	1	1500	Yes

4.3. Questionnaire

We asked the participants three kinds of questionnaires: a demographic questionnaire before the conversation (Table 6), a questionnaire to evaluate each conversation (Table 7), and a questionnaire to evaluate the total impression of the ECA after all conversations (Table 8).

The questionnaire for evaluating the conversation was developed after the chatbot evaluation criteria by Peras [51], where the criteria for evaluating the ECA were separated into 14 categories: user experience [usability, performance, affect, satisfaction], information retrieval [accuracy, accessibility, efficiency], linguistic [quality, quantity, relation, manner, grammatical accuracy], technology [humanity], and business value. We removed the items related to the answer generation system (robustness of the agent, grammatical accuracy, accessibility, linguistic) because a set of answers was manually crafted. We also removed questions related to business value because they were irrelevant to this study. In addition, the naturalness aspect of the agent was divided into the visual and verbal aspects of the agent.

4.4. Number of Utterances in a Dialogue

As described above, we did not restrict the dialogue length; the participants spoke freely with the ECA. Therefore, the number of user utterances differs from dialogue to dialogue. Figure 4 shows the distributions of this number. Its average is 17.24, and its average duration is 287.66 s.

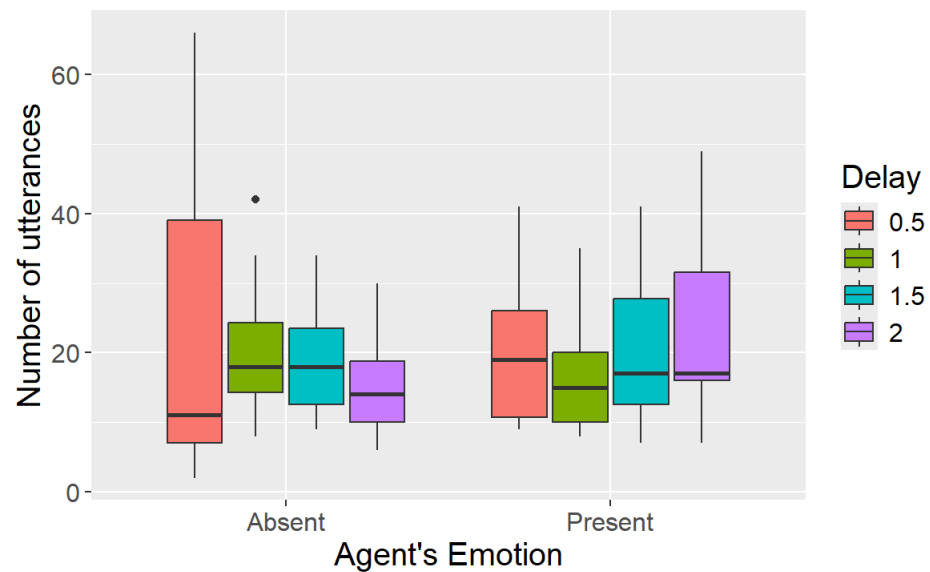


Figure 4. Distributions of the number of user utterances in a dialogue.

4.5. The Real Response Delay

As stated above, we controlled the response delay by inserting the artificial delay after preparing the response speech. However, if the preparation takes longer than the intended delay, the actual delay becomes longer. Figure 5 shows the jitter plots of the actual response delay. The red digits show the average. The figure shows that the average response delays were slightly longer for the 500 ms condition and almost as intended for the longer conditions.

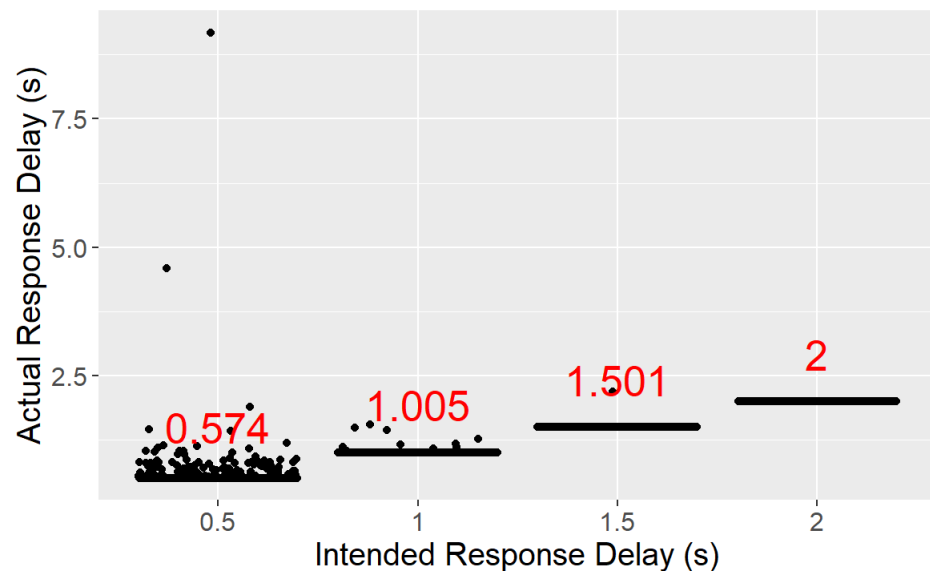


Figure 5. The jitter plot of the actual response delay. The red digits show the average delay.

4.6. Total Impression of the Conversation by the System

Figure 6 shows the mean values of the evaluation shown in Table 7, including all emotion and response delay conditions. The error bars show the standard errors. Values of the items “Interface” and “Visual aspect” were relatively high while “Naturalness” was relatively low. Most of the mean evaluation values were around 3, showing that the participants could use the system easily.

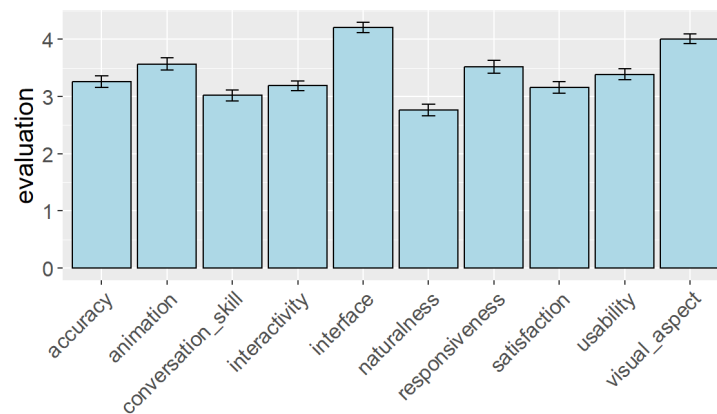


Figure 6. The questionnaire result after each conversation.

4.7. Effect of Response Delay and Emotional Expression

Next, we analyzed the evaluation of each conversation shown in Table 7 to determine whether their evaluation values were affected by the response delay and emotional expression.

Table 10 shows the result of the Multivariate Analysis of Variance (MANOVA) that investigates the effect of the emotional expression and the response delay upon all the evaluation values. As a result, we found a significant interaction between emotional expression and response delay ($p = 0.050$), but the independent factor did not show significance.

Table 11 shows the results of the Analysis of Variance (ANOVA) to evaluate differences among different response delay times and the presence of emotional facial expressions. In this analysis, the presence of emotional expression was treated as a categorical variable, and the response delay was regarded as a continuous variable. From these results, we found statistically significant differences in information accuracy (emotional expression), responsiveness (response delay), and satisfaction (emotional expression). We also found statistically significant differences in the interaction between the response delay and emotion for responsiveness.

Table 10. Results of MANOVA. The independent variables are emotions and different response delays; the response variables are the 10 items shown in Table 7.

Variable	Df	Pillai's Trace	Approx. F	Df1	Df2	<i>p</i>
Response delay	3	0.23772	0.7315	30	255	0.84712
Emotion	1	0.16986	1.6983	10	83	0.09470
Delay $\times$ Emotion	3	0.45136	1.5053	30	255	0.04964

Figures 7–9 show the evaluation results of information accuracy, responsiveness, and overall satisfaction, respectively. As shown in Figures 7 and 9, the participants' perception of information accuracy and overall satisfaction increased when the agent changed facial expressions according to the emotion estimation; however, the response delay did not affect these evaluation results. The result of information accuracy seems to be related to the relationship between the speaker's facial expressions and trustworthiness, where showing a happy face increases the speaker's trustworthiness [52]. The overall satisfaction was not affected by the response delay, either. Chen et al. reported that the satisfaction of dialogue systems monotonically decreased when the response delay increased [53]. Their dialogue system used only speech and did not have facial images; this difference could cause the different responses of the users to the response delay.

Table 11. Results (p -values) of the statistical test (ANOVA) for the evaluation of conversation. We used the response delay (500, 1000, 1500, and 2000 ms) and the emotional facial expression (yes or no) as factors. An asterisk (*) means 5% significance.

Item	Response Delay	Emotional Expression	Interaction
Animation	0.917	0.298	0.093
Conversation skill	0.794	0.192	0.270
Information accuracy	0.529	0.027 *	0.581
Interactivity	0.119	0.237	0.287
Interface	0.771	0.649	0.062
Naturalness	0.336	0.474	0.683
Responsiveness	0.026 *	0.069	0.038 *
Satisfaction	0.938	0.027 *	0.485
Usability	0.781	0.383	0.122
Visual aspect	0.834	0.273	0.258

On the other hand, the evaluation values of responsiveness decreased along with increasing response delay when the agent did not change facial expressions. This is a natural response of the participants because the responsiveness includes the speed of the agent's answering behavior. Interestingly, the responsiveness value did not change for different response delays when the agent changed facial expressions according to emotion.

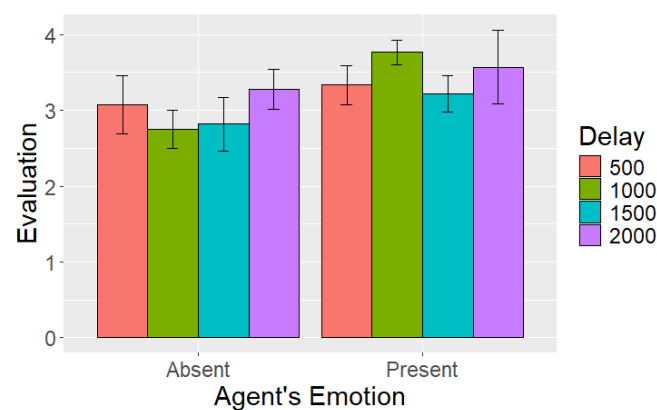


Figure 7. The evaluation result of “information accuracy” with and without the agent's emotional expression. The error bars indicate the standard error. The effect of emotional expression was statistically significant ($p = 0.027$) but that of response delay was not ($p = 0.581$).



Figure 8. Evaluation results of “responsiveness” with and without the agent's emotional expression. The error bars indicate the standard error. The effect of response delay was statistically significant ($p = 0.026$), but that of emotional expression was not ($p = 0.069$). The interaction between the two factors was significant ($p = 0.038$).



Figure 9. Evaluation results of “satisfaction” with and without the agent’s emotional expression. The error bars indicate the standard error. The effect of emotional expression was statistically significant ($p = 0.027$), but that of response delay was not ($p = 0.938$).

Asaka et al. reported that inserting a filler in parallel with the response generation process improved the users’ impression of the system’s responsiveness [54]. In their experiment, the physical response delay was actually shorter than the system without filler insertion; on the other hand, the response delay time of our system was constant regardless of the conditions. The reason the perception of the response delay changed is still unclear. One possible reason is that the users paid more attention to the agent’s facial expression, which made the perception of the response delay less sensitive.

4.8. Effect of Demographic Characteristics

We analyzed how the demographic characteristics of the participants affected the evaluation. Note that because the number of participants (twenty-five) is small and the distribution is not controlled, the inference from the result could be misleading. Therefore, we only show the results and discuss the possible reasons.

In this analysis, we excluded gender as a factor because the gender distribution was unbalanced (18 males and 7 females). We analyzed the effect of English proficiency (three levels: intermediate, fluent, and advanced) and the experience of having used chatbots (two levels: Yes or No; the answers ‘I don’t have much knowledge’ were merged to ‘No’). As a result of two-way ANOVA, we found statistically significant differences for Animation (interaction of English and chatbot use, $p = 0.008$), Naturalness (chatbot use, $p = 0.045$), and Visual aspect (English, $p = 0.00012$). As shown in Table 12, English proficiency and chatbot use do not have a strong bias. Therefore, it is unclear why these factors affect the perception of visual presentation of the ECA.

Table 12. Population of experience of using chatbot and English proficiency.

English	Use of Chatbot	
	No	Yes
Advanced	7	8
Fluent	1	4
Intermediate	4	1

4.9. Questionnaire Results After Four Rounds of Conversation

Finally, we summarize the results of the questionnaire after the four conversations. In Q1 of Table 8, the participant answered that the factors that hindered the conversation were the conversation skill (17 out of 25), responsiveness (4 out of 25), and interactivity (3 out of 25). The ECA sometimes gave inappropriate answers because the current system

employs simple QA-database-based dialogue management. This problem could be solved by employing large language models (LLMs) for dialogue management.

5. Conclusions

In this study, we developed an ECA equipped with multimodal emotion recognition and emotional facial expressions. The agent and the user had a conversation about some artwork. The study examined how the agent's response delay and facial expressions shape the user's impression of the system. The results showed that short response delays did not significantly affect evaluations with the agent except for interactivity and responsiveness; emotional facial expressions positively impacted both satisfaction and information accuracy ratings.

Future research directions include implementing more sophisticated dialogue management systems using large-scale language models (LLMs). Current systems employ simple QA-database-based dialogue management, which may result in inappropriate responses; using LLMs will enable more natural and human-like conversations, further improving user satisfaction. It is also important to conduct a more detailed analysis of the interaction between emotional expressions and response delay. Understanding how emotional expressions change users' perceptions of response delay will enable more effective agent design. Another possible topic for future research is investigating how users' cultural backgrounds and personal characteristics affect their evaluations of agents' emotional expressions.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/app15084256/s1>, Video S1: An example conversation with the developed ECA.

Author Contributions: Conceptualization, S.C.J. and A.I.; methodology, S.C.J.; software, S.C.J.; validation, S.C.J. and A.I.; formal analysis, A.I.; investigation, S.C.J.; resources, S.C.J.; data curation, S.C.J.; writing—original draft preparation, S.C.J.; writing—review and editing, A.I. and T.N.; visualization, A.I.; supervision, A.I.; project administration, A.I.; funding acquisition, A.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by JSPS KAKENHI JP23K20725 and JP21H00895.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Cassell, J. Embodied conversational interface agents. *Commun. ACM* **2000**, *43*, 70–78. [\[CrossRef\]](#)
2. Cassell, J. Embodied conversational agents: Representation and intelligence in user interfaces. *AI Mag.* **2001**, *22*, 67.
3. Allouch, M.; Azaria, A.; Azoulay, R. Conversational agents: Goals, technologies, vision and challenges. *Sensors* **2021**, *21*, 8448. [\[CrossRef\]](#)
4. Santos-Perez, M.; Gonzalez-Parada, E.; Cano-Garcia, J.M. Mobile embodied conversational agent for task specific applications. *IEEE Trans. Consum. Electron.* **2013**, *59*, 610–614. [\[CrossRef\]](#)
5. Miyake, S.; Ito, A. A spoken dialogue system using virtual conversational agent with augmented reality. In Proceedings of the 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC), Hollywood, CA, USA, 3–6 December 2012; IEEE: New York, NY, USA, 2012; pp. 1–4.
6. Fontecha, J.; González, I.; Salas-Seguín, A. Using Conversational Assistants and Connected Devices to Promote a Responsible Energy Consumption at Home. *Proceedings* **2019**, *31*, 32. [\[CrossRef\]](#)
7. André, E. Design and evaluation of embodied conversational agents for educational and advisory software. In *Handbook of Conversation Design for Instructional Applications*; IGI Global: Hershey, PA, USA, 2008; pp. 343–362.

8. Craig, P.; Roa-Seiler, N.; Rosano, F.; Díaz, M. The role of embodied conversational agents in collaborative face to face computer supported learning games. In Proceedings of the 26th International Conference on System Research, Informatics & Cybernetics, Baden-Baden, Germany, 29 July–2 August 2013.
9. Sebastian, J.; Richards, D. Changing stigmatizing attitudes to mental health via education and contact with embodied conversational agents. *Comput. Hum. Behav.* **2017**, *73*, 479–488. [\[CrossRef\]](#)
10. Laranjo, L.; Dunn, A.G.; Tong, H.L.; Kocaballi, A.B.; Chen, J.; Bashir, R.; Surian, D.; Gallego, B.; Magrabi, F.; Lau, A.Y.; et al. Conversational agents in healthcare: A systematic review. *J. Am. Med. Inform. Assoc.* **2018**, *25*, 1248–1258. [\[CrossRef\]](#)
11. Bin Sawad, A.; Narayan, B.; Alnefaie, A.; Maqbool, A.; Mckie, I.; Smith, J.; Yuksel, B.; Puthal, D.; Prasad, M.; Kocaballi, A.B. A systematic review on healthcare artificial intelligent conversational agents for chronic conditions. *Sensors* **2022**, *22*, 2625. [\[CrossRef\]](#)
12. Dumanis, I.; Smith, S. Evaluating the impact of embodied conversational agents (ECAs) attentional behaviors on user retention of cultural content in a simulated mobile environment. In Proceedings of the 7th Workshop on Eye Gaze in Intelligent Human Machine Interaction: Eye-Gaze & Multimodality, Istanbul, Turkey, 16 November 2014; pp. 27–32.
13. Pelachaud, C.; Bilvi, M. Computational Model of Believable Conversational Agents. In *Communication in Multiagent Systems: Agent Communication Languages and Conversation Policies*; Huget, M.P., Ed.; Springer: Berlin/Heidelberg, Germany, 2003; pp. 300–317. [\[CrossRef\]](#)
14. Wirzberger, M.; Schmidt, R.; Georgi, M.; Hardt, W.; Brunnett, G.; Rey, G.D. Effects of system response delays on elderly humans' cognitive performance in a virtual training scenario. *Sci. Rep.* **2019**, *9*, 8291. [\[CrossRef\]](#)
15. Strömbergsson, S.; Hjalmarsson, A.; Edlund, J.; House, D. Timing responses to questions in dialogue. In Proceedings of the Interspeech, Lyon, France, 25–29 August 2013; pp. 2584–2588. [\[CrossRef\]](#)
16. Sacks, H.; Schegloff, E.A.; Jefferson, G. A simplest systematics for the organization of turn-taking for conversation. *Language* **1974**, *50*, 696–735. [\[CrossRef\]](#)
17. Hara, K.; Inoue, K.; Takanashi, K.; Kawahara, T. Turn-Taking Prediction Based on Detection of Transition Relevance Place. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 4170–4174.
18. Yahagi, R.; Chiba, Y.; Nose, T.; Ito, A. Multimodal Dialogue Response Timing Estimation Using Dialogue Context Encoder. In *Conversational AI for Natural Human-Centric Interaction, Proceedings of the 12th International Workshop on Spoken Dialogue System Technology, IWSDS 2021, Singapore, 15–17 November 2021*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 133–141.
19. Sakuma, J.; Fujie, S.; Zhao, H.; Kobayashi, T. Improving the response timing estimation for spoken dialogue systems by reducing the effect of speech recognition delay. In Proceedings of the Interspeech, Dublin, Ireland, 20–24 August 2023; Volume 2023, pp. 2668–2672.
20. Trimboli, C.; Walker, M.B. Switching pauses in cooperative and competitive conversations. *J. Exp. Soc. Psychol.* **1984**, *20*, 297–311. [\[CrossRef\]](#)
21. Devillers, L.; Vasilescu, I.; Vidrascu, L. F0 and pause features analysis for anger and fear detection in real-life spoken dialogs. In Proceedings of the Speech Prosody, Nara, Japan, 23–26 March 2004.
22. Cassell, J.; Bickmore, T.; Billinghurst, M.; Campbell, L.; Chang, K.; Vilhjálmsón, H.; Yan, H. Embodiment in conversational interfaces: Rea. In Proceedings of the the SIGCHI Conference on Human Factors in Computing Systems, CHI '99, New York, NY, USA, 15–20 May 1999; pp. 520–527. [\[CrossRef\]](#)
23. Van Pinxteren, M.M.; Pluymaekers, M.; Lemmink, J.G. Human-like communication in conversational agents: A literature review and research agenda. *J. Serv. Manag.* **2020**, *31*, 203–225. [\[CrossRef\]](#)
24. Diederich, S.; Brendel, A.B.; Morana, S.; Kolbe, L. On the design of and interaction with conversational agents: An organizing and assessing review of human-computer interaction research. *J. Assoc. Inf. Syst.* **2022**, *23*, 96–138. [\[CrossRef\]](#)
25. Cassell, J.; Thorisson, K.R. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Appl. Artif. Intell.* **1999**, *13*, 519–538. [\[CrossRef\]](#)
26. Ochs, M.; Pelachaud, C.; Sadek, D. An empathic virtual dialog agent to improve human-machine interaction. In Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems, Estoril, Portugal, 12–16 May 2008; Volume 1, pp. 89–96.
27. Becker, C.; Kopp, S.; Wachsmuth, I. Simulating the emotion dynamics of a multimodal conversational agent. In Proceedings of the Tutorial and Research Workshop on Affective Dialogue Systems, Kloster Irsee, Germany, 14–16 June 2004; Springer: Berlin/Heidelberg, Germany, 2004; pp. 154–165.
28. Egges, A.; Kshirsagar, S.; Magnenat-Thalmann, N. Generic personality and emotion simulation for conversational agents. *Comput. Animat. Virtual Worlds* **2004**, *15*, 1–13. [\[CrossRef\]](#)
29. Ochs, M.; Sadek, D.; Pelachaud, C. A formal model of emotions for an empathic rational dialog agent. *Auton. Agents -Multi-Agent Syst.* **2012**, *24*, 410–440. [\[CrossRef\]](#)

30. Chiba, Y.; Nose, T.; Kase, T.; Yamanaka, M.; Ito, A. An analysis of the effect of emotional speech synthesis on non-task-oriented dialogue system. In Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue, Melbourne, Australia, 12–14 July 2018; pp. 371–375.
31. Yamanaka, M.; Chiba, Y.; Nose, T.; Ito, A. A study on a spoken dialogue system with cooperative emotional speech synthesis using acoustic and linguistic information. In *Recent Advances in Intelligent Information Hiding and Multimedia Signal Processing: Proceeding of the Fourteenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Sendai, Japan, 26–28 November 2018*; Springer: Berlin/Heidelberg, Germany, 2019; Volume 2, pp. 101–108.
32. Firdaus, M.; Chauhan, H.; Ekbal, A.; Bhattacharyya, P. EmoSen: Generating sentiment and emotion controlled responses in a multimodal dialogue system. *IEEE Trans. Affect. Comput.* **2020**, *13*, 1555–1566. [\[CrossRef\]](#)
33. Loveys, K.; Sagar, M.; Broadbent, E. The effect of multimodal emotional expression on responses to a digital human during a self-disclosure conversation: A computational analysis of user language. *J. Med. Syst.* **2020**, *44*, 143. [\[CrossRef\]](#)
34. Saha, T.; Reddy, S.; Das, A.; Saha, S.; Bhattacharyya, P. A shoulder to cry on: Towards a motivational virtual assistant for assuaging mental agony. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, WA, USA, 10–15 July 2022; pp. 2436–2449.
35. Jokinen, K.; Homma, K.; Matsumoto, Y.; Fukuda, K. Integration and interaction of trustworthy ai in a virtual coach—An overview of EU-Japan collaboration on eldercare. In Proceedings of the Annual Conference of the Japanese Society for Artificial Intelligence, Yokohama, Japan, 13–15 November 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 190–200.
36. Pauw, L.S.; Sauter, D.A.; van Kleef, G.A.; Lucas, G.M.; Gratch, J.; Fischer, A.H. The avatar will see you now: Support from a virtual human provides socio-emotional benefits. *Comput. Hum. Behav.* **2022**, *136*, 107368. [\[CrossRef\]](#)
37. Funk, M.; Cunningham, C.; Kanver, D.; Saikalis, C.; Pansare, R. Usable and acceptable response delays of conversational agents in automotive user interfaces. In Proceedings of the 12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Virtual, 21–22 September 2020; pp. 262–269.
38. Gnewuch, U.; Morana, S.; Adam, M.; Maedche, A. Faster is not always better: Understanding the effect of dynamic response delays in human-chatbot interaction. In Proceedings of the European Conference on Information Systems, Portsmouth, UK, 23–28 June 2018; p. 113.
39. Heeman, P.; Lunsford, R. Turn-Taking Offsets and Dialogue Context. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; Volume 2017, pp. 1671–1675. [\[CrossRef\]](#)
40. Jolibois, S.; Ito, A.; Nose, T. Multimodal Expressive Embodied Conversational Agent Design. In Proceedings of the International Conference on Human-Computer Interaction, Copenhagen, Denmark, 23–28 July 2023; Springer: Cham, Switzerland, 2023; pp. 244–249.
41. Baltrušaitis, T.; Robinson, P.; Morency, L.P. Openface: An open source facial behavior analysis toolkit. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; IEEE: New York, NY, USA, 2016; pp. 1–10.
42. Ekman, P.; Friesen, W.V. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*; Consulting Psychologists Press: Palo Alto, CA, USA, 1978.
43. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, CA, USA, 13–18 June 2010; IEEE: New York, NY, USA, 2010; pp. 94–101.
44. Demszky, D.; Movshovitz-Attias, D.; Ko, J.; Cowen, A.; Nemade, G.; Ravi, S. GoEmotions: A Dataset of Fine-Grained Emotions. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Virtual, 5–10 July 2020; pp. 4040–4054.
45. Nisimura, R.; Lee, A.; Saruwatari, H.; Shikano, K. Public speech-oriented guidance system with adult and child discrimination capability. In Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, QC, Canada, 17–21 May 2004; IEEE: New York, NY, USA, 2004; Volume 1, pp. 433–436.
46. Inoue, R.; Kurosawa, Y.; Mera, K.; Takezawa, T. A question-and-answer classification technique for constructing and managing spoken dialog system. In Proceedings of the 2011 International Conference on Speech Database and Assessments (Oriental COCOSA), Hsinchu, Taiwan, 26–28 October 2011; IEEE: New York, NY, USA, 2011; pp. 97–101.
47. Navarretta, C. Mirroring facial expressions and emotions in dyadic conversations. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia, 23–28 May 2016; pp. 469–474.
48. Stevens, C.J.; Pinchbeck, B.; Lewis, T.; Luerssen, M.; Pfitzner, D.; Powers, D.M.; Abrahamyan, A.; Leung, Y.; Gibert, G. Mimicry and expressiveness of an ECA in human-agent interaction: Familiarity breeds content! *Comput. Cogn. Sci.* **2016**, *2*, 1–14. [\[CrossRef\]](#)
49. Miura, I. Switching pauses in adult-adult and child-child turn takings: An initial study. *J. Psycholinguist. Res.* **1993**, *22*, 383–395. [\[CrossRef\]](#)
50. Mori, H. An analysis of switching pause duration as a paralinguistic feature in expressive dialogues. *Acoust. Sci. Technol.* **2009**, *30*, 376–378. [\[CrossRef\]](#)

51. Peras, D. Chatbot evaluation metrics. In *Economic and Social Development: Book of Proceedings*; Varazdin Development and Entrepreneurship Agency: Varazdin, Croatia, 2018; pp. 89–97.
52. Oosterhof, N.N.; Todorov, A. Shared perceptual basis of emotional expressions and trustworthiness impressions from faces. *Emotion* **2009**, *9*, 128. [[CrossRef](#)] [[PubMed](#)]
53. Chen, X.; Zhou, M.; Wang, R.; Pan, Y.; Mi, J.; Tong, H.; Guan, D. Evaluating Response Delay of Multimodal Interface in Smart Device. In *Design, User Experience, and Usability. Practice and Case Studies*; Marcus, A., Wang, W., Eds.; Springer: Cham, Switzerland, 2019; pp. 408–419.
54. Asaka, S.; Itoyama, K.; Nakadai, K. Improving Impressions of Response Delay in AI-based Spoken Dialogue Systems. In *Proceedings of the 2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, Pasadena, CA, USA, 26–30 August 2024; IEEE: New York, NY, USA, 2024; pp. 1416–1421.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.