Evaluating Data-Driven Co-Speech Gestures of Embodied Conversational Agents through Real-Time Interaction

Yuan he
yuan-he@live.com
KTH Royal Institute of Technology
Stockholm, Sweden

André Pereira
KTH Royal Institute of Technology
Stockholm, Sweden

Taras Kucherenko*
SEED - Electronic Arts (EA)
Stockholm, Sweden

ABSTRACT

Embodied Conversational Agents (ECAs) that make use of co-speech gestures can enhance human-machine interactions in many ways. In recent years, data-driven gesture generation approaches for ECAs have attracted considerable research attention, and related methods have continuously improved. Real-time interaction is typically used when researchers evaluate ECA systems that generate rule-based gestures. However, when evaluating the performance of ECAs based on data-driven methods, participants are often required only to watch pre-recorded videos, which cannot provide adequate information about what a person perceives during the interaction. To address this limitation, we explored use of real-time interaction to assess data-driven gesturing ECAs. We provided a testbed framework, and investigated whether gestures could affect human perception of ECAs in the dimensions of *human-likeness*, *animacy*, *perceived intelligence*, and *focused attention*. Our user study required participants to interact with two ECAs – one with and one without hand gestures. We collected subjective data from the participants' self-report questionnaires and objective data from a gaze tracker. To our knowledge, the current study represents the first attempt to evaluate data-driven gesturing ECAs through real-time interaction and the first experiment using gaze-tracking to examine the effect of ECAs' gestures.

CCS CONCEPTS

• **Human-centered computing** → **HCI design and evaluation methods**; *Interactive systems and tools*; • **Computing methodologies** → Computer graphics; Animation.

KEYWORDS

embodied conversational agent, user study, evaluation instrument, data-driven, gesture generation, gaze tracking

*Work performed while at KTH

1 INTRODUCTION

During interpersonal communication, we convey information not only through speech but also through non-verbal behaviors. It has been proven that non-verbal behaviors have critical impacts on interactions, such as revealing personality, emotions, and intimacy levels [20]. Therefore, numerous studies have focused on enabling virtual agents to mimic gestural human-to-human communicative behaviors in addition to verbal communication capabilities, thus enhancing the user's perception during interactions [11, 12, 38]. Hand gestures, as one of the most commonly observed nonverbal behaviors, have received considerable research attention when designing ECAs [7, 32].

There have been two main approaches to enabling Embodied Conversational Agents (ECAs) to generate co-speech hand gestures. One approach is to identify the patterns produced by human cognition and behavior generation, summarize the correspondence between them, then model them accordingly, which we commonly refer to as *rule-based methods* [5, 6, 16, 32]. The other approach is to employ machine learning algorithms to build models from large amounts of human communication data, which we regard as *data-driven methods* [1, 22, 26, 40]. Both types of methods have their pros and cons. Rule-based methods are more interpretable than data-based approaches, and the generated gestures can accurately reflect the agent's cultural background, gender, and other identifying characteristics. However, rule formulation requires theoretical research and a lot of human labor [8]. On the other hand, while data-driven approaches address the limitation of a lack of domain knowledge, assembling training data and mapping generated gestures to semantics remain the principal challenges.

Even though the techniques for generating gestures have been steadily optimized and improved, the methodology for evaluating their quality remains a research challenge. There is still no standardized method for quantitatively assessing gesture generation results [39]. Researchers have used objective measurements to evaluate a generative model's performance, such as comparing the joint velocity and position of generated and ground truth gestures [27]. Although the required data can be obtained readily, objective evaluation cannot represent human interpretation of gesture performance. Therefore, user studies should also be conducted to evaluate generated gestures subjectively.

Using subjective evaluation methods to evaluate the effectiveness of gesture generation serves two principal purposes: First, to determine if the gestures generated by the proposed model can affect users' perceptions of interactions in some specific dimensions;

and second, to compare the performances of different generative models regarding their perceived qualities in user experience. This study focused on serving the first purpose to evaluate the impact of one selected model on the user experience through real-time user interaction, but the experiment workflow developed in this research can also be applied to compare different models in the future.

Moreover, some objective indicators can be used to help compensate for the qualitative nature of subjective observations. For example, some studies have utilized eye-gazing behavior [9] and body posture [34] to assess user engagement. We also implemented an eye tracker to observe the human's *focused attention* when interacting with the ECA.

In most cases, rule-based systems are designed for existing platforms that feature interactive virtual agents (e.g. Greta [31]). Therefore, evaluating such models in interaction was relatively straightforward since no additional integration work was required. However, data-driven approaches present a different situation: movements are typically generated in 3D space, which is more complicated to integrate into an interactive virtual agent. This is probably why previous systems did not evaluate learning-based models in an interactive setting. Users evaluating data-driven gesture generation models were typically asked to watch clips of ECAs' behaviors, rate the generated stimuli, or perform pairwise comparisons among ECAs [39]. In such approaches, humans are positioned as observers rather than as interactors, which is not as natural as the real-world scenarios faced by ECAs. In this study, we demonstrate how putting humans in interactors' place can be achieved, and we hope to inspire others to test their learning-based models in a similar manner - through interaction.

The main contribution of this study is in evaluating interactively the use of deep-learning models to generate gestures. Our study also examines whether it is practical to use real-time interaction to measure the ECAs' performance and compare the effect of their gesture behaviors on user perception. We also developed a testbed for future researchers to benchmark their models. The link to the video shows the experimental procedure used in our user study: www.yaeh.io/research/hci/presentingbot

2 RELATED WORK

In this section, we review both subjective and objective methods applied in previous studies to evaluate ECA with gestures. Then we discuss effective approaches to user studies, including common experimental settings, measurement tools, and evaluation dimensions, from which we can draw lessons for our own evaluation.

2.1 Evaluation for ECAs with Gestures

Most researchers use both objective and subjective methods to evaluate ECAs. On one hand, the researchers evaluate the performance of data-driven models objectively. On the other hand, researchers can understand how users perceive the system through subjective assessment, which is crucial for implementing systems in real-world applications.

Huang et al. conducted a subjective experiment to compare the human perception of a robot under four modal conditions [17]. The study results indicated that participants perceived higher levels of

naturalness, *effectiveness*, and *likability* when interacting with a robot that generated gestures based on a data-driven model than when interacting with a robot that communicated exclusively through speech.

Levine et al. undertook a user study to evaluate a proposed HMM-based model. Participants were asked to perform side-by-side pairwise comparisons to determine which model produced more realistic gestures [26].

Yoon et al. evaluated the quality of gesture generation models by both objective and subjective methods. Objectively, the researchers compared the generated gestures with the ground truth in the original video. Subjectively, researchers adopted *anthropomorphism* and *likeability* from the Godspeed questionnaire [4] as measurements [40].

In 2020, a group of researchers launched the first gesture generation challenge [24]. This event provided a benchmark for practitioners in the gesture generation domain to compare their models by using a common dataset, visualization setup, and evaluation process. User studies were conducted in the challenge by watching video clips of different model results. In our study, we also focused on conducting user experiments, but through real-time interaction.

2.2 Approaches to Conduct User Studies on ECAs with Gesture Functions

2.2.1 User Study through Interaction. Real-time interactions with users are an effective means of assessing the performance of an ECA system or robot. In evaluating gesture-generation methods, it is also common to have participants rate the interaction experience. As an example, Salem et al. required participants to interact with a humanoid robot in both unimodal (speech only) and multimodal (speech with consistent or inconsistent gestures) conditions. The participants were then asked to rate a variety of aspects, including *human-likeness*, *likeability*, *shared reality* with the robot, judgments of acceptance, and future contact intentions [33]. Asly and Tapus examined personality traits (introversion and extroversion) in the design of gestural representations of robots, and corresponding user studies were conducted via real-time interactions [2].

Nevertheless, our literature reviews found that real-time interaction studies tend to focus on rule-based methods. In the case of measuring data-driven methods, user research is often conducted only in a one-way manner without the ability to communicate in real-time with the users. As an example, in Le and Pelachaud's perceptual experiment, participants were asked to rate the performance of gesture generation after viewing a video of an NAO robot narrating a fairy tale [25]. GENEA 2020 Gesture-Generation Challenge [24] organizers provided participants with a user interface for measuring performance. This allowed them to rate their performance after viewing a video of the ECA, although the experiment did not involve interaction [19]. A potential reason why data-driven methods were barely evaluated in real-time interactions could be the absence of appropriate testbeds. Due to the fact that data-driven methods are applied to generate gestures, the feedbacks from the ECAs are not fully predictable based on user input, which makes it more difficult to bind animations to ECAs. To tackle this limitation, Nagy et al. developed a modular framework in the Unity3D environment that enabled the ECA to communicate with data-driven

models in real-time [28]. Hence, it provided an infrastructure for future researchers to evaluate data-driven gesture generation models by adding modules. In this study, we adopted this framework and modified it to suit our scenario.

2.2.2 Measurements to Evaluate Gesturing ECAs. Gestures have a variety of effects on humans, including cognitive, emotional, and behavioral, as well as on their performance of tasks [35]. However, because human communication is inherently multimodal, it is difficult to measure the contribution of gestures to human perception in isolation. To the best of our knowledge, there is still no unified questionnaire that can directly evaluate the quality of gestures in ECAs. Still, existing questionnaires can indirectly measure the role that gestures play in the system by assessing people's overall perception of ECA [14].

Researchers typically choose the tools for evaluation based on their research interests. Aly and Tagus' focus was on checking whether people can accurately identify robots that match their personality through non-verbal cues, and on testing whether gestures can improve expressiveness. In this study, experimenters used the Big 5 Inventory Test to distinguish between introverted and extroverted participants. Then experimenters adopted a questionnaire including 24 questions on a seven-point Likert scale to evaluate the gestures generation model, which is mapped by human personality traits. The survey covered user preferences, robot personality congruence, user engagement, robot expressiveness, robot gesture, voice synchronization, etc. [2]. Salem et al. investigated the impact of robot gestures on humans' perceptions of anthropomorphism and likability. Researchers developed two sets of questionnaires to assess participants' perception of the robot and their performance on task-accomplishing interactions in the experiment. Perception-related questions included *humanlikeness*, *likeability*, *shared reality* [10], and *future contact intentions*. To assess task performance, the authors used both objective and subjective ratings. The subjective indicator was a five-point Likert-scale question asking participants to rate their capacity to solve the challenge. The objective metric was the error rate of task completion [33].

Due to the lack of a standard evaluation scale, most studies have used self-developed questionnaires to assess gesture generation. Fitrianie et al. reviewed 81 papers relating to intelligent virtual agents, and found that more than 76% of these studies employed measurement instruments that were only applied once [13]. Self-designed questionnaires do not have any guarantee of reliability. Additionally, constantly creating new measurement tools rather than reusing well-tested instruments makes it harder to replicate and compare experiments. Therefore, we applied validated and commonly-used measurements to our study to ensure its rigour and to facilitate comparison with other studies.

2.2.3 Evaluating ECAs through Behavior Observation. Researchers can complement questionnaires with behavioral observations. Kooijmans et al. developed a software that can collect a variety of objective data in human-robot interactions, including sound, vision, person identification, motion, body contact, and robotic behavior [21]. Some other commonly utilized behavioral indicators include proxemics, postures [34], eye-gazes, etc. Bailenson et al. measured the social distance between a person and a virtual agent to determine co-presence in a virtual environment [3]. Nakano et al. used

eye-gaze behavior as an objective indicator of a person's attention when talking to a virtual agent [29]. Behavioral indicators have been shown to be an effective complementary investigation method in conjunction with subjective questionnaires.

3 SYSTEM DESIGN AND IMPLEMENTATION

The design and development of our ECA system were based on a realistic scenario of a virtual robot making a presentation. The virtual agent can present to participants six classical Roman monuments according to the participant's voice command. The script and the experiment system will be open-sourced.

3.1 System Architecture

The system was based on the modular framework developed by Nagy et al [28] [1]. The modified system architecture could be transplanted to any situation that involves a presenter, such as a lecturer or a museum guide. The system architecture consisted of four modules (see Figure 1). Module 1 was the gesture generation model known as *Gesticulator* developed by Kucherenko et al. [23]. *Gesticulator* is a data-driven method that can automate hand gestures based on both audio and text. Module 2 contained the speech synthesis model developed by Székely et al. [36], which was used to generate the agent's audio output. This synthesizer was selected because it shared the same training dataset with our chosen gesture generation model, *Gesticulator* [23]. Module 3 was in charge of communication with other modules and contained presentation content. We developed a database in the Unity3D engine using C#. The database allowed system users to customize the presentation's content by providing corresponding command words, pictures, audio, and speech text. Module 4 provided the virtual scene, which could be replaced by any customized 3D model. In this scene, the Gesticulator can communicate with the virtual robot through the ActiveMQ message broker. The current scene contains a virtual robot standing in front of a painting frame. The paintings are some famous Roman monuments. During interaction with participants, the painting content would swap like a presentation slider.
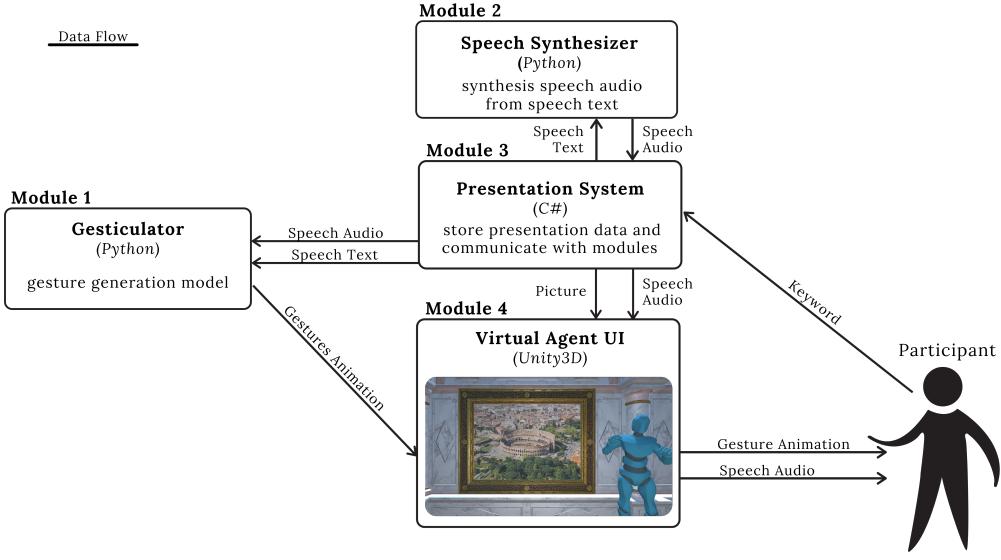
3.2 Modules Communication and Information Flow

To avoid the system latency in response to participants' voice commands, we generated gestures in advance rather than in real-time. We input both presentation text and audio files into *Gesticulator*, and *Gesticulator* exported the generated gestures as joint angle rotations in CSV files.

Before starting the experiment, we prepared four groups of presentation data. First, we wrote the presentation contents and saved them as text files. Second, we synthesized speech from the text using the speech synthesizer [36] and saved the outputs as audio files. Third, we selected pictures that would be shown in the presentation. Fourth, we decided on the keywords that can invoke ECA's actions. We stored the four groups of resources and generated gestures in the database of Module 3.

When the presentation system (Module 3) receives a keyword from a participant, it would send the related audio clip and image

---

[1]https://github.com/nagyrajmund/gesturebot

Figure 1: The system architecture.

to the user interface (Module 4), while enabling the ECA to perform the corresponding gesture motion.

4 USER STUDY

4.1 Experiment Conditions and Hypotheses

The purpose of this study was to evaluate the effects of co-speech gestures of the ECA system on user perception through interacting in real-time. Therefore, we identified two experimental conditions and referred to them as 'Gesturing Condition' and 'Idle Condition'.

In both conditions, the virtual agent presents the Roman monuments selected by the participant's voice command. During the 'Gesturing Condition', the virtual agent produces gestures while introducing the monument displayed on the painting. During 'Idle Condition,' the virtual agent's hands merely hang naturally without gestures, accompanied by natural breathing and body microdynamics when presenting.

In order to collect each participant's subjective opinion towards both conditions, we applied a within-subjects experimental method, where each subject was exposed to two conditions. Instead of comparing different data-driven models, this study focused on examining the feasibility and approaches for evaluating data-driven gesture generation models in real-time, but the system design left room for future comparisons of multiple models.

In the study, each participant was assigned to both conditions. To exclude potential bias caused by ordering, we controlled the occurrence order of the two conditions. Fourteen participants were exposed to the gesturing condition first, followed by the idle condition; the remaining 14 participants were exposed to the opposite condition sequence. The robot body colors in both experimental conditions were set to blue and green, respectively, so that users could differentiate the two conditions. In order to avoid potential bias resulting from color preference, we counterbalanced the two colors and the two experimental conditions.

There was no significant difference in all those dimensions. The exact analytical values are shown in Figure 2.

We assumed that the ECA's hand gestures could improve human perception of the agent's realism and attract more attention during the interaction. To test this assumption, we propose two major hypotheses.
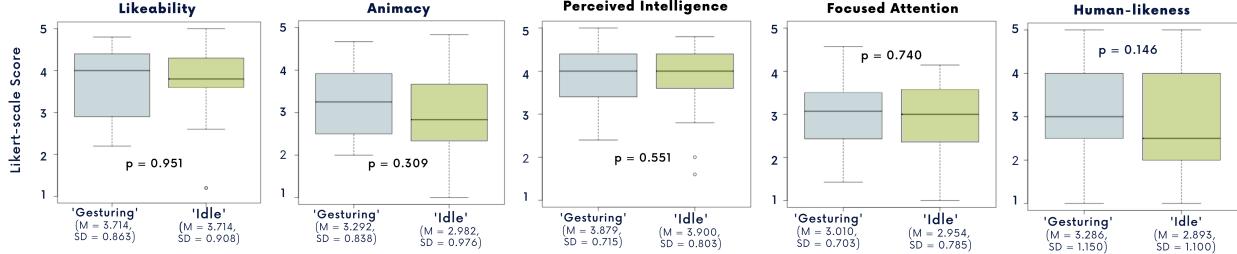
H1. The ECA with data-driven generated gestures is perceived better in *likeability*, *animacy*, *perceived intelligence*, and *human-likeness* than the ECA standing with idle posture.

H2. The ECA with data-driven generated gestures can attract more participant's attention than the ECA standing with idle posture.

4.2 Selection of Measuring Instruments

To ensure our study's rigour, validity, and reproducibility, we chose to use a validated scale as our primary measurement tool.

To examine hypothesis 1, we selected *likeability*, *animacy*, and *perceived intelligence* dimensions from the Godspeed questionnaire [4] and asked a direct question regarding *human-likeness*, which we took from the GENEA Challenge 2020 [24]. We asked participants to rate the interaction task subjectively in each condition. To evaluate hypothesis 2, both subjective and objective methods were adopted. Subjective assessment was measured using seven entries in the *focused attention* dimension of User Engagement Scale (UES) [30]; Objective assessment was inferred from the gaze data collected by the eye tracker. It is believed that a longer gaze duration attracts more focused attention.

Figure 2: The questionnaire results on *likeability, animacy, perceived intelligence, focused attention* and *human-likeness*

4.3 Experiment Procedure

Participants were exposed to two experimental conditions in succession. First, a selection menu guided the participant to select a keyword of interest. The participant would use a voice command to execute the selection. After receiving the command, the virtual robot would start presenting. Each speech segment lasted about 40 seconds. Each condition consisted of two selection interactions. After completing each experimental condition, the participant rated the interaction with each agent. From the start of the system, the user's eye gaze data was continuously recorded by a Tobii eye tracker at the bottom of the screen [37].

After the experiment, we interviewed each participant about preferences and perceived differences between the two conditions.

At the end, the experimenter gave a short debriefing to each participant to explain the intention of the study.

4.4 Participants

A total of 28 participants (10 female, 18 male) were recruited by online and offline posters published in the university community (Facebook forums, Whatsapp groups, Telegram groups, advertisement boards and building doors on campus). After completing the experiment, participants were rewarded with a voucher from a local supermarket.
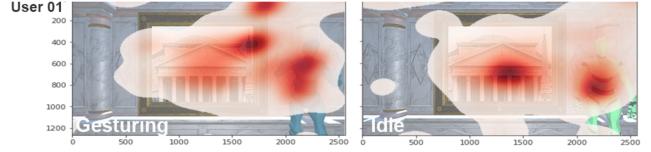
Participants were diverse in age (18–36, Mean=24.39, SD=13.71) and nationality. To ensure the quality of the experiment, all participants were required to speak and listen to English at an excellent level.

5 RESULTS AND FINDINGS

5.1 Quantitative Results

5.1.1 Questionnaire. The *animacy* subscale consisted of six items (α = .90), the *likeability* subscale consisted of five items (α = .91), the *perceived intelligence* subscale consisted of five items (α = .85), and the *focused attention* subscale consisted of seven items (α = .88). The measurements in all dimensions had good internal consistency.

We used a t-test to rule out a possible ordering bias to the data. Then, we performed the Shapiro-Wilk normality test on the mean values of each dimension. The results showed that *animacy* and *focused attention* could be considered to obey a normal distribution and could be tested by the paired t-test. In contrast, the distributions of the other three dimensions did not have normality and were suitable to be tested by a non-parametric test.

Figure 3: The heatmap shows gazing time (User 1 as example).

Figure 4: The division of presented object (painting) and the presenter (robot)

A paired samples t-test was performed to compare *animacy* and *focused attention* in 'Gesturing' and 'Idle' conditions.

An exact Wilcoxon-Pratt Signed-Rank Test was performed to compare *likeability, perceived intelligence* and *human-likeness* in the 'Gesturing' and 'Idle' conditions.
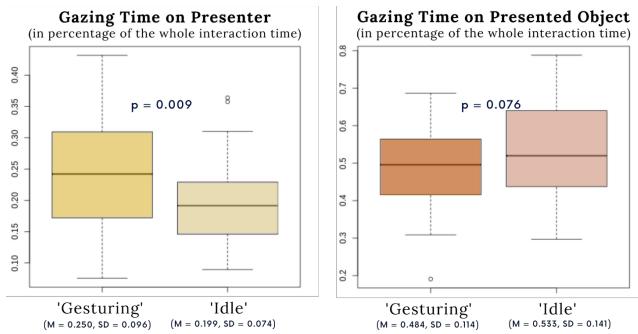
5.1.2 Eye-gazing Analysis. The original gaze data was recorded in the format of x, y coordinates for each frame. To establish an overall understanding of the subject's gaze behavior, we generated a separate heatmap (see Figure 3) for the two sets of coordinate data corresponding to each subject. Then, we divided the coordinate area corresponding to the screen into two parts, the presented object (painting) and the presenter (robot) (see Figure 4). We obtain the ratio of gaze time corresponding to the two areas to the total interaction time and we subsequently express this proportion in terms of gazing time.

Figure 5 displays the distribution of results of eye gazing time on the virtual agent (presenter) and the painting (presented object) respectively.

A paired samples t-test was performed to compare eye gazing time on virtual agent (presenter) in 'Gesturing' and 'Idle' conditions. There was significant difference in eye gazing time on virtual agent

(presenter) between 'Gesturing' (M = 0.250, SD = 0.096) and 'Idle' (M = 0.199, SD = 0.074); t(df) = 2.794, p = 0.009.

A paired samples t-test was performed to compare eye gazing time on the painting (presented object) in 'Gesturing' and 'Idle' conditions. There was no significant difference in eye gazing time on the painting (presented object) between 'Gesturing' (M = 0.484, SD = 0.114) and 'Idle' (M = 0.533, SD = 0.141); t(df) = -1.844, p = 0.076.

Figure 5: The boxplots of eye gazing time on presenter and presented object.

5.1.3 Findings.

Likeability. As the result of our analysis in the *Likeability* dimension shows, the participants did not prefer the virtual agent with gestures over the one with idle micro-movements. Figure 2 shows more participants rated the gesturing robot with low scores than the idle robot in *likeability* dimension, though the mean score of the gesturing robot was still higher. The result was consistent with the opinions received from the after-experiment interviews.

Human-likeness. We found no significant difference between the two conditions. Both agents received extreme ratings in *human-likeness* dimension.

Animacy. In the analysis, the virtual agent with gestures received higher mean scores in *animacy*. However, these differences did not reach statistical significance.

Perceived Intelligence. Perceived intelligence was the only dimension, where the 'Idle' robot received slightly higher mean score than the 'Gesturing' robot. But the difference was not significant.

Focused Attention. According to the questionnaire results, neither condition received more attention than the other. Nevertheless, eye gazing data analysis showed a significant difference when comparing gazing time on the agents and on the screen. Figure 2 illustrates that participants spent more time on the presenter's body than presented object under 'Gesturing' condition and vice versa.

5.2 Qualitative Result

Two questions were asked in the after-experiment interviews:
   Q1. Have you noticed the difference between two robots?
   Q2. Which one do you prefer and why?

Regarding Q1, out of 28 participants, 24 mentioned color differences, 17 thought the two virtual agents' voice (pitch and tone) was different, and 15 called attention to the difference in hand movements.

Regarding the Q2, 13 participants preferred the virtual agent with gestures because it was more vivid, more natural, more expressive, friendlier, and that the movement could reduce boredom. Nine participants preferred the virtual agent without gestures because it looked calmer and less distracting. Six participants reported that it was hard to say which robot they preferred or that they liked neither of them. One of the participants explained that in this experiment, the virtual agent with gestures was too intensive and not natural enough. He argued that human-like gestures should only appear at specific times and not be continuously and uninterruptedly output. The virtual agent in the idle condition, on the other hand, was too stiff. One participant pointed out that the gesture animation appeared random and not anthropomorphic enough due to the virtual agent's inability to make iconic and deictic gestures.

As the result of our analysis in the *likeability* dimension, the participants did not prefer the virtual agent with gestures over the one with idle micro-movements. Figure 2 showed more participants rated the gesturing robot with low scores than the idle robot in *likeability* dimension, though the mean score of the gesturing robot was still higher. The result was consistent with the opinions received from the after-experiment interviews.

We found no significant difference between the two conditions. For both agents, participants gave either very high or very low scores on the *human-likeness* dimension.

In the analysis, the virtual agent with gestures received higher mean scores in *animacy*. However, these differences did not reach statistical significance.

Perceived intelligence was the only dimension, where the 'Idle' robot received slightly higher mean score than the 'Gesturing' robot. But the difference was not significant.

According to the questionnaire results, neither condition received more attention than the other. Nevertheless, eye gazing data analysis showed a significant difference when comparing gazing time on the agents and on the screen. Figure 2 illustrates that participants spent more time on the presenter's body than the presented object under 'Gesturing' condition and vice versa.

6 DISCUSSION

In this section, we discuss the results within the context of the proposed hypotheses and discuss the validity and usefulness of the suggested experimental methodology.

6.1 H1: Improvement of *likeability, animacy, perceived intelligence,* and *human-likeness*

Although the mean scores showed a slight indication that the ECA with data-driven generated gestures was perceived better in *likeability*, *animacy* and *human-likeness* dimensions and that the ECA with idle posture received higher mean rate in the dimension of *perceived intelligence*, the results were not statistically significant. Hence, we could not confirm H1.

Other researchers, however, experienced different findings. For instance, in Salem et al.'s study, participants perceived the robot

that gestured while it talked to be more likeable than the robot that only spoke verbally [33]. Huang and Mutlu found that robots with gestures were more likeable than those without gestures [17]. Salem et al. [33] and Ishi et al. [18] reported that a robot that could make co-speech gestures was considered more human-like than a standstill robot.

By combining the participants' interview feedback with our analytical results, we speculate that the following factors may affect participants' ratings. First, the gesture-generation model we selected produced too many gestures that did not pause appropriately. In the presentation scenario, the presenters' excessive body movements may have distracted people from focusing on the presentation's content. Second, the dialogue system we adopted was not natural enough, and unexpected delays and lags could have affected the participants' *likeability*.

6.2 H2: Increased attention

Even though H2 was not supported by questionnaire results, it was confirmed by the result from eye gaze data. As the eye gazing results indicated, the gesturing agent attracted more attention to its body. On the other hand, the idle agent enabled participants to focus on the object being presented.

We were aware that using two different measurement tools could lead to contrasting results. The reason might be that when the user interface is complicated, the information obtained by the questionnaire is merely a general impression. We cannot know exactly which element on the interface is making the perceptual difference. In this case, the behavioral observation tool can be more sensitive for identifying the aspects that affect human perception.

6.3 Experimental Methodology

Our study also demonstrated the feasibility of evaluating data-driven motion generation models through real-time interactions. The experiment system worked smoothly and without unexpected interruption. From our experience in this study, we concluded three advantages to conducting evaluation practices through real-time interaction rather than watching video clips.

First, evaluating through real-time interaction is more natural than watching video clips because it simulates real-world communication with humans. Participants notice more details when they interact with agents. For example, the way a virtual agent reacts to some human behaviors can be perceived as an indication of intelligence.

Second, the interaction process can be designed to adapt to the actual use scenarios. Evaluating specific use scenarios could lead to varied results. For example, *likeability* could be highly dependant on the use scenario. A virtual agent that is designed to be extremely vivacious might be perceived as likeable when presented as a dance teacher but less likeable as a technical expert.

Third, involving real-time interaction in evaluation practices allows for more possibilities to employ behavioral observation methods. By analyzing human responses to the system during the interaction, such as posture and eye gazing, experimenters could obtain richer results.

7 LIMITATIONS AND FUTURE WORK

The current study has several limitations. The small number of participants meant that we could not get a sufficiently diverse sample. Our literature review shows that different cultures and sexes perceive gestures differently. In light of this, we speculate we will obtain richer results if we can recruit more participants. If there is a chance for a follow-up study, we will launch the existing test system online to reach more participants.

This study evaluated only one data-driven model for its impact on user perception due to time constraints. The existing system design permits the integration of multiple virtual robots and gesture models, enabling future comparisons between the models.

The experiment was conducted in a desktop environment, where users could not interact in a fully immersive way with the virtual robot. As indicated by our findings, objective user behavior data can be beneficial for supplementary and productive outcomes. For further study, it could be useful to enable the system to collect multiple types of data. Virtual reality environments may be able to immerse users while also collecting objective data on user behavior (e.g., social distance).

8 CONCLUSION

In this paper, we proposed an approach for evaluating data-driven gesture generation models in an interaction. To test the framework we compared two conditions: one with idle gestures and one with gestures generated by a machine-learning model. Even though the questionnaire results did not indicate a significant difference in all the dimensions, the analysis of eye gazing data has supported the second hypothesis — **the ECA with data-driven generated gestures can attract more participants' attention in certain areas than the ECA standing with idle posture**.

The finding can be used as an empirical basis for the design of ECA applications in presentation scenarios. For example, a math teacher agent would require students to pay more attention to the slides than to the agent, so gestures would be less of a factor. In contrast, an NPC in a game that tells the history of a particular world may benefit from incorporating gestures to attract the player's attention and make the character more expressive.

In addition to subjective questionnaires, objective data can provide valuable insight. Our study discussed and validated the use of objective behavioral data to complement structured questionnaire data. Our study used gaze time to infer *focused attention* — longer gaze times are considered to indicate greater focus [15].

Our experimental setup serves as an example of how real-time evaluations of data-driven gesture models can be performed efficiently and effectively. The results of our study suggest that conducting user studies with behavioral observation methods, such as eye-tracking, can be beneficial.

REFERENCES

[1] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 487–496.

[2] Amir Aly and Adriana Tapus. 2013. A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human-robot interaction. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 325–332.

[3] Jeremy N Bailenson, Eyal Aharoni, Andrew C Beall, Rosanna E Guadagno, Aleksandar Dimov, and Jim Blascovich. 2004. Comparing behavioral and self-report measures of embodied agents' social presence in immersive virtual environments. In *Proceedings of the 7th Annual International Workshop on PRESENCE*. IEEE, 1864–1105.

[4] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics* 1, 1 (2009), 71–81.

[5] Justine Cassell, Tim Bickmore, Lee Campbell, Hannes Vilhjalmsson, and Hao Yan. 2000. Human conversation as a system framework: Designing embodied conversational agents. *Embodied conversational agents* (2000), 29–63.

[6] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 1994. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*. 413–420.

[7] Justine Cassell and Hannes Vilhjálmsson. 1999. Fully embodied conversational avatars: Making communicative behaviors autonomous. *Autonomous agents and multi-agent systems* 2, 1 (1999), 45–64.

[8] Chung-Cheng Chiu and Stacy Marsella. 2011. How to train your avatar: A data driven approach to gesture generation. In *International Workshop on Intelligent Virtual Agents*. Springer, 127–140.

[9] Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M Rehg. 2018. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *Proceedings of the European conference on computer vision (ECCV)*. 383–398.

[10] Gerald Echterhoff, E Tory Higgins, and John M Levine. 2009. Shared reality: Experiencing commonality with others' inner states about the world. *Perspectives on Psychological Science* 4, 5 (2009), 496–521.

[11] Marc Fabri, DJ Moore, and DJ Hobbs. 2002. Expressive agents: Non-verbal communication in collaborative virtual environments. *Proceedings of Autonomous Agents and Multi-Agent Systems (Embodied Conversational Agents)* (2002).

[12] Jasper Feine, Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2019. A taxonomy of social cues for conversational agents. *International Journal of Human-Computer Studies* 132 (2019), 138–161.

[13] Siska Fitrianie, Merijn Bruijnes, Deborah Richards, Amal Abdulrahman, and Willem-Paul Brinkman. 2019. What are We Measuring Anyway? -A Literature Survey of Questionnaires Used in Studies Reported in the Intelligent Virtual Agent Conferences. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. 159–161.

[14] Siska Fitrianie, Merijn Bruijnes, Deborah Richards, Andrea Bönsch, and Willem-Paul Brinkman. 2020. The 19 unifying questionnaire constructs of artificial social agents: An iva community analysis. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–8.

[15] Alexandra Frischen, Andrew P Bayliss, and Steven P Tipper. 2007. Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychological bulletin* 133, 4 (2007), 694.

[16] Chien-Ming Huang and Bilge Mutlu. 2013. Modeling and Evaluating Narrative Gestures for Humanlike Robots.. In *Robotics: Science and Systems*. 57–64.

[17] Chien-Ming Huang and Bilge Mutlu. 2014. Learning-based modeling of multimodal behaviors for humanlike robots. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 57–64.

[18] Carlos T Ishi, Daichi Machiyashiki, Ryusuke Mikata, and Hiroshi Ishiguro. 2018. A speech-driven hand gesture generation method and evaluation in android robots. *IEEE Robotics and Automation Letters* 3, 4 (2018), 3757–3764.

[19] Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, and Gustav Eje Henter. 2021. HEMVIP: Human evaluation of multiple videos in parallel. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 707–711.

[20] Mark L Knapp, Judith A Hall, and Terrence G Horgan. 2013. *Nonverbal communication in human interaction*. Cengage Learning.

[21] Tijn Kooijmans, Takayuki Kanda, Christoph Bartneck, Hiroshi Ishiguro, and Norihiro Hagita. 2007. Accelerating robot development through integral analysis of human–robot interaction. *IEEE Transactions on Robotics* 23, 5 (2007), 1001–1012.

[22] Taras Kucherenko, Dai Hasegawa, Naoshi Kaneko, Gustav Eje Henter, and Hedvig Kjellström. 2021. Moving fast and slow: Analysis of representations and post-processing in speech-driven automatic gesture generation. *International Journal of Human–Computer Interaction* (2021), 1–17.

[23] Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 242–250.

[24] Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. 2021. A large, crowdsourced evaluation of gesture generation systems on common data: The GENEA Challenge 2020. In *26th International Conference on Intelligent User Interfaces*. 11–21.

[25] Quoc Anh Le and Catherine Pelachaud. 2012. Evaluating an expressive gesture model for a humanoid robot: Experimental results. In *Submitted to 8th ACM/IEEE International Conference on Human-Robot Interaction*.

[26] Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun. 2010. Gesture controllers. In *ACM SIGGRAPH 2010 papers*. 1–11.

[27] Mina Marmpena, Fernando Garcia, and Angelica Lim. 2020. Generating robotic emotional body language of targeted valence and arousal with conditional variational autoencoders. In *Companion of the 2020 ACM/IEEE international conference on human-robot interaction*. 357–359.

[28] Rajmund Nagy, Taras Kucherenko, Birger Moell, André Pereira, Hedvig Kjellström, and Ulysses Bernardet. 2021. A Framework for Integrating Gesture Generation Models into Interactive Conversational Agents. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '21)*. 1779–1781.

[29] Yukiko I Nakano and Ryo Ishii. 2010. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In *Proceedings of the 15th international conference on Intelligent user interfaces*. 139–148.

[30] Heather L O'Brien, Paul Cairns, and Mark Hall. 2018. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies* 112 (2018), 28–39.

[31] Catherine Pelachaud. 2015. Greta: an interactive expressive embodied conversational agent. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*. 5–5.

[32] Brian Ravenet, Catherine Pelachaud, Chloé Clavel, and Stacy Marsella. 2018. Automating the production of communicative gestures in embodied characters. *Frontiers in psychology* 9 (2018), 1144.

[33] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2013. To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics* 5, 3 (2013), 313–323.

[34] Jyotirmay Sanghvi, Ginevra Castellano, Iolanda Leite, André Pereira, Peter W McOwan, and Ana Paiva. 2011. Automatic analysis of affective postures and body motion to detect engagement with a game companion. In *Proceedings of the 6th international conference on Human-robot interaction*. 305–312.

[35] Shane Saunderson and Goldie Nejat. 2019. How robots influence humans: A survey of nonverbal communication in social human–robot interaction. *International Journal of Social Robotics* 11, 4 (2019), 575–608.

[36] Éva Székely, Gustav Eje Henter, Jonas Beskow, and Joakim Gustafson. 2019. Spontaneous Conversational Speech Synthesis from Found Data.. In *INTERSPEECH*. 4435–4439.

[37] Tobii Pro AB. 2014. Tobii Pro Lab. Computer software. http://www.tobiipro.com/

[38] Alessandro Vinciarelli, Maja Pantic, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, Francesca D'Errico, and Marc Schroeder. 2011. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing* 3, 1 (2011), 69–87.

[39] Pieter Wolfert, Nicole Robinson, and Tony Belpaeme. 2022. A review of evaluation practices of gesture generation in embodied conversational agents. *IEEE Transactions on Human-Machine Systems* (2022).

[40] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 4303–4309.