DEGREE PROJECT IN INFORMATION AND COMMUNICATION TECHNOLOGY,
SECOND CYCLE, 30 CREDITS
STOCKHOLM, SWEDEN 2022

Evaluating Generated Co-Speech Gestures of Embodied Conversational Agent(ECA) through Real-Time Interaction

Yuan He

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

Swedish Title

Utvärdering av genererade samspråkliga gester hos Embodied Conversational Agent (ECA) genom interaktion i realtid

Author

Yuan He <yuanhe@kth.se>
Master in Human-Computer Interaction & Design
KTH Royal Institute of Technology

Place for Project

Department of Intelligent Systems
KTH Royal Institute of Technology

Examiner

Jonas Beskow
KTH Royal Institute of Technology

Supervisor

Taras-Svitozar Kucherenko
André Tiago Abelho Pereira
KTH Royal Institute of Technology

Abstract

Embodied Conversational Agents (ECAs)' gestures can enhance human perception in many dimensions during interactions. In recent years, data-driven gesture generation approaches for ECAs have attracted considerable research attention and effort, and methods have been continuously optimized. Researchers have typically used human-agent interaction for user studies when evaluating systems of ECAs that generate rule-based gestures. However, when evaluating the performance of ECAs that generate gestures based on data-driven methods, participants are often required to watch prerecorded videos, which cannot provide an adequate assessment of human perception during the interaction. To address this limitation, we proposed two main research objectives: First, to explore the workflow of assessing data-driven gesturing ECAs through real-time interaction. Second, to investigate whether gestures could affect ECAs' human-likeness, animacy, perceived intelligence, and humans' focused attention in ECAs. Our user study required participants to interact with two ECAs by setting two experimental conditions with and without hand gestures. Both subjective data from the participants' self-report questionnaire and objective data from the gaze tracker were collected. 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 ECA gestures. The eye-gazing data indicated that when an ECA can generate gestures, it would attract more attention to its body.

Keywords

Embodied Conversational Agent(ECA); User Study; Evaluation Instrument; Gesture Generation; Gaze-tracking

Sammanfattning

Förkroppsligade konversationsagenter (Embodied Conversational Agents, ECAs) gester kan förbättra människans uppfattning i många dimensioner under interaktioner. Under de senaste åren har datadrivna metoder för att generera gester för ECA:er fått stor uppmärksamhet och stora ansträngningar inom forskningen, och metoderna har kontinuerligt optimerats. Forskare har vanligtvis använt sig av interaktion mellan människa och agent för användarstudier när de utvärderat system för ECA:er som genererar regelbaserade gester. När man utvärderar prestandan hos ECA:er som genererar gester baserat på datadrivna metoder måste deltagarna ofta titta på förinspelade videor, vilket inte ger en adekvat bedömning av människans uppfattning under interaktionen. För att åtgärda denna begränsning föreslog vi två huvudsakliga forskningsmål: För det första att utforska arbetsflödet för att bedöma datadrivna ECA:er för gester genom interaktion i realtid. För det andra att undersöka om gester kan påverka ECA:s människoliknande, animerade karaktär, upplevd intelligens och människors fokuserade uppmärksamhet i ECA:s. I vår användarstudie fick deltagarna interagera med två ECA:er genom att ställa in två experimentella villkor med och utan handgester. Både subjektiva data från deltagarnas självrapporterande frågeformulär och objektiva data från gaze tracker samlades in. Såvitt vi vet är den aktuella studien det första försöket att utvärdera datadrivna ECA:er med gester genom interaktion i realtid och det första experimentet där man använder blickspårning för att undersöka effekten av ECA:s gester. Uppgifterna om blickspårning visade att när en ECA kan generera gester skulle den locka mer uppmärksamhet till sin kropp.

Nyckelord

Förkroppsligad konversationsagent (ECA); Användarstudie; Utvärderingsinstrument; Generering av gester; Spårning av blickar

Acknowledgements

I would like to express my deep appreciation to my primary supervisor, Taras Kucherenko, who guided me throughout this project with great patience. I would also like to thank my supervisor André Tiago Abelho Pereira, who supported me a lot in conducting the experiment and offered deep insight into the user study.

I would particularly like to thank my examiner, Professor Jonas Beskow, for providing final advice on adjustments and coordination of the presentation.

I wish to acknowledge Rajmund Nagy for providing technical support. I wish to extend my special thanks to Éva Székely for offering the speech synthesis system.

I would like to show my gratitude to the EIT foundation for providing the multi-disciplinary education program and giving me a second chance to change my career. All the courses and assistance provided by the two host universities, KTH and the University of Paris Saclay, were greatly appreciated.

Finally, I wish to extend my special thanks to my family and friends for all the emotional support during my master's period.

He, Yuan
Stockholm, November,2021

Contents

Chapter 1

Introduction

1.1 Research Scope

Interpersonal communication does not only convey information through speech but also non-verbal behaviors. It has been proven that non-verbal behaviors have critical impacts on interactions, such as revealing personality, emotions, and intimacy levels [49]. Therefore, numerous studies have focused on enabling virtual agents to mimic human-to-human communicative behaviors in addition to verbal communication capabilities, thus enhancing the user's perception during interactions [23][84][22]. Gestures, as one of the most commonly observed nonverbal behaviors, have received considerable research attention [71][14].

Researchers have mainly attempted two ways to enable Embodied Conversational Agents (ECAs) to generate co-speech gestures. One way is to identify the cognitive patterns between the human mind and behavior generation, summarize the correspondence between them, then model them accordingly, which we commonly refer to as *rule-based methods*. The other way is to employ machine learning algorithms to build models from large amounts of human communication data, which we regard as *data-driven methods*. Both types of methods have their pros and cons. Rule-based methods have stronger interpretability than data-based approaches, and the resulting gestures may accurately match the cultural background, gender, and other identifying characteristics of the agent should display. However, rule formulation requires theoretical research and a lot of human labor [16]. The Data-driven approach addresses the limitation of lack of domain knowledge, but the

collection and collation of training data and the mapping of the generated gestures to semantics remain the main challenges. In this study, we employed Gesticulator, which is a generative model trained from both speech and text modalities [58].

As gesture generation methods were progressively optimized and improved, the methodology to evaluate the quality of generated gestures remains an open research area. There is still a lack of a common standard method for quantitative assessment of gesture generation results. Some studies employed objective methods to evaluate the performance of the generation model, for instance, comparing the joint velocities and positions of the generated gesture with the ground truth human gesture [66]. Although the required data can be obtained readily, objective evaluation could not represent the humans' sense of gesture performance [87]. Hence, subjective evaluation of the generated gestures by Humans through user studies is also essential.

Evaluating the effectiveness of gesture generation through subjective evaluation methods serves two principal purposes: First, to inspect that the gestures generated by the proposed model can impact users' perceptions of interactions in some specific dimensions; and second, to compare the performances of different generative models regarding their performance in user experience. Our study focused on serving the first purpose to evaluate the impact of one selected model on the user interaction experience through real-time user interaction, but the experiment workflow developed in this research can be applied to compare the models in the future.

Evaluating human-ECA interactions through user study requires clarification of two crucial questions. One of them is, "What to measure?".In addition to the two most frequently rated metrics, namely Speech-gesture correlation/appropriateness and human-likeness, several studies also concerned human perception of other attributes of systems, such as user engagement [70], perceived likeability [85], and trust [77]. The dimensions of evaluation also depend on the use scenarios. Nowadays, ECAs are applied in a variety of fields, such as service [72], digital health [53] and training [62].User concerns about the characteristics that ECAs need to exhibit will vary depending on the purpose of the application.For example, when an ECA acts as a customer service agent, its friendliness can be highly valued; when it plays the role as a teacher, its ability to capture the user's attention and the trustworthiness it demonstrates may be of greater importance.Since the ECA established in this study was a presenter in a museum, we selected animacy, likeability, perceived intelligence,

and focused attention, which were of greater interest to us, as the main evaluation dimensions.

The second question for subjective evaluation through user study is ”how to measure?”. The instrument typically applied for subjective assessment is questionnaires. However, until we conducted this study, there still lacked a standard questionnaire to evaluate ECAs. Under this situation, some researchers have begun developing a unified scale that can cover multiple dimensions [26]. However, the development of the unified questionnaire is still in progress of validation. So we employed two validated questionnaires as alternatives, which are the Godspeed Questionnaire [6] and User Engagement Scale(UES) [70] to evaluate our results. Moreover, specific indicators can be objectively observed as compensation to the subjective measurements. For example, some studies have utilized eye-gazing behavior [17] and body posture [75] to assess user engagement. We also implemented an eye tracker to observe the human's focused attention when interacting with the ECA in this study.

1.2 Motivation

Users in previous experiments evaluating data-driven gesture generation models were typically asked to watch clips of ECAs' behaviours, rate the generated stimuli, or perform pairwise comparisons among ECAs[87]. Humans are positioned as observers in such approaches, rather than as interactors, which is not as natural as the scenarios faced by ECAs in real-world situations. The aim of this study is to evaluate deep-learning-based gesture generation models in interaction for the first time.

1.3 Research Questions and Approach

In our experiments, we applied the ECA into a presentation scenario and involved the participants interacting with the ECA. We implemented two parts of interactions, which correspond to two conditions, respectively:(1) enable the ECA to make data-driven generated gestures(the 'Gesturing' condition) (2) the ECA's body swing slightly to from side to side while breathing (the 'Idle' condition). The experiment participants were required to make some simple voice commands to the ECA, such as to select the topic they would like to listen to. After each interaction, participants needed to rate a 5 point Likert scale according to their experience. Through this experiment, we sought

to investigate the following question.

Can the selected data-driven gesture model improve users' perception of the ECA during real-time interaction?

Our study also aims to discuss whether it is practical to evaluate the ECA performances in real-time interaction and compare the effect of ECA's gesture behavior on user perception.

1.4 Outline

The remainder of this thesis is organized as follows. Chapter 2 will introduce the theoretical background and knowledge underlying the study. Chapter 3 will discuss the related literature. We will describe our experimental methods and procedures in Chapter 4. In Chapter 5, we will analyze the experimental results qualitatively and quantitatively. Chapter 6 will discuss the contributions and limitations of the study.

Chapter 2

Theoretical Background

This Chapter will demonstrate extended theoretical background related to gesture and ECA related study, which provided inspiration for our study. We will first outline the importance of co-speech gestures in social communication, then describe the previous practices of developing and evaluating gestures in ECAs.

2.1 The Importance of Co-speech Gestures in Social Communication

2.1.1 Universality and Classification of Gestures

Nonverbal behaviors play a crucial role in interpersonal interactions. Among the various types of nonverbal behaviors, gestures are the most frequently observed. Gesturing accompanied by speech is prevalent in people of all cultures, genders, and ages. Making gestures while speaking seems to be innate. Even children who are congenitally blind will move their hands while speaking, even if they have never seen another person [31]. The gesturing phenomenon does not disappear even when it cannot be seen by the person with whom one is communicating. It is noticeable that when people talk on the phone, the speaker still gestures even if the person cannot see him or her [7].

Humans tend to use gestures as an extra-verbal channel to aid or enhance communication. For example, when a speaker describes information that the listener does not understand, he or she makes more and larger gestures to help him or her

communicate [45] [39].

Gestures can convey multiple kinds of information. According to McNeill's classification, gestures can be classified into five types. The first type is emblematic gestures, also known as quotable gestures, which refer to direct translations of short verbal communications. For instance, waving the hand can indicate hello or bye. The second type is called affect displays, as the name suggests, and is used to express emotions or intentions, often with some personal, subtle movements. The third type, called regulators, is used to maintain the interaction between the communicating parties. The fourth type is adaptors, usually an unconscious habitual action. The fifth type is called illustrators, which are usually associated with the semantics of speech and can be used to point (deictic), to represent concrete objects or actions (iconic), to translate abstract concepts (metaphoric), or to correspond to rhythm and pause (beat) [4].

2.1.2 How gestures impact the effectiveness of information communication

The impact of gestures on communication has also been extensively studied. Many studies have shown that nonverbal behaviors may convey no less information than words in communication [8].

In many situations, gestures can enhance the intentional information conveyed [40]. For example, when the speaker describes the size of an object, matching the gesture to the description enhances the listener's impression of the object's size [54]. Gestures can also be used as a supplement to speech information. Gestures provide an additional channel when speech is not clearly articulated or when the environment is too noisy for the conversation to be heard. Gestures can also reveal unintentional information; people reveal information through gestures that are not expressed in speech or even reveal cues when lying [21].

Similarly, hand movements can mislead the listener. For example, in one interview, the interviewer made a hat gesture during the conversation, and then when the interviewee was asked, "What color hat did he wear today?" the interviewee would answer. However, the interviewer was not wearing a hat that day [10]. This phenomenon tends to occur when gestures convey different information than speech.

2.1.3 How gestures express emotions and personality

Many studies on Kinesics have shown that human emotional states can be bodily expressed. Some objective studies based on motion capture data have indicated that velocity, acceleration, and jerk are closely related to emotional states. Most of the studies have focused on the overall dynamics of the whole body. It has been found that when a person is angry, he or she moves faster and over a larger spatial area. When a person is fearful, his body contracts into a smaller spatial area and moves more slowly [12]. Some studies focused on the correlation between gestures and emotions. These studies have found that the magnitude of gestural movements is related to speed and underlying emotion [33]. For example, when a person feels joy, there is a high peak in elbow flexion and extension velocity, with the arm extended forward. In contrast, elbow motion amplitude and extension velocity are small and slow when a person feels sad [12].

Studying the mapping relationship between gestures and emotional states would facilitate the optimization of social robot and virtual agent gestures so that the robot or agent can better interact with people in accordance with social guidelines.

2.1.4 Cultural and Gender Variations in Gestural Expressions

Just as people from different cultures have different verbal languages, body language varies across cultures. S. Kita classified the factors that lead to cross-cultural variation in gesture into four categories [47]: The first is the association between form and meaning of culture-specific conventions. This difference is usually found in the expression of emblematic gestures. A typical example is forming a circle with the thumb and index finger and the other three fingers up. In Northern Europe, England, and Southern Italy, this gesture means OK/good; however, this gesture may be very offensive in Brazil, Greece, and Turkey.

The second factor is spatial cognition. Levinson summarized two ways in which people encode spatial relations, one using relative positions to their bodies (e.g., left, right, front, back) and the other using absolute positions in universal space (e.g., east, west, south, north) [34]. Since the preference for these two ways of encoding spatial information differs across cultures, the corresponding gestural motion also varies.

The third factor is linguistic differences. That is, the mapping between lexical and syntactic expressions to gestural expressions varies across cultures. For example, when describing the movement of an object in English, it is customary to put the ”trajectory” of the movement and the ”manner” of the movement in one clause (e.g., It is rolling down); however, when describing it in Japanese, people use two clauses to describe the same movement (e.g., It is getting down. It is getting down. It is rolling. Accordingly, English speakers would make one gesture when gestures are performed, while Japanese speakers would produce two. This difference usually appears in the expression of deictic gestures [48].

The fourth factor is the gestural pragmatics in a particular culture. More specifically, it refers to some of the conventions of using gestures agreed upon in that culture. For example, when Italians gesture, they take up more space, and their whole arm moves. However, when Jews gesture, the body occupies a more compact space [20].

Human communication is multimodal. Hence recognizing the cultural variations in gestures is as critical as understanding differences of various languages.

Gender differences in the use of gestures were also observed. Researchers found that women use more gestures than men to indicate their thoughts and emotions in the same task [52]. J. Liu’s study demonstrated that gender differences in language and gesture exist both internally (in cognitive processing and brain activity) and externally (in the presentation of language and gesture) [64]. Since our study focused more on the transfer of human gestural behavior to ECAs, we concentrated only on the external presentation of gestures. Freeman’s study pointed out that men are accustomed to expressing themselves with obvious gestures in social communication, whereas women tend to use subtle and restrained gestures [27]. The gender differences in human gestural expression imply that when assigning gestural functions to ECAs, designing different gestural strategies according to the gender settings of the ECA may enable the system to achieve better performance. However, care should also be taken to avoid enhancing gender-related stereotypes.

2.1.5 The effect of gestures on the perception and comprehension of information recipients

From the viewpoint of the information recipient, the speaker’s gestures can contribute to the perception and understanding of the information.

A.L. Hubbard et al. used the technology of functional magnetic resonance imaging to scan the brains of subjects exposed to speech. They identified that the addition of beat gestures to a speech could make several brain areas more active [43].

Information recipients' perception of gestures is related to the gesture types. It has been shown that illustrator gestures contribute to the listener's attention, ability to remember content, and comprehension accuracy [9]. Some studies have also suggested that adaptors gestures are associated with perceptions of anxiety, tension [37] and are negatively related to perceived persuasiveness [11][65].

The motion amplitude of the gesture in space also affects the receiver's perception. Research by G. Ferré revealed that listeners perceive more prosodic emphasis when they see larger magnitude gestures than when they see smaller magnitude gestures [24].

Many of the above examples have mentioned that speakers make gestures as if they intend them to enhance the listener's comprehension. However, whether gestures actually contribute to listeners' comprehension is still a question that needs to be studied. Some studies suggest that gestures do not affect listeners' understanding of the content of the speech [46][54]. Other studies have demonstrated that gestures can help listeners understand speech content when they are visible to them [32][83].

Some researchers have proposed a pertinent answer to this paradox: Although gestures do not always enhance the receiver's understanding of the content, they can still benefit the listener under certain conditions. Hostetter's research indicated that in order for gestures to enhance the listener's understanding of the content of the message, three influencing factors need to be considered. The first factor is the topics expressed by the gesture. When spatial and movement information is expressed with gestures, it maximizes the listener's understanding, while when abstract information is expressed with gestures, it contributes least to understanding. The second factor is whether the content of the gesture overlaps with speech. When gestures are used to convey some information that is not contained in speech, they are more helpful in promoting the listener's understanding and vice versa. The third factor is the age of the listener, with children benefiting more from gestures than adults [40].

Clarifying the impact of gestures on information receivers can help researchers better capture the focus when designing gestural features for ECAs.

2.2 ECAs with Gestures

Given the role of gestures in human social interaction, it is not surprising that researchers attempted to empower ECAs with gestural expressions. In this section, we will first provide an overview of the concept of ECAs, followed by examples of their application in real-world scenarios in order to allow the reader to develop an understanding of ECAs; finally, we will present the two main classes of methods that enable ECAs to generate gestures and their respective advantages and disadvantages.

2.2.1 What are ECAs

The term *conversational agent(CA)* refers to a software system capable of communicating with humans through natural language. CA applications have become increasingly popular with the development of natural language processing technology. CAs can be classified into three different types, namely audio-based, text-based, and body-based. Audio-based CAs are typically voice assistants such as Apple's Siri and Amazon's Echo smart speakers. Generally, text-based CAs can be regarded as chatbots, such as automated customer support on e-commerce websites. The body-based CA can also be referred to as *embodied conversational agents (ECAs)*, which have received increasing research attention due to their ability to interact with humans multimodally through nonverbal behaviors.

When discussing embodied conversational agents, a variety of terms are often used interchangeably. However, there are slight differences between these terms. *Virtual characters* often refer to virtual figures in games or animations whose behavior is pre-animated, such as NPCs (Non-Player Characters) in video games. An *avatar* usually refers to the digital representation of a real human being in the physical world, for example, in online games, players can manipulate the action of their avatars through the joystick or mouse to make them walk, run or jump; A *virtual human* is a term more focused on the human-like appearance.

In this study, we use the term *embodied conversational agent (ECA)* to describe a virtual robot-like agent with a body and with a simple conversational system.

2.2.2 Applications of ECAs

Presently, ECAs are broadly applied in a variety of scenarios, such as health management [53], training [62] and customer service [78]. According to the directions of interaction between ECA and users, we can classify the application scenarios into the following three categories. 1. one-way interaction; 2. two-way interaction; 3. multi-party interaction. Below, we will introduce some application examples in each of these three categories.

1. One-way Interaction: Agents with one-way interaction capabilities are capable of generating appropriate expressions and actions based on the content of the communication, but are not capable of receiving input from the user.

   For example, a virtual customer service agent in front of the reception desk providing information and guidance can be considered an one-way interaction [87].

2. Two-way Interaction: Two-way interaction is the most common manner that ECAs interact with users. This type of agent can not only output text, speech and non-verbal behavior, but also receive multi-modal inputs from the user and then provide feedback on the user's input.

   For instance, when an intelligent health agent intervening in an user's drinking habits based on her/his facial expressions can be refer to as a two-way interaction [63].

   The other instance is Anna, IKEA's virtual agent,which can help users to find appropriate products based on their text input. Anna can also make facial expressions related to the output [79].

3. Multi-party Interaction: Agents with multi-party interaction capabilities receive signals not only from a single user input, but also act on inputs from multiple users or other agents.

   For example, D.Traum and J.Rickel presented an immersive virtual worlds where multi-party conversation can be taken place among different characters and users.The system can be used for military training [82].

Among the three categories of interaction scenarios, two-way interaction is applied most extensively. The interaction designed in this study also belongs to two-way

interaction.

2.2.3 Methods to Generate Gestures for ECAs

As robots and virtual agents have become more prevalent in real-world applications, many HRI studies have discussed how to enable the nonverbal behavior of robots/agents to conform better to social communication patterns.The major question to enable an ECA to generate co-speech gestures is to align gesture animation to speech content [71]. Many researchers have dedicated their efforts to Speech-to-Gesture mapping methods to enable their agents better mimic natural communication with humans [13][56][88][2].

The mapping methods can be categorized into two groups: the rule-based methods that requires pre-animation of the agent's body movements; and the data-driven methods that can generate gestures from pre-trained machine learning models.

Rule-based methods require human work to associate speech to gesture animation. J. Cassell et al. [13] used a gesture dictionary containing specific motions to represent related semantic contents. S.Kopp et al. [51] proposed an XML based language called Behavior Markup Language(BML).Human body parts are assigned to different BML elements such as <head>,<torso> etc. that contain attributes demonstrating possible motions.

Rule-based methods have notable limitations: first, they require significant labor costs and time to research and design rules; second, they are more unlikely to produce natural gestures when some rules cannot be clearly defined. Data-driven methods greatly compensate for these limitations.

For the data-driven approach, it is crucial to choose the types of training data. In previous studies, inputs are usually in single modality, either audio or text. In the case of *audio-driven approach*, Hasegawa et al. [36] adopted Deep Neural Networks(DNNs) to synthesize gestures from speech audio. Kucherenko et al. [56] added representation learning onto the model to simplify the learning tasks. Levine et al. proposed Gesture Controllers, a method that employs reinforcement learning methods to construct an optimal policy, thus enabling the mapping of acoustic features in speech to motion animation [61]. Ginosar et al. [30] translated raw audio of speech to smooth and plausible hand gesture motion in a GAN-CNN system. Alexanderson

et al. [2] proposed a probabilistic model to synthesize style-controllable gestures with audio input, which based on normalizing flows with autoregression(MoGlow) [38]. As far as *text-driven* approaches, Neff et al. used speech text as a feature and speech themes, rhymes, etc. obtained from video recordings as tags to empower the model to produce gestures of given styles [69]. Yoon et al. [88] presented an end-to-end model containing an encoder and decoder, to generate gestures from speech text.

Researchers also attempted to combine text- and audio-driven approaches by employing both input formats together. Kucherenko et al. [58] suggested *Gesticulator*, a deep-learning based model to generate co-speech gestures from both semantic and acoustic input. Thangthai et al. [81] built a model on an encoder-decoder bidirectional LSTM neural network architecture, which generate gestures from text and audio features.

Generally, both rule-based and data-driven approaches have advantages and limitations. For example, rule-based approaches may be more precise in representing specific personalities and emotions [1]. In situations where gestures complement speech content, for instance with deictic, iconic or linguistically-specific metaphoric gestures, rule-based methods can be more efficient in automating gestures, as illustrated in the research conducted by Ravenet et al. [71]. Purely based on data, it is difficult to create an accurate gesture representation. However, rule-based approaches do require expert knowledge to design the model and are less flexible to implement in situations that are not well defined. By using data-driven methods, smoother gestures are achieved without the need for a deep understanding of the mechanism of gesticulation from a cognitive viewpoint. The main limitation is the database that can be used to train the model [57].

Chapter 3

Related Work

This chapter will first briefly describe how previous studies have evaluated ECA with gestures, both subjectively and objectively. Next, we will discuss effective approaches to user studies, including commonly used experimental settings, measurement tools, and evaluation dimensions, from which we will be able to take lessons for our own evaluation.

3.1 Evaluation for ECAs with Gestures

According to the theoretical background chapter, gestures might increase the listener's comprehension and perception by introducing an additional modality in some circumstances. Do gestures, therefore, enhance human interaction with ECAs? Furthermore, how about adding gesture generation functions to ECAs? How good are the generated gestures? On what dimensions does gesture behavior affect human perception of ECAs? To answer these questions, we must evaluate ECAs systems. As we evaluate ECA systems, two questions need to be addressed: "How to evaluate?" referring to the evaluation method, and "What to evaluate?" which involves different evaluation dimensions and indicators.

From our review of the studies, it appears that most researchers use both objective and subjective methods to evaluate ECAs. On the one hand, the researchers evaluate the performance of data-driven models objectively or compare the results of multiple models. 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. The following sections present some case studies to illustrate their evaluations.

Huang et al. conducted a subjective experiment to compare the human perception of the robot under four modal conditions (1. gestures generated by data-driven model; 2. no gesture; 3. gestures randomly assigned; 4. gestures generated from rules) [41]. The study results indicate that participants experienced 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.

According to study by Salem et al. , participants felt a greater degree of human-likeness, likability, and shared reality when co-speaking with the robot, even when the gestures were inconsistent with the content intent [74][73]. According to the same group of researchers, incongruent gestures implemented by robots may adversely affect task-related performance. The perception of gestures may also differ by gender. Interestingly, the researchers found that male and female respondents viewed the robots differently when they performed the four categories of gestures described by McNeill (deictic, beat, metaphoric, and iconic) [42].

Yoon et al. implemented both objective and subjective evaluations. To objectively evaluate the quality of gesture generation, the researchers compared the generated gestures with the ground truth in the original video. Furthermore, researchers in this group also undertook a user study to obtain a subjective assessment of the proposed model. Anthropomorphism and Likeability were chosen from the Godspeed questionnaires, plus a self-defined Speech-Gesture Correlation dimension. Researchers compared the proposed model to a neural network-generated model, a manually defined model, and a random model using the selected instruments. Results show that the proposed model outperforms other models on the chosen dimensions [88].

Levine et al. undertook a user study to evaluate the proposed HMM-based model. Participants were asked to perform side-by-side pairwise comparisons to determine which model produced more realistic gestures. The results of the user tests were similar to the results of the objective evaluation - the gestures generated by the proposed models were found more realistic [61].

Ishi et al. evaluated the quality of models by comparing generated animations with

human action videos and adjusting the model based on the variation of mean squared error. Additionally, the researchers conducted subjective experiments on 20 subjects. Four different conditions were compared: no action, directly mapped human action, text-based generated gesture action, and text- and prosody-based gesture action. Human-likeness, Speech suitability, Gesture Naturalness, Frequency, and Timing of Gestures have been selected as indicators. The researchers customized a 7-point Likert scale question for the study. The results indicated that the two proposed methods were more natural than the other approaches [44].

In 2020, a group of researchers launched the first gesture generation challenge [59]. This event set a benchmark for the practitioners in the gesture generation domain to compare their models by using the common dataset, common visualization setup, and common evaluation procedures. In the challenge, both objective and subjective methods were used, but the main emphasis was on subjective evaluation.

According to our literature review, user studies are essential for the successful evaluation of ECAs. In the following section, we will discuss in more detail the approaches related to user studies.

3.2 Approaches to Conduct User Studies on ECAs with Gesture Functions

3.2.1 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 [73]. 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 [3].

Nevertheless, in our literature reviews, we found that studies of real-time interaction

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 [60]. GENEA 2020 organizers provided participants with a user interface for measuring performance. This allowed them to rate their performance after viewing a video of the ECA, however, the experiment did not involve human interaction [59]. 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 makes it possible to experiment with different gesture generation models on a virtual conversational agent [67].

3.2.2 Experimental Setups

The purpose of user studies is to measure performance under different experimental conditions to assess the design's effectiveness. A typical experimental design faces two major challenges once the conditions have been established: how to assign participants to different conditions, and how to collect data.

Typically, researchers use within-subject or between-subject designs to assign participants to experimental conditions. Wolfert et al. reviewed 22 studies that included user experiments. They have found that a majority of studies (15 of 22) conducted within-subject experiments and fewer (7 of 22) employed between-subject experiments [87]. In comparison with between-subject experiments, within-group experiments require fewer participants, and their internal validity is not affected by random assignment. Within-subject experiments should be conducted with the awareness that the order of exposure may affect perceptions. The "demand effect" could also be a pitfall, where participants may guess the tester's intention and purposefully evaluate the item in a way that pleases the tester [15].

For data collection, researchers often employ direct scoring or pairwise comparisons. Generally, direct scoring requires participants to complete a questionnaire after being

exposed to an experimental condition in order to rate it. For example, participants would be asked to rate the degree of human-likeness of ECA on a scale ranging from 1 to 5, from non-human-like to human-like. The pairwise comparison method allows participants to compare two or more conditions at the same time. The participant is then asked to select the experimental condition that is closer to the criteria provided by the tester. For example, participants would be asked whether ECA1 or ECA2 is more human-like. Wolfert et al. compared these two methods using a data-driven gesture automation method as a test-bed [86]. The results obtained using the two methods did not differ significantly in terms of validity. The paired comparison approach was considered faster and provided greater reliability as compared to direct scoring. In the opinion of the researcher, the paired comparison is more effective when fewer conditions are present. On the other hand, the direct score method allows researchers not only to rank conditions but also to compare different dimensions within a single condition, thus providing a more refined assessment [86].

3.2.3 Measurements to Evaluate Gesturing ECAs

Gestures have a variety of effects on humans, including cognitive, emotional, behavioural, as well as on their performance of tasks [76]. However, to our knowledge, there is still no unified questionnaire that can directly evaluate the quality of gestures in ECAs. Researchers usually select the features for evaluation based on their own research interests. The existing questionnaires serve as indirect measures of the role gestures play within the system by assessing people's perception of ECAs as a whole [26]. Nevertheless, since human communication is inherently multimodal, indirect measures are still valid and relevant. Below are some typical cases to illustrate how evaluation dimensions and tools are used in existing studies.

Aly and Tagus developed a model for generating gestures based on mapping human personality traits to nonverbal behaviours. The model was applied to the NAO robot. An autonomously developed questionnaire including 24 questions on a 7-point Likert scale was used to evaluate the results. The survey covered user preferences, robot personality congruence, user engagement, robot expressiveness, robot gesture, voice synchronization, etc. In this study, testers used the Big 5 Inventory Test to distinguish between introverted and extroverted participants and then conducted two rounds of experiments. The objective of the first experiment was to check whether people

were able to accurately identify a robot that matched their personality through verbal and non-verbal cues. In this way, each subject was exposed to introverted and extroverted robots. The second experiment was designed to test whether gestures improved expressiveness. Participants would interact only with the robots matching their personalities, one robot communicating unimodally (by spoken words only) while another communicating bimodally (by spoken words and gestures) [3].

Salem et al. investigated the impact of robot gestures on humans' perceptions of anthropomorphism and likability. Researchers employed two sets of questionnaires to assess participants' perception of the robot and their performance on task-related tasks in the experiment. Perception-related questions included Humanlikeness, Likeability, Shared Reality [19], and Future Contact Intentions. The researchers customised the questionnaire and performed reliability testing. To assess task performance, the authors used both objective and subjective ratings. The subjective indicator was a 5-point Likert-scale question asking participants to rate their capacity to solve the challenge. The objective metric was the error rate of task completion [73].

Due to the lack of a standard evaluation scale, many studies have used self-developed questionnaires to assess gesture generation. Autonomously designed questionnaires, however, might not have guaranteed internal reliability. Additionally, constantly creating new measurement tools rather than reusing well-tested instruments makes it harder to replicate and compare experiments. S. 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 [25]. Because of the need for a unified measurement tool, Fitrianie et al. analyzed 189 constructs drawn from 89 questionnaires adopted in previous researches and involved many domain experts in categorizing these constructs. As a result of integrating constructs with overlapping content, 19 different sets of dimensions were derived, which served as the basis for the development of a unified scale in the future [26]. Since this scale was still in development at the time of our study, we regret that it could not be used as our measuring tool.

3.2.4 Evaluation Instruments for ECAs

Human perception and cognition are difficult to measure, especially for researchers unfamiliar with the psychology discipline. Engineers often develop questionnaires

that are not properly validated. To facilitate researchers in the HCI/HRI field to evaluate their systems, C.Bartneck et al. developed a set of standardized measurement instruments known as the Godspeed Questionnaire. The questionnaire contains five dimensions: Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety [6]. Correlations between the dimensions were examined. In this questionnaire, questions were asked using 5-point semantic differential scales, which are similar to Likert scales but can better avoid acquiescence bias [28].

Godspeed Questionnaire has been widely used in social robot studies. Many studies used the entire Godspeed Questionnaire set. For instance, A. Deshmukh et al. analyzed how Godspeed scores varied with gesture amplitude and speed in an experiment involving 30 observers [18]. J.R.C. Ham used the whole Questionnaire to measure different combinations of gestures and gaze behaviors [35]. Some studies have only used specific dimensions of the Godspeed Questionnaire. For example, A. Korgesager employed only the Perceived Intelligence dimension to examine the effect of head nodding when the robot possessed it [55]. In our srtudy, we selected Animacy, Likeability, and Perceived Intelligence from the Godspeed questionnaire based on our research interests.

Besides the dimensions covered by the Godspeed Questionnaire, User Engagement is also a common concern in ECA systems. There is widespread agreement among researchers that user engagement can be classified into three categories: emotional, cognitive, and behavioral. O'Brien developed the User Engagement Scale (UES) by identifying six dimensions of engagement, namely Focused Attention, Perceived Usability, Aesthetic Appeal, Endurability, Novelty, and Felt Involvement. UES consists of 31 questions rated on a Likert scale of 1 to 5 [70]. Researchers rarely use all dimensions simultaneously to evaluate a proposed system but instead utilize certain sub-dimensions. Due to our study's focus on how an agent's gestures impact humans' attention, we only assessed the Focused Attention dimension on the UES.

3.2.5 Evaluation of ECAs through Behavior Observation

Researchers consider behavior observation to be more rigorous than indirect surveys due to its objectivity. Researchers can complement questionnaires with behavioral observation methods. T. Kooijmans et al. developed a software called Interaction Debugger that collects objective data in human-robot interactions. This system can

capture a variety of data, including sound, vision, person identification, motion, body contact, and robotic behavior [50]. Some other commonly utilized behavioral indicators include proxemics, postures [75], eye-gazes, etc. J.N. Bailenson measured the social distance between a person and a virtual agent to determine copresence in a virtual environment [5]. Y.I. Nakano et al. used eye-gaze behavior as an objective indicator of a person's attention when talking to a virtual agent [68].

Chapter 4

Study Method

4.1 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 6 classical Roman monuments according to the participant's voice command. The video and source code of the study can be found on the author's website[1].

4.1.1 System Architecture

The system was modified and simplified based on the modular framework developed by Nagy et al. [67]. The modified system architecture can be transplanted to any situation that involves a presenter, such as a lecturer or a museum guide. The system architecture consists of 4 modules (see figure 4.1.1). Module 1 is the gesture generation model known as *Gesticulator* developed by Kucherenko et al. [58]. *Gesticulator* is a data-driven method that can automate hand gestures based on both prosody and text. The current gesture generation methods can also be replaced by other data-driven models that output gesture coordinations by frame. Module 2 contains the speech synthesis model developed by Székely et al. [80], which is used to generate the agent's audio output. Module 3 contains the database of presentation content (see figure 4.1.2). The author developed this database in the Unity3D engine using C#. The database allows system users to customize the presentation's content by providing corresponding command words, pictures, audio, and speech text [80]. Module 4

[1] https://www.yaeh.io/research/hci/presentingbot

provides the virtual scene that can be replaced by any customized 3D model. In this scene, the Gesticulator can communicate with the virtual robot through the ActiveMQ message broker.

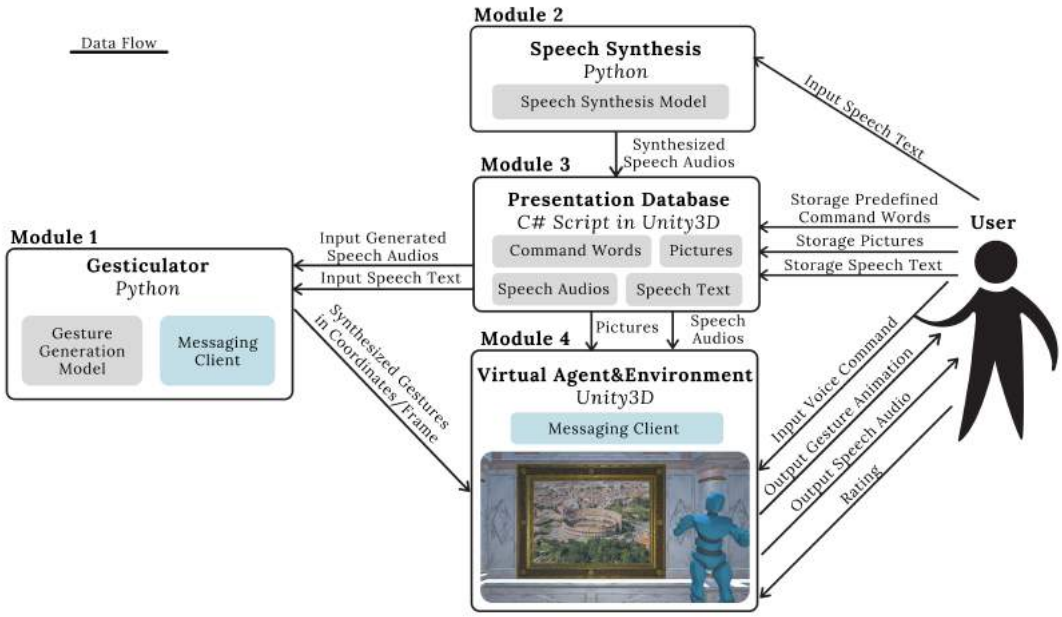

Figure 4.1.1: The system architecture.

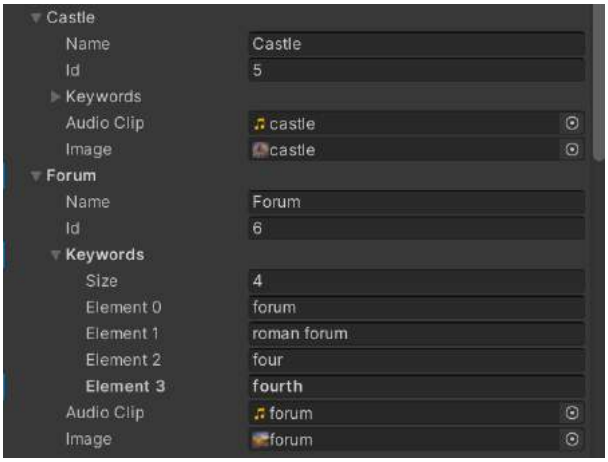

Figure 4.1.2: The user interface of database.

4.1.2 Database Structure

In this experiment, the topic of our presentation was ”Roman Classical Monumental Architecture.” The database includes images of six monuments, audio clips, and the presentation text. The gesture files were saved in CSV format as coordinates for

every frame. Because of the unpredictability of speech recognition, we simplified the speech commands as much as possible and implemented them using the Microsoft speech recognition API included with Unity3D. We defined two forms of speech input, one by speaking keywords and the other by saying which number corresponds to a picture.

4.1.3 Gesture Generation

First, we wrote the presentation contents and saved them as text files. Next, we synthesized speech from the text using the speech synthesizer [80]. The synthesizer was selected because it shared the same training dataset with our chosen gesture generator - *Gesticulator* [58]. We then decided to generate gestures in advance rather than in real-time in order to minimize the latency of the system in response to a voice command. We input both text and audio files into Gesticulator, and Gesticulator exported the generated gestures saved in the format of coordinates/frame in CSV files. Finally, we imported text, audio, image, and gesture files into the database. Thus, after receiving a voice command during the experiment, the system would present the referred audio clip and image, as well as play related gesture animation.

4.2 Participants

A total of 28 participants (10 female, 18 male) were recruited by online and offline posters published in the KTH community (Facebook forums, Whatsapp groups, Telegram groups, advertisement boards and building doors on campus). After completing the experiment, participants were rewarded with a voucher worth 70 kr from a local supermarket (ICA).

Participants were diverse in age (18–36, Mean=24.39, SD=13.71) and nationality. To ensure the quality of the experiment procedure, all participants had to be very competent in both listening and speaking English.

The experiment was carried out using a within-subject approach, where each subject was exposed to two experimental conditions.

4.3 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 by 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 presented the related Roman monuments followed 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 only hang naturally without gestures, accompanied by natural breathing and body microdynamics when presenting.

Due to the limited number of participants in our study, we selected a within-subjects experimental design. Instead of comparing different models, this study will examine the feasibility and approaches to evaluating data-driven gesture generation models in real-time.

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. 14 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 colours 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 colour preference, we counterbalanced the two colours and the two experimental conditions.

We assumed that ECAs' hand gestures could improve human perception 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, human-likeness, animacy, and perceived intelligence than the ECA standing with idle posture.

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

4.4 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.

We selected three dimensions from the Godspeed questionnaire [6] and then customized an item regarding Human-likeness and asked participants to rate the interaction task subjectively in each condition (see table 4.4.1). Both subjective and objective methods were adopted to evaluate hypothesis 2. Subjective assessment was measured using seven entries in the Focused Attention dimension of User Engagement Scale (UES) [70] (see Table 4.4.2); Objective assessment was inferred from the gaze data collected by the eye tracker. It is believed that when the gaze duration is relatively longer, more focused attention is attracted.

Table 4.4.1: Questions taken from Godspeed Questionnaire

Dimension	**Item**
Human-likeness	How human-like does the gesture motion appear? 1 2 3 4 5
Animacy	Please rate your impression of the agent on these scales: Dead 1 2 3 4 5 Alive Stagnant 1 2 3 4 5 Lively Mechanical 1 2 3 4 5 Organic Artificial 1 2 3 4 5 Lifelike Inert 1 2 3 4 5 Interactive Apathetic 1 2 3 4 5 Responsive
Likeability	Please rate your impression of the agent on these scales: Dislike 1 2 3 4 5 Like Unfriendly 1 2 3 4 5 Friendly Unkind 1 2 3 4 5 Kind Unpleasant 1 2 3 4 5 Pleasant Awful 1 2 3 4 5 Nice
Perceived Intelligence	Please rate your impression of the agent on these scales: Incompetent 1 2 3 4 5 Competent Ignorant 1 2 3 4 5 Knowledgeable Irresponsible 1 2 3 4 5 Responsible Unintelligent 1 2 3 4 5 Intelligent Foolish 1 2 3 4 5 Sensible

Table 4.4.2: Questions taken from User Engagement Scale

Dimension	Item
Focused Attention	I lost myself in this experience. Strongly Disagree 1 2 3 4 5 Strongly Agree
	I was so involved in this experience that I lost track of time. Strongly Disagree 1 2 3 4 5 Strongly Agree
	I blocked out things around me when I was interacting with the agent. Strongly Disagree 1 2 3 4 5 Strongly Agree
	When I was interacting with the agent, I lost track of the world around me Strongly Disagree 1 2 3 4 5 Strongly Agree
	The time I spent interacting with the agent just slipped away. Strongly Disagree 1 2 3 4 5 Strongly Agree
	I was absorbed in this experience. Strongly Disagree 1 2 3 4 5 Strongly Agree
	During this experience I let myself go. Strongly Disagree 1 2 3 4 5 Strongly Agree

4.5 Experiment Procedure

Participants were provided with a consent form before the study explaining the purpose and type of data that would be collected. After leading the participant into the experimental area, the experimenter instructed them to put on the headphones, adjust the microphone, and sit in the proper position. The experimenter then started the experiment system. Participants would see a general instruction on the screen, and the experimenter would explain the experiment briefly while the participant reads the detailed instructions. The experimenter would leave the experimental area during the experiment to ensure distraction-free conditions for the participant. Afterwards, the participant should follow the system's automated instructions without any intervention from the experimenter. From the start of the system, the user's eye gaze data was continuously collected by the Tobii eye tracker at the bottom of the screen.

Participants were exposed to two experimental conditions in succession. Firstly, a selection menu guided the participant to select a keyword of interest. The participant

would need to 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. Upon finishing the presentation, the interface jumped back to the selection menu, and the participant was required to choose another topic. Each condition consisted of two selection interactions. After completing each experimental condition, the system automatically enters the questionnaire mode, and the participant should give a rating according to their interaction experience.

Following the completion of both experimental conditions, the system entered a farewell page, prompting the user to end the interaction and notifying the experimenter. After the experiment, the experimenter conducted a brief interview with the participant. The following questions were asked:

1. Have you noticed the difference between two robots?

2. Which one do you prefer and why?

Finally, the experimenter gave a short debriefing to the participant to explain the intention of the study.

Chapter 5

Result

This chapter will first demonstrate how we process the data retrieved from the user study. Afterwards, we will present the results of statistical analysis and qualitative interviews. Finally, we will discuss the results obtained concerning the proposed hypothesis and compare our results with those obtained by other researchers.

5.1 Data Preparation

5.1.1 Questionnaire Data

A questionnaire of 5 dimensions was sent to 28 participants after each experiment. All questions in this questionnaire contained items based on a 5-point Likert scale or semantic scale, ranging from 1 to 5, representing the weakest to the strongest attribute measured by this indicator. The Animacy subscale consisted of 6 items ($\alpha = .90$), the Likeability subscale consisted of 5 items ($\alpha = .91$), the Perceived Intelligence subscale consisted of 5 items ($\alpha = .85$), and the Focused Attention subscale consisted of 7 items ($\alpha = .88$). The measurements in all dimensions have good internal consistency.

We then used a t-test to rule out whether the ordering introduces bias to the data. Finally, we performed the Shapiro-Wilk normality test on the mean values of each dimension separately. 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 the non-parametric test.

5.1.2 Eye-gazing Data

We received the original gaze data in the format of x, y coordinates of each frame. First, we removed the data corresponding to when the system enters the questionnaire mode and consider only the data for the visual interaction between the subject and the virtual robot. In this way, each subject corresponds to two coordinate data sequences in two sets of experimental conditions. Afterward, we generate a heatmap (see figure 5.1.2) of the two sets of coordinate data corresponding to each subject separately to establish an overall understanding of their gaze behavior. Then, we divided the coordinate area corresponding to the screen into two parts, the presented object (painting) and the presenter (robot) (see figure 5.1.1). We filter out the coordinates of the gaze points falling in these two areas and calculate the proportion of the corresponding time to the total time for subsequent statistical calculations.

Figure 5.1.1: The area division of presenter and presented object

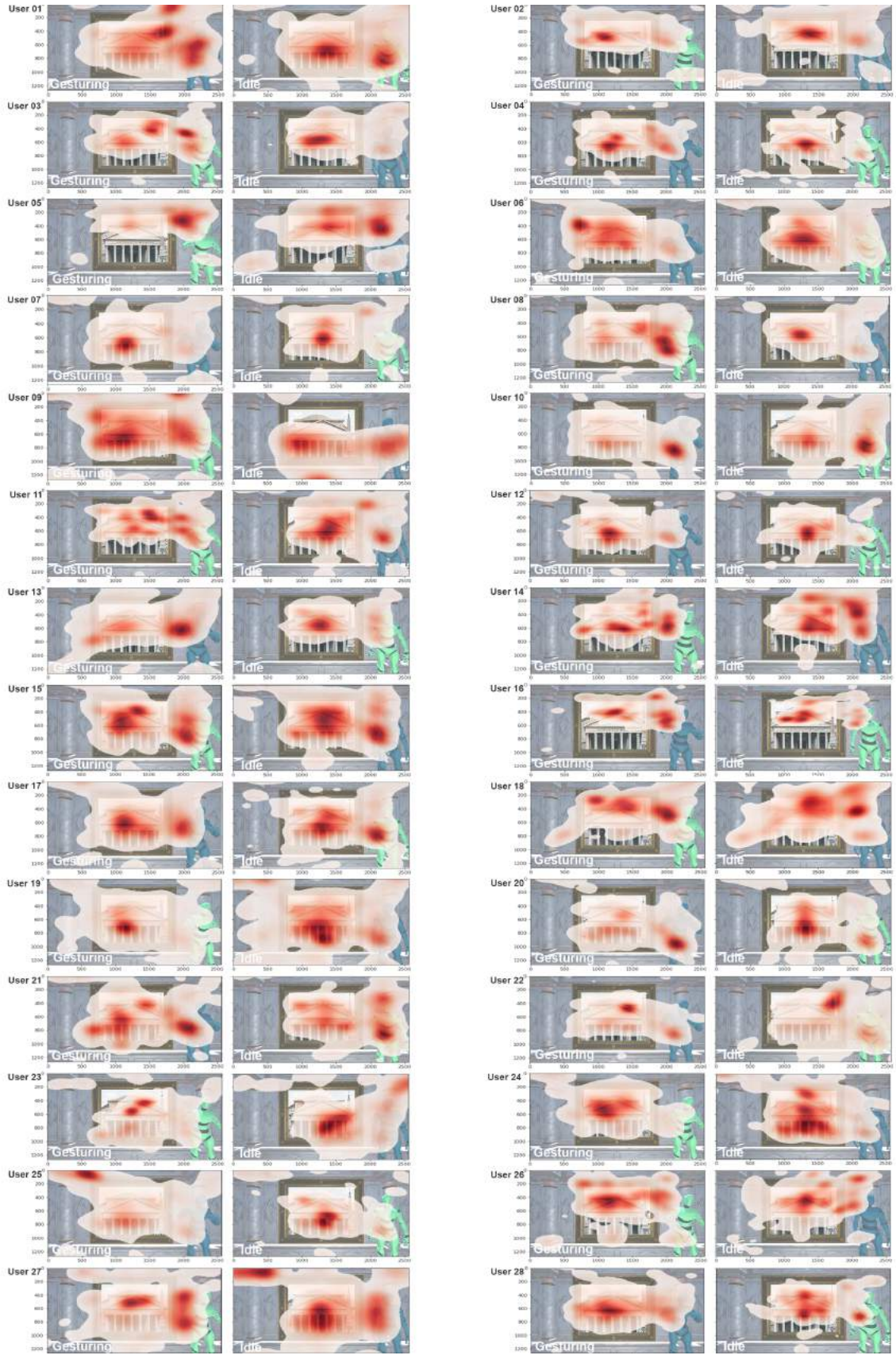

Figure 5.1.2: The eye-gazing heatmap generated based on each participant's focused time. The left-side image in each group is under condition 'Gesturing' and the right-side image is under condition 'Idle'

5.2 Analytical Result

5.2.1 Questionnaire Results

A paired samples t-test was performed to compare animacy in 'Gesturing' and 'Idle' conditions. There was no significant difference in animacy between 'Gesturing'(M = 3.292, SD = 0.838) and 'Idle' (M = 2.982, SD = 0.976); t(df) = 1.633, p = 0.309. The figure 5.2.1 displays the distribution of results in Animacy dimension.

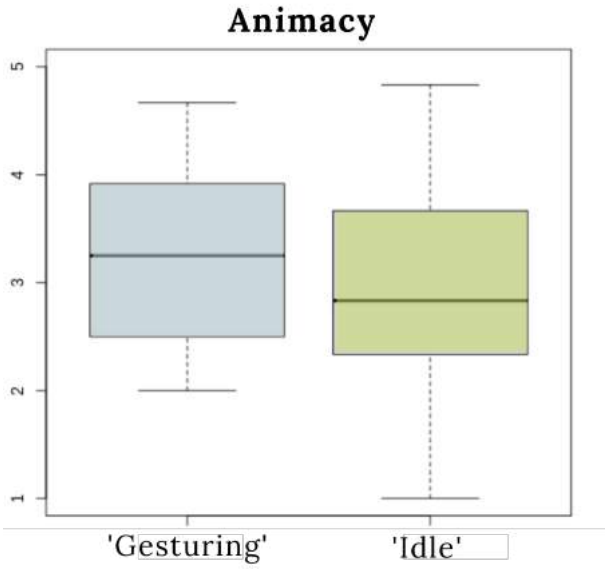

Figure 5.2.1: The boxplots of Animacy

A paired samples t-test was performed to compare focused attention in 'Gesturing' and 'Idle' conditions. There was no significant difference in focused attention between 'Gesturing'(M = 3.010, SD = 0.703) and 'Idle' (M = 2.954, SD = 0.785); t(df) = 0.335, p = 0.740. The figure 5.2.2 displays the distribution of results in Focused Attention dimension.

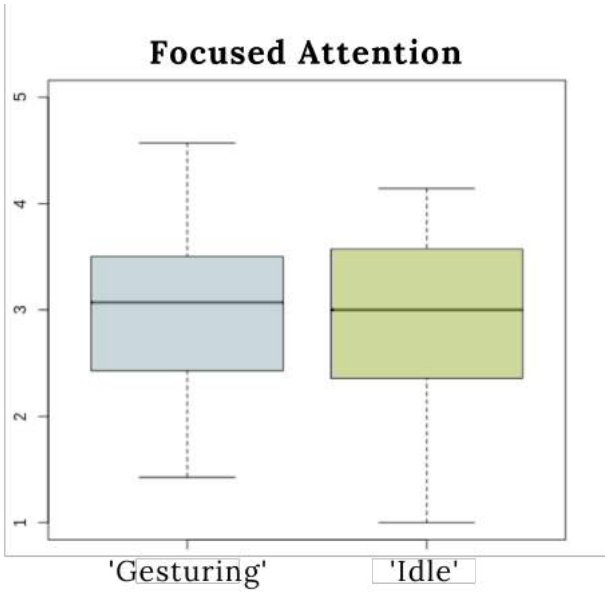

Figure 5.2.2: The boxplots of Focused Attention

An exact Wilcoxon-Pratt Signed-Rank Test was performed to compare likeability in 'Gesturing' and 'Idle' conditions. There was no significant difference in likeability between 'Gesturing'(M = 3.714, SD = 0.863) and 'Idle' (M = 3.714, SD = 0.908); Z = -0.068, p = 0.951. The figure 5.2.3 displays the distribution of results in Likeability dimension.

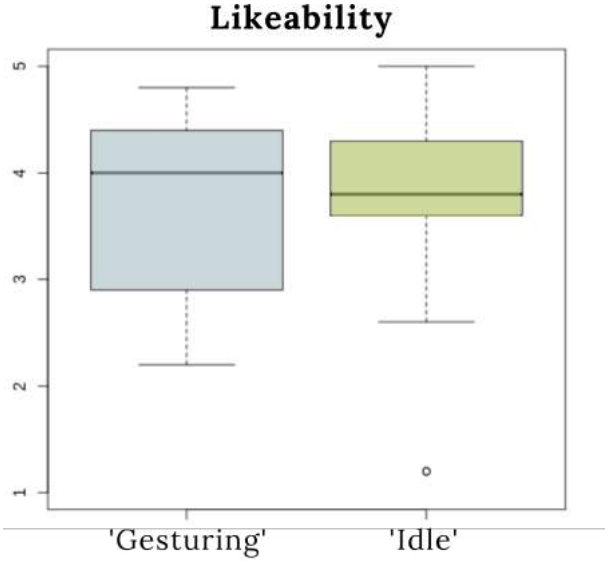

Figure 5.2.3: The boxplots of Likeability

An exact Wilcoxon-Pratt Signed-Rank Test was performed to compare perceived intelligence in 'Gesturing' and 'Idle' conditions.There was no significant difference in

perceived intelligence between 'Gesturing'(M = 3.879, SD = 0.715) and 'Idle' (M = 3.900, SD = 0.803); Z = -0.610, p = 0.551. The figure 5.2.4 displays the distribution of results in Perceived Intelligence dimension.

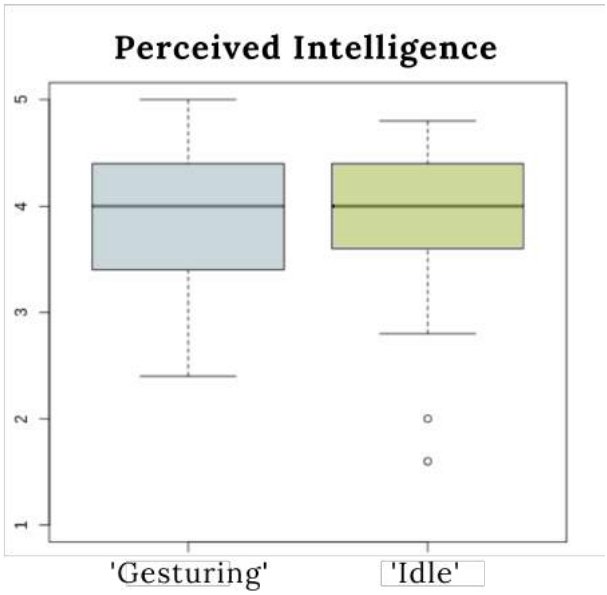

Figure 5.2.4: The boxplots of Perceived Intelligence

An exact Wilcoxon-Pratt Signed-Rank Test was performed to compare human-likeness in 'Gesturing' and 'Idle' conditions.There was no significant difference in human-likeness between 'Gesturing'(M = 3.286, SD = 1.150) and 'Idle' (M = 2.893, SD = 1.100); Z = 1.502, p = 0.146. The figure 5.2.5 displays the distribution of results in Human-likeness dimension.

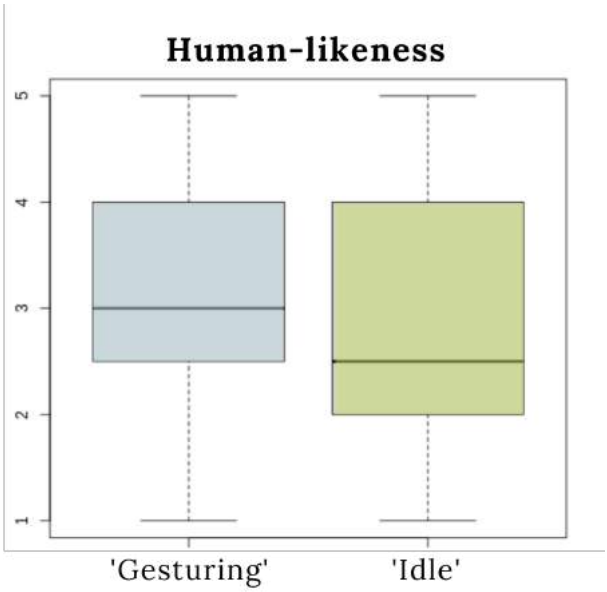

Figure 5.2.5: The boxplots of Human-likeness

5.2.2 Eye-gazing Analysis Results

The second part analyzes the objective evaluation results based on eye gazing data. After dividing the screen into two parts(see figure 5.1.1): presented object (painting) and presenter (robot), we obtain the percentage of gaze time corresponding to the two areas to the total interaction time and we subsequently express this proportion in terms of gazing time(see figure 5.2.6 and figure 5.2.7).

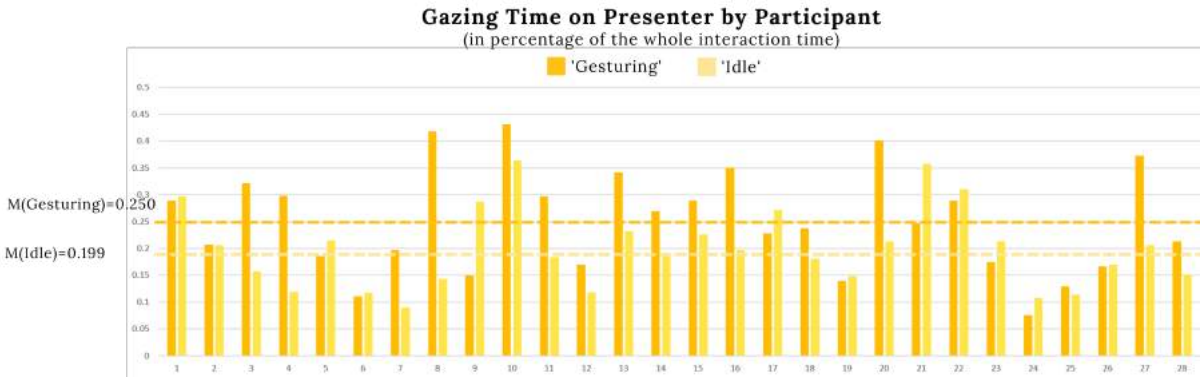

Figure 5.2.6: The eye-gazing time on presenter in percentage by each participant and average gazing time under both conditions.

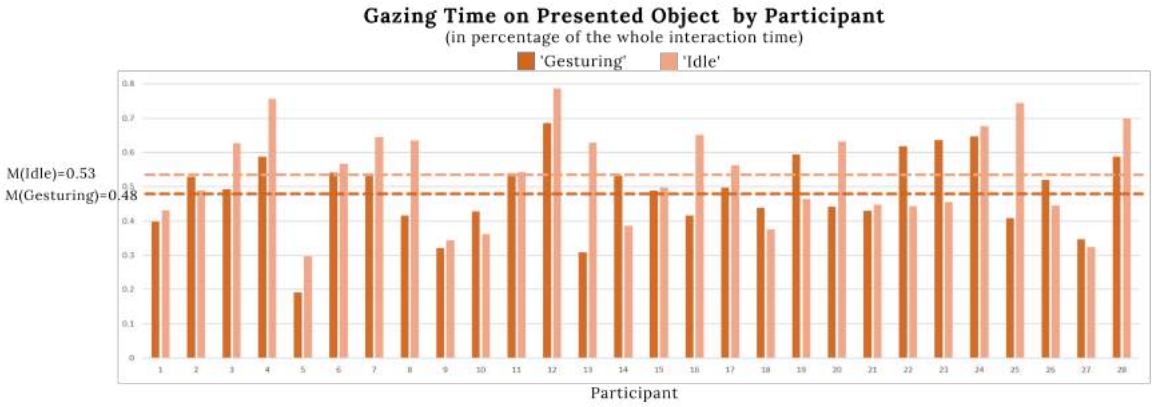

Figure 5.2.7: The eye-gazing time on presented object in percentage by each participant and average gazing time under both conditions.

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. The figure 5.2.8 displays the distribution of results of eye gazing time on virtual agent(presenter).

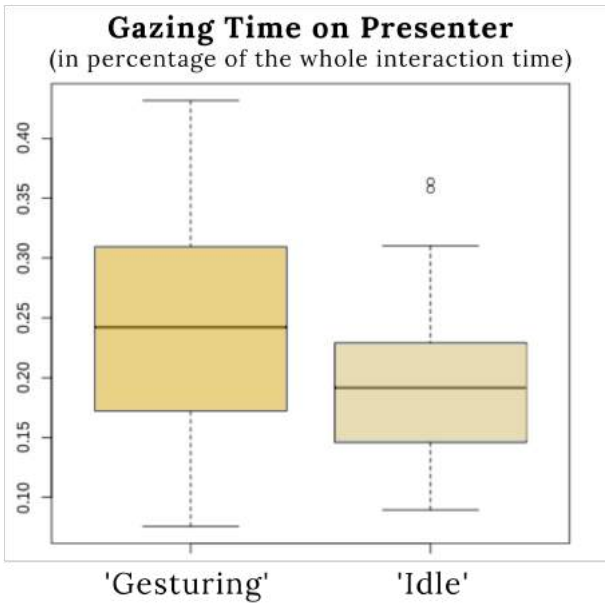

Figure 5.2.8: The boxplots of eye gazing time on presenter.

A paired samples t-test was performed to compare eye gazing time on painting (presented object) in 'Gesturing' and 'Idle' conditions. There was no significant difference in eye gazing time on 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. The figure 5.2.9 displays the distribution of results of eye gazing time on painting(presented object).

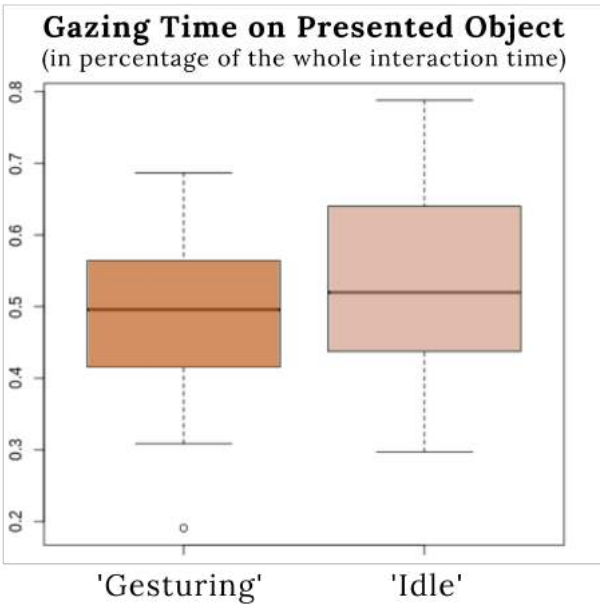

Figure 5.2.9: The boxplots of eye gazing time on presented object.

5.3 Qualitative Result

After each participant finished the experiment, the experimenter would ask them two questions. 1. Have you noticed the difference between two virtual agents? What are the differences? 2. Which virtual agent do you prefer and why?

Regarding the first question, out of 28 participants, 24 mentioned colour differences, 17 thought the two virtual agents' voice (pitch and tone) was different, and 16 came forward with the difference in hand movements.

Regarding the second question, 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. 9 participants preferred the virtual agent without gestures because it looked calmer and less distracting. 6 participants reported no preference or none at all. 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.

5.4 Discussion of the Results

Likeability 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. The result is consistent with the opinions received from the after-experiment interviews.

Other researchers, however, came up with different results. 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 [73]. Huang and Mutlu found that robots with gestures were more likeable than those without [41].

By combining the participants' feedback with our own, we speculate that the following reasons may affect the participants' preferences. First, the model we selected produced too many gestures that did not pause appropriately. In the presentation

scenario, the presenters' excessive body movements may distract people from focusing on the presentation's content. Second, the dialogue system we adopted was not sufficiently intelligent, and unexpected delays and lags could impact the participants' likeability.

Human-likeness We found no significant difference between the two conditions.

Nevertheless, Salem et al. [73] and Ishi et al. [44] reported that a robot that could make co-speech gestures was considered more human-like than a standstill robot.

Possibly, this is caused by the fact that we included only one question related to *human-likeness* dimension, whereas in other studies, there were multiple items contributing to the *human-likeness* dimension.

Animacy and Perceived Intelligence In the analysis, the virtual agent with gestures received higher mean scores in animacy and perceived intelligence. However, these differences did not reach statistical significance, which might be due to a low number of participants.

Focused Attention According to the questionnaire results, neither condition received more attention than the other. Nevertheless, eye gazing data analysis showed a significant difference. In the experiment, a virtual agent acted as a presenter, and some images served as the presented objects. As the eye gazing results indicate, the gesturing agent could attract more attention to its body. On the other hand, the idle agent enabled humans to focus on the object being presented.

Chapter 6

Conclusion

Even though the questionnaire results did not indicate a significant difference in all the dimensions, the analysis of eye gazing data has proved 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 will require students to pay more attention to the slides so that gestures will 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 in order to attract the player's attention and make the character more expressive.

The results of our study suggest that conducting user studies with behavioural observation methods, such as eye-tracking, could be beneficial in obtaining different results. In addition to subjective questionnaires, objective data could provide valuable insight. It is pertinent to discuss and validate whether objective behavioural data can support measurement dimensions. Our study used gaze time to infer focused attention - longer gaze times are considered to indicate greater focus [29]. It is always important to consider the theoretical connection between objective data and evaluation metrics.

6.1 Contributions of this Study

6.1.1 Proved Feasibility of Evaluating ECA in Real-time Interaction

Our study provided an example of evaluating ECA systems 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 would simulate real-world communication with humans. Participants could notice more details when they interact with agents. For example, the way a virtual agent reacts to some human behaviors would be considered a contribution to intelligence. However, participants could hardly realize it if they could not interact with the system.

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 highly depend on the use scenario. When a virtual agent is designed very vivacious, people could consider it likeable as a dance teacher but less likeable as a technical expert.

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

6.1.2 Examining Multiple Measurement Techniques to Evaluate Gesture Performance

As of today, there is no standard method for assessing the performance of the ECA's gesture functionality. In our study, we derived dimensions from two validated scales to conduct a subjective analysis. One of these was the *Godspeed Questionnaire* developed by Bartneck et al. [6], the other was the *UES (User Engagement Scale)* by O'Brien et al. [70].

However, neither of the two questionnaires was specifically designed to evaluate ECA's ability to perform hand gestures, as the *Godspeed Questionnaire* is commonly used in the field of human-robot interaction and is used to evaluate robot performance in general. The *UES*, on the other hand, has a more extensive application domain, and may be used to measure user engagement across a variety of media, such as websites, games, and videos. Once we selected these two scales to serve our study, we sought to determine if the dimensional entries in the chosen domains were internally consistent. Our analysis revealed that the dimensions that we selected from both questionnaires presented good internal consistency in the sample data. In our knowledge, no other study has attempted to examine the effects of gestures on ECA in terms of *animacy*, *focused attention*, and *perceived intelligence* before this study.

Furthermore, an eye-tracking device was used to observe the gaze behavior of the participants. Therefore, we were able to objectively examine the results of the subjective questionnaire regarding the dimension of focused attention. Additionally, the collected eye-gazing data can be used for a more accurate analysis of the area of attention of the participants in each of the two experimental conditions. As far as we know, this is the first study that uses eye-tracking for assessing gesture performance.

6.2 Consideration on Sustainability and Ethics

Before starting the experiment, we distributed an informed consent form to each participant, notifying them of the type of data being collected and the purpose of the study. In order to make the experimental process as paperless as possible, we embedded the questionnaire as a module in the software system. This ensured a smooth experimental process and provided the possibility of minimising waste when conducting large-scale experiments in the future.

6.3 Consideration on Societal Aspects

Gestures are highly prevalent in social activities. Considering that ECAs with gestures can be used in a variety of contexts, such as games, customer service, teaching activities, etc., the acceptance of gestures by different cultures is an important societal consideration when designing gestural animations. As an example, it is vital to

minimize the frequency of gestures in cultures where they are less commonly used and to avoid gestures that could be offensive or misleading within a particular socio-cultural context.

6.4 Limitations and Future Work

The current study has several limitations. We could not get a sufficiently diverse sample of participants considering the small number of participants. 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 by using non-verbal behaviours. As indicated by our findings, objective user behaviour data can be beneficial for supplementary and prosperous outcomes. For further study, it could be vital to have the system able to collect multiple types of data. VR environments may be able to both immerse users while also collecting objective data on user behaviour (e.g., social distance).

After participating in the experiment, several participants commented that the current insufficiently intelligent dialogue system affected their grading of ECA on the dimension of *Perceived Intelligence*. Participants tended to focus too much on the inadequacies of the dialogue system, scoring higher in conditions in which the robot recognized their voice commands correctly or in which the robot pronounced the words correctly but paying little attention to the performance of the gestural animations. Therefore, we should also strive to improve the performance of the speech system in subsequent studies so that participants will instinctively ignore the speech component during the experiment.

Bibliography

[1] H. Admoni, T. Weng, B. Hayes, and B. Scassellati. “Robot nonverbal behavior improves task performance in difficult collaborations”. In: *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2016, pp. 51–58.

[2] S. Alexanderson, G. E. Henter, T. Kucherenko, and J. Beskow. “Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows”. In: *Computer Graphics Forum*. Vol. 39. 2. Wiley Online Library. 2020, pp. 487–496.

[3] A. Aly and A. Tapus. “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. 2013, pp. 325–332.

[4] R. Arnheim. “Hand and Mind: What Gestures Reveal about Thought by David McNeill”. In: *Leonardo* 27.4 (1994), pp. 358–358.

[5] J. N. Bailenson, E. Aharoni, A. C. Beall, R. E. Guadagno, A. Dimov, and J. Blascovich. “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. 2004, pp. 1864–1105.

[6] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi. “Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots”. In: *International journal of social robotics* 1.1 (2009), pp. 71–81.

[7] J. Bavelas, J. Gerwing, C. Sutton, and D. Prevost. “Gesturing on the telephone: Independent effects of dialogue and visibility”. In: *Journal of Memory and Language* 58.2 (2008), pp. 495–520.

[8] G. Beattie. *Visible thought: The new psychology of body language*. Routledge, 2004.

[9] K. W. Berger and G. R. Popelka. “Extra-facial gestures in relation to speechreading”. In: *Journal of Communication Disorders* 3.4 (1971), pp. 302–308.

[10] S. C. Broaders and S. Goldin-Meadow. “Truth is at hand: How gesture adds information during investigative interviews”. In: *Psychological Science* 21.5 (2010), pp. 623–628.

[11] J. K. Burgoon, T. Birk, and M. Pfau. “Nonverbal behaviors, persuasion, and credibility”. In: *Human communication research* 17.1 (1990), pp. 140–169.

[12] R. Calvo, S. D’Mello, J. Gratch, A. Kappas, M. Lhommet, and S. Marsella. *Expressing emotion through posture and gesture*. 2015.

[13] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone. “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*. 1994, pp. 413–420.

[14] J. Cassell and H. Vilhjálmsson. “Fully embodied conversational avatars: Making communicative behaviors autonomous”. In: *Autonomous agents and multi-agent systems* 2.1 (1999), pp. 45–64.

[15] G. Charness, U. Gneezy, and M. A. Kuhn. “Experimental methods: Between-subject and within-subject design”. In: *Journal of economic behavior & organization* 81.1 (2012), pp. 1–8.

[16] C.-C. Chiu and S. Marsella. “How to train your avatar: A data driven approach to gesture generation”. In: *International Workshop on Intelligent Virtual Agents*. Springer. 2011, pp. 127–140.

[17] E. Chong, N. Ruiz, Y. Wang, Y. Zhang, A. Rozga, and J. M. Rehg. “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)*. 2018, pp. 383–398.

[18] A. Deshmukh, B. Craenen, A. Vinciarelli, and M. E. Foster. “Shaping Robot Gestures to Shape Users’ Perception: The Effect of Amplitude and Speed on Godspeed Ratings”. In: *Proceedings of the 6th International Conference on Human-Agent Interaction*. 2018, pp. 293–300.

[19] G. Echterhoff, E. T. Higgins, and J. M. Levine. “Shared reality: Experiencing commonality with others’ inner states about the world”. In: *Perspectives on Psychological Science* 4.5 (2009), pp. 496–521.

[20] D. Efron, J. M. Efron, and S. Van Veen. *Gesture, race and culture: A tentative study of the spatio-temporal and” linguistic” aspects of the gestural behavior of eastern Jews and southern Italians in New York City, living under similar as well as different environmental conditions*. Vol. 9. Mouton, 1972.

[21] P. Ekman. “Why lies fail and what behaviors betray a lie”. In: *Credibility assessment*. Springer, 1989, pp. 71–81.

[22] M.
Fabri, D. Moore, and D. Hobbs. “Expressive agents: Non-verbal communication in collaborative virtual environments”. In: *Proceedings of Autonomous Agents and Multi-Agent Systems (Embodied Conversational Agents)* (2002).

[23] J. Feine, U. Gnewuch, S. Morana, and A. Maedche. “A taxonomy of social cues for conversational agents”. In: *International Journal of Human-Computer Studies* 132 (2019), pp. 138–161.

[24] G. Ferré. “Gesture/speech integration in the perception of prosodic emphasis”. In: *Proceedings of Speech Prosody 2018* (2018), pp. 35–39.

[25] S. Fitrianie, M. Bruijnes, D. Richards, A. Abdulrahman, and W.-P. Brinkman. “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*. 2019, pp. 159–161.

[26] S. Fitrianie, M. Bruijnes, D. Richards, A. Bönsch, and W.-P. Brinkman. “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*. 2020, pp. 1–8.

[27] J. E. Freeman. “Women: A feminist perspective.” In: (1975).

[28] O. Friborg, M. Martinussen, and J. H. Rosenvinge. “Likert-based vs. semantic differential-based scorings of positive psychological constructs: A psychometric comparison of two versions of a scale measuring resilience”. In: *Personality and Individual Differences* 40.5 (2006), pp. 873–884.

[29] A. Frischen, A. P. Bayliss, and S. P. Tipper. “Gaze cueing of attention: visual attention, social cognition, and individual differences.” In: *Psychological bulletin* 133.4 (2007), p. 694.

[30] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik. “Learning individual styles of conversational gesture”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3497–3506.

[31] S. Goldin-Meadow and M. W. Alibali. “Gesture’s role in speaking, learning, and creating language”. In: *Annual review of psychology* 64 (2013), pp. 257–283.

[32] J. A. Graham and M. Argyle. “A cross-cultural study of the communication of extra-verbal meaning by gestures”. In: *International Journal of Psychology* 10.1 (1975), pp. 57–67.

[33] M. M. Gross, E. A. Crane, and B. L. Fredrickson. “Methodology for assessing bodily expression of emotion”. In: *Journal of Nonverbal Behavior* 34.4 (2010), pp. 223–248.

[34] J. J. Gumperz and S. C. Levinson. “Rethinking linguistic relativity”. In: *Current Anthropology* 32.5 (1991), pp. 613–623.

[35] J. Ham, R. Bokhorst, R. Cuijpers, D. v. d. Pol, and J.-J. Cabibihan. “Making robots persuasive: the influence of combining persuasive strategies (gazing and gestures) by a storytelling robot on its persuasive power”. In: *International conference on social robotics*. Springer. 2011, pp. 71–83.

[36] D. Hasegawa, N. Kaneko, S. Shirakawa, H. Sakuta, and K. Sumi. “Evaluation of speech-to-gesture generation using bi-directional LSTM network”. In: *International Conference on Intelligent Virtual Agents*. ACM. 2018, pp. 79–86.

[37] D. D. Henningsen, K. S. Valde, and E. Davies. “Exploring the effect of verbal and nonverbal cues on perceptions of deception”. In: *Communication Quarterly* 53.3 (2005), pp. 359–375.

[38] G. E. Henter, S. Alexanderson, and J. Beskow. “Moglow: Probabilistic and controllable motion synthesis using normalising flows”. In: *ACM Transactions on Graphics (TOG)* 39.6 (2020), pp. 1–14.

[39] J. Holler and R. Stevens. “The effect of common ground on how speakers use gesture and speech to represent size information”. In: *Journal of Language and Social Psychology* 26.1 (2007), pp. 4–27.

[40] A. B. Hostetter. “When do gestures communicate? A meta-analysis.” In: *Psychological bulletin* 137.2 (2011), p. 297.

[41] C.-M. Huang and B. Mutlu. “Learning-based modeling of multimodal behaviors for humanlike robots”. In: *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2014, pp. 57–64.

[42] C.-M. Huang and B. Mutlu. “Modeling and Evaluating Narrative Gestures for Humanlike Robots.” In: *Robotics: Science and Systems*. 2013, pp. 57–64.

[43] A. L. Hubbard, S. M. Wilson, D. E. Callan, and M. Dapretto. “Giving speech a hand: Gesture modulates activity in auditory cortex during speech perception”. In: *Human brain mapping* 30.3 (2009), pp. 1028–1037.

[44] C. T. Ishi, D. Machiyashiki, R. Mikata, and H. Ishiguro. “A speech-driven hand gesture generation method and evaluation in android robots”. In: *IEEE Robotics and Automation Letters* 3.4 (2018), pp. 3757–3764.

[45] N. Jacobs and A. Garnham. “The role of conversational hand gestures in a narrative task”. In: *Journal of Memory and Language* 56.2 (2007), pp. 291–303.

[46] S. D. Kelly and L. H. Goldsmith. “Gesture and right hemisphere involvement in evaluating lecture material”. In: *Gesture* 4.1 (2004), pp. 25–42.

[47] S. Kita. “Cross-cultural variation of speech-accompanying gesture: A review”. In: *Language and cognitive processes* 24.2 (2009), pp. 145–167.

[48] S. Kita, A. Özyürek, S. Allen, A. Brown, R. Furman, and T. Ishizuka. “Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production”. In: *Language and cognitive processes* 22.8 (2007), pp. 1212–1236.

[49] M. L. Knapp, J. A. Hall, and T. G. Horgan. *Nonverbal communication in human interaction*. Cengage Learning, 2013.

[50] T. Kooijmans, T. Kanda, C. Bartneck, H. Ishiguro, and N. Hagita. “Accelerating robot development through integral analysis of human–robot interaction”. In: *IEEE Transactions on Robotics* 23.5 (2007), pp. 1001–1012.

[51] S. Kopp, B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, K. R. Thórisson, and H. Vilhjálmsson. “Towards a common framework for multimodal generation: The behavior markup language”. In: *International workshop on intelligent virtual agents*. Springer. 2006, pp. 205–217.

[52] C. Kramer. “Perceptions of female and male speech”. In: *Language and speech* 20.2 (1977), pp. 151–161.

[53] L. L. Kramer, S. Ter Stal, B. C. Mulder, E. de Vet, and L. van Velsen. “Developing embodied conversational agents for coaching people in a healthy lifestyle: scoping review”. In: *Journal of medical Internet research* 22.2 (2020), e14058.

[54] R. M. Krauss, R. A. Dushay, Y. Chen, and F. Rauscher. “The communicative value of conversational hand gesture”. In: *Journal of experimental social psychology* 31.6 (1995), pp. 533–552.

[55] A. Krogsager, N. Segato, and M. Rehm. “Backchannel head nods in danish first meeting encounters with a humanoid robot: The role of physical embodiment”. In: *International Conference on Human-Computer Interaction*. Springer. 2014, pp. 651–662.

[56] T. Kucherenko, D. Hasegawa, G. E. Henter, N. Kaneko, and H. Kjellström. “Analyzing input and output representations for speech-driven gesture generation”. In: *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. 2019, pp. 97–104.

[57] T. Kucherenko, D. Hasegawa, N. Kaneko, G. E. Henter, and H. Kjellström. “Moving fast and slow: Analysis of representations and post-processing in speech-driven automatic gesture generation”. In: *International Journal of Human–Computer Interaction* (2021), pp. 1–17.

[58] T. Kucherenko, P. Jonell, S. van Waveren, G. E. Henter, S. Alexandersson, I. Leite, and H. Kjellström. “Gesticulator: A framework for semantically-aware speech-driven gesture generation”. In: *Proceedings of the 2020 International Conference on Multimodal Interaction*. 2020, pp. 242–250.

[59] T. Kucherenko, P. Jonell, Y. Yoon, P. Wolfert, and G. E. Henter. “A large, crowdsourced evaluation of gesture generation systems on common data: The GENEA Challenge 2020”. In: *26th International Conference on Intelligent User Interfaces*. 2021, pp. 11–21.

[60] Q. A. Le and C. Pelachaud. “Evaluating an expressive gesture model for a humanoid robot: Experimental results”. In: *Submitted to 8th ACM/IEEE International Conference on Human-Robot Interaction*. 2012.

[61] S. Levine, P. Krähenbühl, S. Thrun, and V. Koltun. “Gesture controllers”. In: *ACM SIGGRAPH 2010 papers*. 2010, pp. 1–11.

[62] J. Linssen, M. Theune, M. Bruijnes, et al. “Beyond simulations: serious games for training interpersonal skills in law enforcement”. In: *Social simulation conference*. 2014.

[63] C. Lisetti, R. Amini, U. Yasavur, and N. Rishe. “I can help you change! an empathic virtual agent delivers behavior change health interventions”. In: *ACM Transactions on Management Information Systems (TMIS)* 4.4 (2013), pp. 1–28.

[64] J. Liu et al. “Analysis of gender differences in speech and hand gesture coordination for the design of multimodal interface systems”. In: (2014).

[65] F. Maricchiolo, A. Gnisci, M. Bonaiuto, and G. Ficca. “Effects of different types of hand gestures in persuasive speech on receivers’ evaluations”. In: *Language and cognitive processes* 24.2 (2009), pp. 239–266.

[66] M. Marmpena, F. Garcia, and A. Lim. “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. 2020, pp. 357–359.

[67] R. Nagy, T. Kucherenko, B. Moell, A. Pereira, H. Kjellström, and U. Bernardet. "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. 2021, pp. 1779–1781.

[68] Y. I. Nakano and R. Ishii. "Estimating user's engagement from eye-gaze behaviors in human-agent conversations". In: *Proceedings of the 15th international conference on Intelligent user interfaces*. 2010, pp. 139–148.

[69] M. Neff, M. Kipp, I. Albrecht, and H.-P. Seidel. "Gesture modeling and animation based on a probabilistic re-creation of speaker style". In: *ACM Transactions on Graphics (TOG)* 27.1 (2008), pp. 1–24.

[70] H. L. O'Brien, P. Cairns, and M. Hall. "A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form". In: *International Journal of Human-Computer Studies* 112 (2018), pp. 28–39.

[71] B. Ravenet, C. Pelachaud, C. Clavel, and S. Marsella. "Automating the production of communicative gestures in embodied characters". In: *Frontiers in psychology* 9 (2018), p. 1144.

[72] P. Sajjadi, L. Hoffmann, P. Cimiano, and S. Kopp. "A personality-based emotional model for embodied conversational agents: Effects on perceived social presence and game experience of users". In: *Entertainment Computing* 32 (2019), p. 100313.

[73] M. Salem, F. Eyssel, K. Rohlfing, S. Kopp, and F. Joublin. "To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability". In: *International Journal of Social Robotics* 5.3 (2013), pp. 313–323.

[74] M. Salem, S. Kopp, I. Wachsmuth, K. Rohlfing, and F. Joublin. "Generation and evaluation of communicative robot gesture". In: *International Journal of Social Robotics* 4.2 (2012), pp. 201–217.

[75] J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P. W. McOwan, and A. Paiva. "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*. 2011, pp. 305–312.

[76] S. Saunderson and G. Nejat. "How robots influence humans: A survey of nonverbal communication in social human–robot interaction". In: *International Journal of Social Robotics* 11.4 (2019), pp. 575–608.

[77] K. Schaefer. "The perception and measurement of human-robot trust". In: (2013).

[78] J. Sebastian and D. Richards. "Changing stigmatizing attitudes to mental health via education and contact with embodied conversational agents". In: *Computers in Human Behavior* 73 (2017), pp. 479–488.

[79] H. Shah. "Malware and cybercrime: the threat from artificial dialogue systems". In: *Written Evidence in UK House of Commons Parliamentary Science and Technology Select Committee Report on Malware and Cybercrime* (2012).

[80] É. Székely, G. E. Henter, J. Beskow, and J. Gustafson. "Spontaneous Conversational Speech Synthesis from Found Data." In: *INTERSPEECH*. 2019, pp. 4435–4439.

[81] A. Thangthai, K. Thangthai, A. Namsanit, S. Thatphithakkul, and S. Saychum. "The Nectec gesture generation system entry to the GENEA Challenge 2020". In: *Proc. GENEA Workshop. https://doi. org/10.5281/zenodo*. Vol. 4088629. 2020.

[82] D. Traum and J. Rickel. "Embodied agents for multi-party dialogue in immersive virtual worlds". In: *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2*. 2002, pp. 766–773.

[83] L. Valenzeno, M. W. Alibali, and R. Klatzky. "Teachers' gestures facilitate students' learning: A lesson in symmetry". In: *Contemporary Educational Psychology* 28.2 (2003), pp. 187–204.

[84] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder. "Bridging the gap between social animal and unsocial machine: A survey of social signal processing". In: *IEEE Transactions on Affective Computing* 3.1 (2011), pp. 69–87.

[85] A. Weiss and C. Bartneck. “Meta analysis of the usage of the godspeed questionnaire series”. In: *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2015, pp. 381–388.

[86] P. Wolfert, J. M. Girard, T. Kucherenko, and T. Belpaeme. “To rate or not to rate: Investigating evaluation methods for generated co-speech gestures”. In: *Proceedings of the 2021 International Conference on Multimodal Interaction*. 2021, pp. 494–502.

[87] P. Wolfert, N. Robinson, and T. Belpaeme. “A review of evaluation practices of gesture generation in embodied conversational agents”. In: *arXiv preprint arXiv:2101.03769* (2021).

[88] Y. Yoon, W.-R. Ko, M. Jang, J. Lee, J. Kim, and G. Lee. “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. 2019, pp. 4303–4309.

TRITA -EECS-EX-2022:49