applied sciences

MDPI

Article

Exploring the Effectiveness of Evaluation Practices for Computer-Generated Nonverbal Behaviour

Pieter Wolfert [1,2,*], Gustav Eje Henter [3] and Tony Belpaeme [1]

[1] IDLab-Airo, Ghent University-imec, 9052 Ghent, Belgium; tony.belpaeme@ugent.be
[2] Donders Institute for Brain, Cognition & Behaviour, Radboud University, 6500 HB Nijmegen, The Netherlands
[3] KTH Royal Institute of Technology, 10044 Stockholm, Sweden; ghe@kth.se
* Correspondence: pieter.wolfert@donders.ru.nl

Abstract: This paper compares three methods for evaluating computer-generated motion behaviour for animated characters: two commonly used direct rating methods and a newly designed questionnaire. The questionnaire is specifically designed to measure the human-likeness, appropriateness, and intelligibility of the generated motion. Furthermore, this study investigates the suitability of these evaluation tools for assessing subtle forms of human behaviour, such as the subdued motion cues shown when listening to someone. This paper reports six user studies, namely studies that directly rate the appropriateness and human-likeness of a computer character's motion, along with studies that instead rely on a questionnaire to measure the quality of the motion. As test data, we used the motion generated by two generative models and recorded human gestures, which served as a gold standard. Our findings indicate that when evaluating gesturing motion, the direct rating of human-likeness and appropriateness is to be preferred over a questionnaire. However, when assessing the subtle motion of a computer character, even the direct rating method yields less conclusive results. Despite demonstrating high internal consistency, our questionnaire proves to be less sensitive than directly rating the quality of the motion. The results provide insights into the evaluation of human motion behaviour and highlight the complexities involved in capturing subtle nuances in nonverbal communication. These findings have implications for the development and improvement of motion generation models and can guide researchers in selecting appropriate evaluation methodologies for specific aspects of human behaviour.

Keywords: human–computer interaction; embodied conversational agents; subjective evaluations

check for updates

Citation: Wolfert, P.; Henter, G.E.; Belpaeme, T. Exploring the Effectiveness of Evaluation Practices for Computer-Generated Nonverbal Behaviour. *Appl. Sci.* **2024**, *14*, 1460. https://doi.org/10.3390/app14041460

Academic Editor: Daeho Lee

Received: 18 December 2023
Revised: 24 January 2024
Accepted: 2 February 2024
Published: 10 February 2024

1. Introduction

In this paper, we address the subjective evaluation of generated nonverbal behaviour for embodied conversational agents. We specifically focus on two common methods for assessing speech gesture behaviour and listening behaviour. Additionally, we introduce a newly designed questionnaire for the evaluation of speech gesture behaviour as well as listening behaviour based on recommendations provided in a recent review [1].

Nonverbal behaviour, encompassing elements such as eye gaze, blinking, and co-speech gestures, plays a vital role in facilitating effective human communication [2]. Among the various forms of nonverbal behaviour, co-speech gestures take up a major role [3]. Research has shown that incorporating co-speech gestures can improve human-human communication. For example, Holler et al. [4] discovered that when questions were accompanied by gestures, the time between questions and answers was reduced. Similar effects can be observed when humans interact with embodied conversational agents (ECAs), as the inclusion of nonverbal behaviours such as co-speech gestures and full-body motion has been found to improve the interaction between humans and ECAs [5,6]. Furthermore, researchers found that including nonverbal behaviours increased the willingness to cooperate with an ECA [7].

Due to the potential of nonverbal behaviour to enhance the human-likeness and communicative efficacy of ECAs, the automatic generation of nonverbal behaviour has become a major research focus. For the field of human motion generation, data-driven methods have gained popularity over recent years [8–11].

Since the aim of nonverbal behaviour generation for ECAs is to enhance ECAs' interactions with people, proper assessment of these technologies must involve people as well (i.e., use subjective evaluation methods). This is often accomplished through Likert scales [1], in interactive scenarios through listening comprehension [12], through direct questions [13], or using behavioural metrics [14]. In related fields such as human–robot interaction (HRI), aspects such as anthropomorphism, animacy, and likeability are evaluated through the use of questionnaires, of which the Godspeed questionnaire is the most common one [15]. Providing these scales with indirect items instead of asking users to evaluate stimuli directly has the benefit that more nuanced information becomes available, and it becomes easier to compare the outcomes of different studies due to the questionnaire being used across different studies.

This paper is an extension of a recently accepted late-breaking report [16] where we focused on generated listening behaviour, while the focus in this paper is more on evaluation methodologies. In comparison with the conference publication, we introduce four new user studies.

In this study, we explore three different approaches to evaluating generated co-speech gestures and listening behaviour, aiming to assess their effectiveness and understand their qualities. Previous research commonly relied on questionnaires and Likert scales to evaluate motion and behaviour quality [1]. Following the recommendations and insights from that review, we introduce a new questionnaire that enables us to compare methodological results to previously used methodologies for assessing human-likeness (introduced in [17,18]) and appropriateness (through mismatching, as introduced in [19,20]). Our objective is to determine the potential benefits of using a standardised questionnaire in this field and examine how well it correlates with the commonly used direct subjective measurements.

A particular goal of this article is to study the effect of different evaluation methods in scenarios where the behaviours evaluated and the differences between them are quite subtle. To this end, we trained two existing motion generation models using dyadic conversational data to generate head, arm, and body motion corresponding to listening, which typically is less vivid than gesture motions performed during active speaking. By evaluating these same models using a number of different user study methodologies, different evaluation methods can be compared. Additionally, we can compare the outputs of these two models for both full-body speech and full-body listening motion and gain insight into how well current and new evaluation methods are able to distinguish differences for subtle forms of generated behaviour. Specifically, we conducted six user studies, namely two appropriateness studies, two human-likeness studies, and two studies that incorporated our newly designed questionnaire. In the human-likeness studies, the participants provided direct ratings on a scale from 0 to 100 to assess the human-likeness of the stimuli, following the HEMVIP paradigm [18]. The appropriateness studies employed the matching/mismatching paradigm previously utilised in [13,19–21].

The proposed questionnaire, used in the last two studies, includes questions covering the constructs of appropriateness, human-likeness, and intelligibility. The selection of these constructs is based on their recurrence in previous studies identified in [1].

In conclusion, we recommend using direct rating or side-by-side comparisons of computer-generated nonverbal behaviour. These are to be preferred over questionnaires, as questionnaires tend to not pick up subtle qualitative differences in behaviour and are not calibrated between raters.

This paper is organised as follows. Section 2 covers the related work on nonverbal behaviour generation and evaluation. In Section 3, we delve into the rationale behind the selection of constructs for our questionnaire, highlighting the considerations and motivations that guided our choices. Section 4 presents our methodology and details about the data set, stimuli, and models. We present our results in Section 5 and discuss our findings in Section 6, where we also draw conclusions with previous research on this topic.

2. Related Work

This section provides an overview of gesture generation, including a discussion of existing research and the current state of the art in generating gestures, followed by a review of the literature on listening behaviour, and concluding with an exploration of evaluation strategies for generated human behaviour.

2.1. Generating and Evaluating Gesture Behaviour

Numerous studies have focused on generating speech-driven motion for embodied conversational agents. For instance, Kucherenko et al. [22] leveraged representation learning to map audio to motion, while Yoon et al. [23] used input text with word-level timestamps to generate motion without using speech audio. Subsequently, other researchers have combined both audio and text representations of speech along with speaker identity in their gesture generation models, such as those in [24–26]. Since the goal of gesture generation for ECAs is to facilitate effective human–agent interaction, some research has furthermore explored generating nonverbal behaviour while considering the interlocutor in the interaction [27,28]. Others have applied more generative approaches to gesture generation, with the aim of learning a probabilistic distribution [24]. More recently, researchers started picking up diffusion models for gesture synthesis [10,29,30]. However, comparing different gesture generation models is challenging, as highlighted by Wolfert et al. [1]. The GENEA Challenge [13,17] was set up to address this issue by allowing multiple teams to build models on a shared data set and submit motions from their models to a shared evaluation. A more in-depth review on the field of gesture generation, especially considering deep learning, can be found in [31].

In contrast to the work discussed here, we train an unsupervised probabilistic model on a datas et containing dyadic interactions. These interactions are not cut to only include speech and gesticulation; instead, the interactions contain the full range of nonverbal motion one would expect to see during a one-on-one interaction.

2.2. Listening Behaviour

Listening is an essential aspect of human–agent interaction, and studies have shown that virtual agents who pretend to listen can enhance engagement during an interaction [32]. For instance, Buschmeier et al. [33] showed that when humans interacted with an attentive agent, they were more likely to provide listener feedback and rated the agent as more helpful. Maatman, Gratch, and Marsella [34] proposed a model that generates listening behaviour based on available features during a conversation. Their system extracts audio and body posture features to drive listening behaviour. Another approach by Gillies et al. [35] utilised input audio from the speaker to generate listening behaviour through motion graphs, where existing motion clips were combined to match new audio input. Mlakar [36] introduced a framework and scripting method to synthesise both verbal and nonverbal motion which entails both gestures and listening. Poppe et al. [37] developed rule-based strategies for generating listening behaviour based on the speaker's speech and gaze, including vocal back channelling. A similar approach in terms of selecting new listening behaviours and sequences can be found in [38]. They used a multi-modal corpus of interviews to generate listening behaviour in a virtual agent conducting interviews. The participants perceived the interviewer as affiliative when the interviewer would mirror their posture. An example of generating listening head behaviour is the work by Jonell et al. [19]. They generated interlocutor-aware facial gestures using nonverbal and verbal input from both the interlocutor and agent using a generative approach. In our work, we include full conversational data from dyadic interactions to generate listening behaviour based on the audio of both participants.

2.3. Evaluation Strategies for Generated Human Behaviour

The previous two sections described the state of the art in the fields of both gesture generation and listening behaviour generation. But how is the generated behaviour in these works compared, and how is quality assessed? A recent review by Wolfert et al. [1] discusses subjective evaluation methodologies used in the field of gesture generation for embodied conversational agents. Their work identified that a majority of studies make use of questionnaires to assess the quality of generated motion. Questionnaires, employing Likert scales, are a widely used tool to assess one's attitude towards a concept [39]. One such example is the Godspeed questionnaire [15], which originates from and is used in the field of human–robot interaction and measures the concepts of anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety. Wolfert et al. [1] called for standardisation given the large variety in reported constructs for the field of gesture generation (which consist of multiple items). In support of this, a 2019 review by Fitrianie et al. [40] that looked into questionnaire usage at intelligent virtual agent conferences (IVAs) found that for 76% of the studies, questionnaires were unique for those studies and were not reused in other studies. Having matching constructs and questionnaire items would make it easier to compare between studies. Efforts for standardisation are underway for the evaluation of virtual agents, but these efforts do not specifically include the evaluation of nonverbal generated behaviour for ECAs [41]. Among the constructs that they reported are human-likeness, appropriateness, naturalness or effectiveness, and understanding. Even though there is a degree of overlap in the constructs between studies, this does not mean that they are measured in the same way or that they contain the same items (statements) and response scales.

There have been other, more direct methods proposed for measuring these constructs that do not involve questionnaires or answering multiple statements. For appropriateness, recent studies introduced a new evaluation paradigm that makes use of matched and mismatched gesture motion [13,19,20]. In appropriateness for speech, the concept involves combining two motion fragments and one audio segment. Both motion fragments should come from the same condition or system. One of the motion fragments is directly associated with the audio and represents the intended match, while the other fragment is randomly selected and does not have a direct connection to the corresponding audio, resulting in a mismatch. The measured subjective preference for matched motion stimuli over mismatched ones then quantifies how specifically appropriate the motion in question is for the speech. For naturalness or human-likeness, recent studies have relied on directly asking participants to indicate human-likeness using a fine-grained slider [13,17,42]. These studies accomplished this using the HEMVIP framework [18], which labels the scale with five anchors (bad, poor, fair, good, and excellent) and associates each position of the slider with a score between 0 and 100 [18]. This framework makes it possible to evaluate multiple different gesture videos for the same speech in parallel. Wolfert et al. compared the HEMVIP framework with pairwise comparisons and found that under certain circumstances, pairwise comparisons could be a faster way of assessing quality, but it has the downside of scaling badly with multiple conditions [42]. As an example of studies to evaluate aspects of gesture understanding, He et al. conducted a study to evaluate the effectiveness and comprehension of generated gestures in an interactive scenario using a virtual avatar [14]. Due to the challenges of online assessments, the researchers employed various methods, including dimensions from the Godspeed questionnaire and direct participant evaluations of human-likeness, along with gaze data. Interestingly, only the behavioural data provided support for their findings, as the results from the questionnaire did not show significant differences.

In conclusion, we can observe that in certain cases, questionnaire constructs are used to measure the attitudes of participants towards a specific construct (such as human-likeness or appropriateness). However, due to the lack of a standardised questionnaire, it is near impossible to compare the results of different studies. Other methodologies of direct measurements have a downside in that they only quantify one thing at a time.

3. Designing a Questionnaire for ECAs

We put forward a questionnaire that instead covers the concepts of appropriateness, human-likeness, and intelligibility, which are key foci of nonverbal behaviour generation for the ECA community. The proposed questionnaire consists of three Likert scales with five Likert items each, which can be found in Figure 1. Wolfert et al. [1] identified multiple studies that included questions on appropriateness or speech–gesture correlation. We based the statements that are part of the constructs on previous included statements as identified in [1]. The GENEA Challenge 2020 used a direct question related to appropriateness. We decided to include the construct of 'appropriateness', with five Likert items (statements) related to the appropriateness of the motion behaviour for the conversation. Since the concept of human-likeness often comes up in subjective evaluations and has also been used for direct questioning by the GENEA Challenge, we came up with five Likert items related to human-likeness of a gesture motion. Lastly, we want to evaluate the intelligibility of the agent or speaker motion, as this also regularly appears in subjective evaluations of synthesised gesture motion (such as 'content' or 'utilisation of gesture' per Table 3 in the work by Wolfert et al. [1]).

Appropriateness of the motion

- The motion seemed appropriate for the context of the conversation.
- The motion felt out of place or irrelevant to the interaction.
- The motion did not distract from the conversation.
- The motion was in sync with the pace of the interaction.
- The motion was in synchronisation with the agent's tone of voice and emotion or his or her active listening.

Human-likeness of the motion

- The motion did not look like it was produced by a human.
- The motion appeared smooth and effortless.
- The motion had the same characteristics as human motion.
- The motion seemed forced or robotic.
- The speed of the motion looked human-like.

Intelligibility of the agent

- The motion enhanced the understanding of the interaction.
- The motion captured what the character was trying to express well.
- The meaning of the motion was easy to interpret.
- The motion helped me understand what the person was saying or showed that he or she was actively listening.
- The motion added to the perception of the agent's strong communication skills.

Figure 1. Participants were asked to rate each statement in the questionnaire on a scale from 1 to 5 using the following anchors: (1) disagree, (2) slightly disagree, (3) neither agree nor disagree, (4) slightly agree, and (5) agree.

4. Materials and Methods

In this section, we first discuss the data, processing, and models used in our studies. We finish with a description of the six user studies.

4.1. Data and Preprocessing

To ensure that the SG model was applicable to a wider range of conversational interactions, we opted to train it on a data set that included human dyadic interactions rather than just a single speaker. Our data set of choice was 'Talking with Hands 16.2', which provides a rich source of dyadic conversational data. This data set includes both motion capture and audio, totalling 50 h of recorded interactions. As the baseline model relies on text input for generating co-speech gestures, we made use of annotations provided by the GENEA Challenge 2022 [13]. For our experiments, we utilised a subset of 10 h of conversation data from the combined data set. We opted to only include conversational takes that included the speaker labelled 'deep5' in the original data as a participant, since this was the single speaker with the most data in the data set. Furthermore, we conducted a thorough manual inspection of the data set to exclude takes that exhibited significant motion errors.

By adhering to these selection and inspection processes, we aimed to create a reliable and high-quality data set for training and evaluation purposes. The audio channel was transformed into a 27 channel mel-frequency representation following the original paper on SG [8]. The resulting features were downsampled to 30 frames per second (FPS) to match up with the frame rate of the motion. Poses (joint rotations) were represented using exponential maps, which prevented discontinuities [43], and full-body motion was used, excluding finger and facial information. The input data for the model consisted of the concatenated audio and speaker identity as well as the motion of the interlocutor.

4.2. Models

To evaluate the three evaluation methods introduced earlier, we generated stimuli from two motion models and the ground truth motion and rendered this motion on an avatar. The StyleGestures model was taken as one of the models for its generative capabilities, and we adapted it to work with dyadic conversational data. As we aimed for a fair comparison in relation to the ground truth motion data, we also trained another model named 'baseline'.

4.2.1. StyleGestures

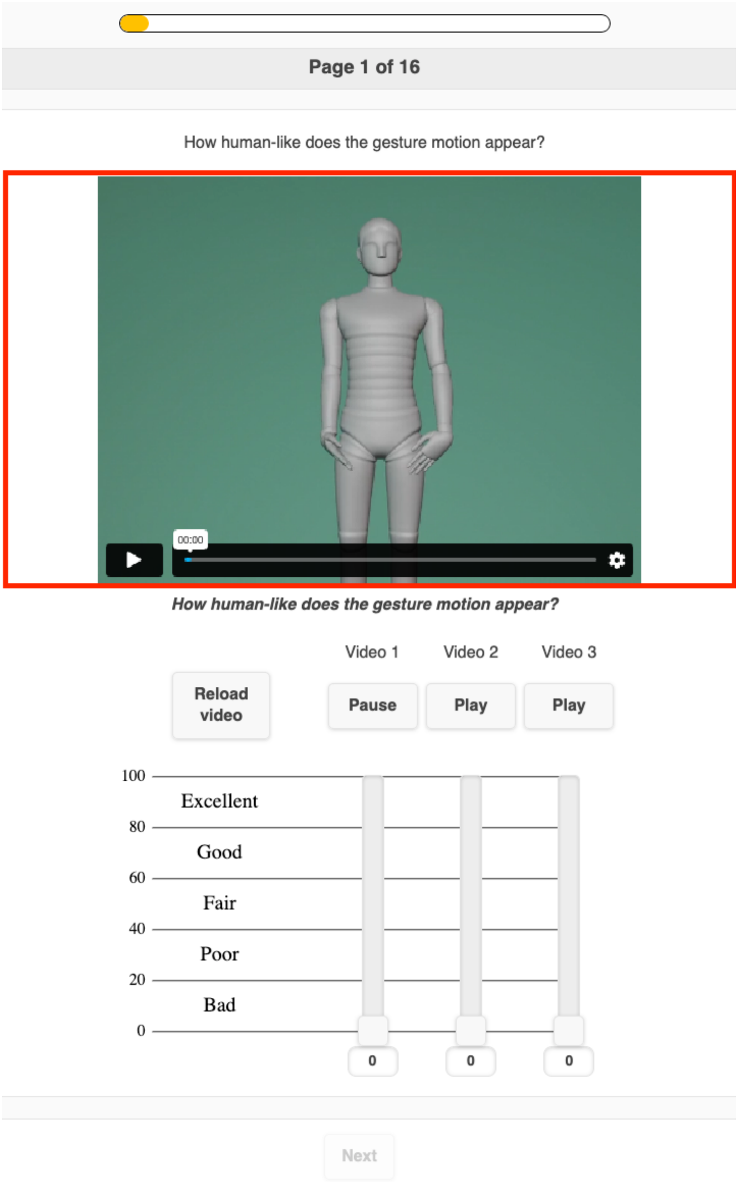
The StyleGestures (SG) model [8] is a probabilistic generative sequence model based on MoGlow which uses normalizing flows [44,45]. The model was modified to accept dyadic input (speaker 1 and speaker 2), with the input being a concatenation of two audio streams, a one-hot encoding of the speaker identity, and the motion stream of the interlocutor (speaker 2). The output of the model was joint angles using the exponential map for speaker 1. The modified SG model was trained using the standard parameters from the SG paper, with a batch size of 120, noam_learning_rate_decay with 3000 warm up steps, and a minimum learning rate of 0.00015. The optimiser used was Adam, with a learning rate of 0.0015. Since the input data for this version of SG deals with dyadic information, they are much larger than the original dimensions of the input data in [8], which only featured processed audio. Therefore, the model was trained for 160,000 steps before test motion was generated. We applied postprocessing to the motion data to improve the quality of our generated listening behaviour. Specifically, we used a Butterworth low-pass filter to smooth the rotation data and filter out minor motion glitches. The cutoff frequency was set to 3.0 Hz, and the filter order was set to 4, as this was found to provide good motion results. We conducted user studies to compare the output of this model to the ground truth.

4.2.2. Baseline

We wanted to compare our results to a model that had already been applied to the data set we used. For this, we selected 'The IVI Lab entry to the GENEA Challenge 2022', since the code for this entry was openly available and tested by others, winning the reproducibility award at the challenge [46]. The baseline model is based on the Tacotron2 architecture from speech synthesis with a locality constraint attention mechanism, and it takes text and speech audio as input to generate motion data [47]. It was trained on only the text and speech input data from the speaker whose motion we were predicting, namely speaker 1 (in contrast to our SG model, which was trained on full dyadic data). For the training parameters, we relied on the values used in [46].

4.3. Visualisation

We rendered the generated (or recorded for the ground truth condition) motion on a faceless avatar, which was provided by the GENEA Challenge 2022. The hands used a fixed pose since we did not attempt to learn the finger motion (which would increase complexity, and the finger motion data in the data set are of poor quality). A screenshot including the avatar we used can be seen in Figure 2. This screenshot displays the three rating bars (one for each video), following the standardised evaluation strategy introduced in [18]. To play one of the other videos, the user has to press play for the other video, which will then be loaded. For the video stimuli for the six studies, see 'Ground truth motion (for both listening and speaking)' (https://drive.google.com/file/d/1sRdbgNrxAB6WMnciMjXJJvi-sxiSOkOA/view?usp=sharing), ('Baseline speech motion on the left') (https://drive.google.com/file/d/1ODB1x6SMBzsVWGrrbmG2Vj23aT4vRiIF/view?usp=sharing), 'Baseline listening motion on the left' (https://drive.google.com/file/d/1mwZblDMB6eOGKPBDrtZMD_N1n04GAmIz/view?usp=sharing), 'StyleGestures speech motion on the left' (https://drive.google.com/file/d/17IpQQZEM8btcbOqH5Xe1cFluqaguj7ZU/view?usp=sharing), and 'StyleGestures listening motion on the left' (https://drive.google.com/file/d/19M8EkufozXGEAfI0wDos0BmdHfDmag7q/view?usp=sharing).



Page 1 of 16

How human-like does the gesture motion appear?

How human-like does the gesture motion appear?

Figure 2. A screenshot that displays the avatar in the HEMVIP interface [18]. This interface was used for studies 1 and 3. Each play button is linked to one video, and only one video was shown at the same time. The user had to rate each video before being able to continue to the next page.

4.4. User Studies

4.4.1. General Set-Up

For each study, we recruited participants through Prolific. We restricted the pool of subjects using the built-in features of the platform. The participants were required to be located in either the USA, UK, Ireland, Australia, Canada, or New Zealand. In addition, we required the participants to be native English speakers. We paid between GBP 1.80 and 2.25 per experiment, resulting in an average hourly wage between GBP 7.50 and 9.50. The median time to complete a study was between 11 and 20 min (this differed per study). Each participant could only complete a study once and was required to perform the study on a PC and not on a small mobile device such as a phone or tablet. The rendered videos were hosted on Vimeo, and the file names were randomised so that the participants could not infer conditions from the file names by looking at the source code of the interface. Right before being sent back to Prolific, the participants were asked to fill out a short questionnaire that asked for their demographics and their experience with the experiment. We piloted the studies locally by sending the studies to peers, and based on their feedback, we launched the final studies on Prolific. These studies were approved by the ethical committee of the Faculty of Psychology and Pedagogical Sciences at Ghent University in Belgium. Table 1 shows an overview of the studies we will discuss in this section.

Table 1. Overview of the studies, the study goal per study, and the question that the participants were asked.

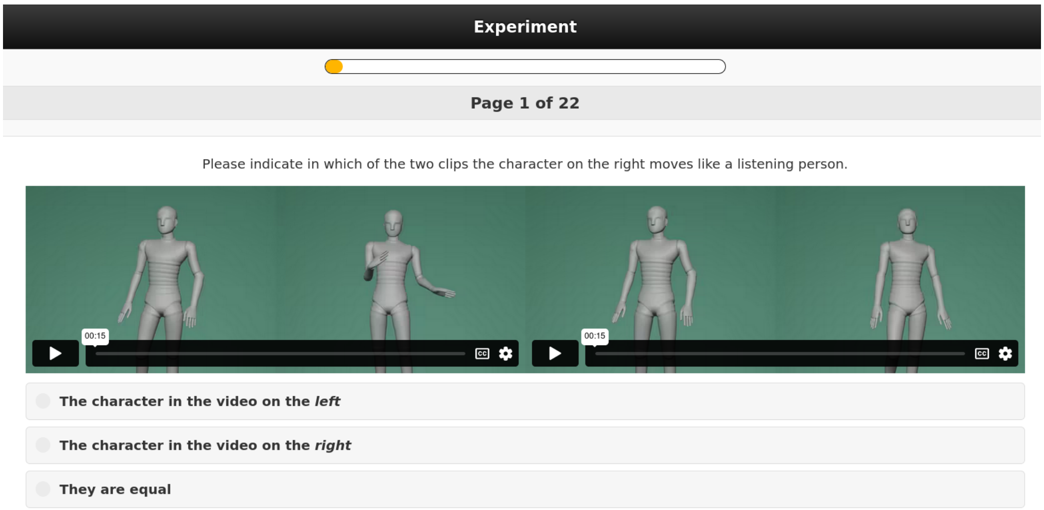
Study	Study Goal	Question Posed
1	Human-likeness of the generated gesture motion	How human-like does the gesture motion appear?
2	Appropriateness of the generated gesture motion for the conversation	Indicate which character's motion best matches speech in terms of rhythm, intonation, and meaning.
3	Human-likeness of the generated listening motion	How human-like does the listening motion appear?
4	Appropriateness of the generated listening motion for the conversation	Indicate which character's motion shows the most appropriate listening behaviour, considering the speaker's motion.
5	Agreement on three constructs for the generated gesture motion: appropriateness, human-likeness, and intelligibility	See Figure 4.
6	Agreement on three constructs for the generated listening motion: appropriateness, human-likeness, and intelligibility	See Figure 4.

4.4.2. Study 1: Human-Likeness for Gesticulation

The purpose of this study was to investigate how human-like a gesture motion was when generated by SG in comparison with the baseline and ground truth motion. We recruited 22 participants. From the test set, 30 segments were randomly selected in which the speaker was talking, and for each stimulus, gesture motion was synthesised from SG and the baseline or taken from the ground truth. Each segment had a length of between 6 and 12 s. The participants were asked the following question: 'How human-like does the gesture motion appear?' We did not include audio when rendering the videos, as the inclusion of audio would affect the ability of the participants to rate the stimuli for human-likeness. Three videos were placed on one evaluation screen using the HEMVIP framework for evaluating the stimuli, which has been validated and used by others before [13,17,18,42] based on webMUSHRA [48]. The participants could press play to start one of the videos, and at all times, only one video was visible (see also Figure 2 for the interface). The order of the videos on the screen was randomised, as well as the order in which the screens were presented to the participant. Each participant was presented with two attention check videos inserted randomly during the experiment. Both attention checks would ask the participant to rate the video with a specific score. The text for the attention check would only appear halfway through the video (so that it would take out participants that did not watch the complete video or that were not paying attention to the video at all). The participants were asked to rate the human-likeness on a scale from 0 to 100, where a score of 100 would mean the gesture motion was completely human-like and excellent. The rating scale was accompanied by five anchors that were equally spaced, namely bad, poor, fair, good, and excellent. Each participant rated 14 screens with 3 stimuli per screen, totalling 42 ratings per participant and 308 ratings per condition.

4.4.3. Study 2: Appropriateness for Gesticulation

For this study, we examined the appropriateness of the gesture motion for the speech generated by the model. We followed the appropriateness paradigm introduced by Rebol et al. [20], in which matching and mismatching stimuli are put on one screen side-by-side. We recruited 27 participants. To form our stimuli, we took the same 30 segments used in study 1 and chose 30 additional segments as mismatching stimuli. These segments were then paired with the interlocutor, resulting in two avatars being visible in each video. The speaker was placed on the left side, whereas the interlocutor was placed on the right side. For each of the 30 videos, we provided a mismatching video with motion unrelated to that part of the conversation. These videos were paired with the matching interlocutor. To establish an appropriateness baseline, we included matched and mismatched videos from the ground truth. We hypothesised that the participants would be able to identify the correct segments for direct motion-captured gesticulation. Both videos were placed on the same page, and the participants were asked to indicate in which of the two clips the character on the left moved appropriately for the speech. The interface for the user study followed the one that was designed in [42] for their study involving pairwise comparisons. See Figure 3 for a screenshot of the interface. The participants had a choice between three options: the left video, the right video, or both being equal. Throughout the experiment, each participant encountered two attention checks inserted into random places during the experiment. One attention check was text-based, and the other one was audio-based. Halfway through the video, it would ask the participant to select the button belonging to that specific video. We used Barnard's test for identifying statistically significant differences between conditions at the level of $\alpha = 0.05$. Additionally, we applied the Holm–Bonferroni method to correct for multiple comparisons.



Experiment

Page 1 of 22

Please indicate in which of the two clips the character on the right moves like a listening person.

○ **The character in the video on the *left***

○ **The character in the video on the *right***

○ **They are equal**

Figure 3. A screenshot of the pairwise interface (as introduced in [42]) used in studies 2 and 4.

4.4.4. Study 3: Human-Likeness for Listening

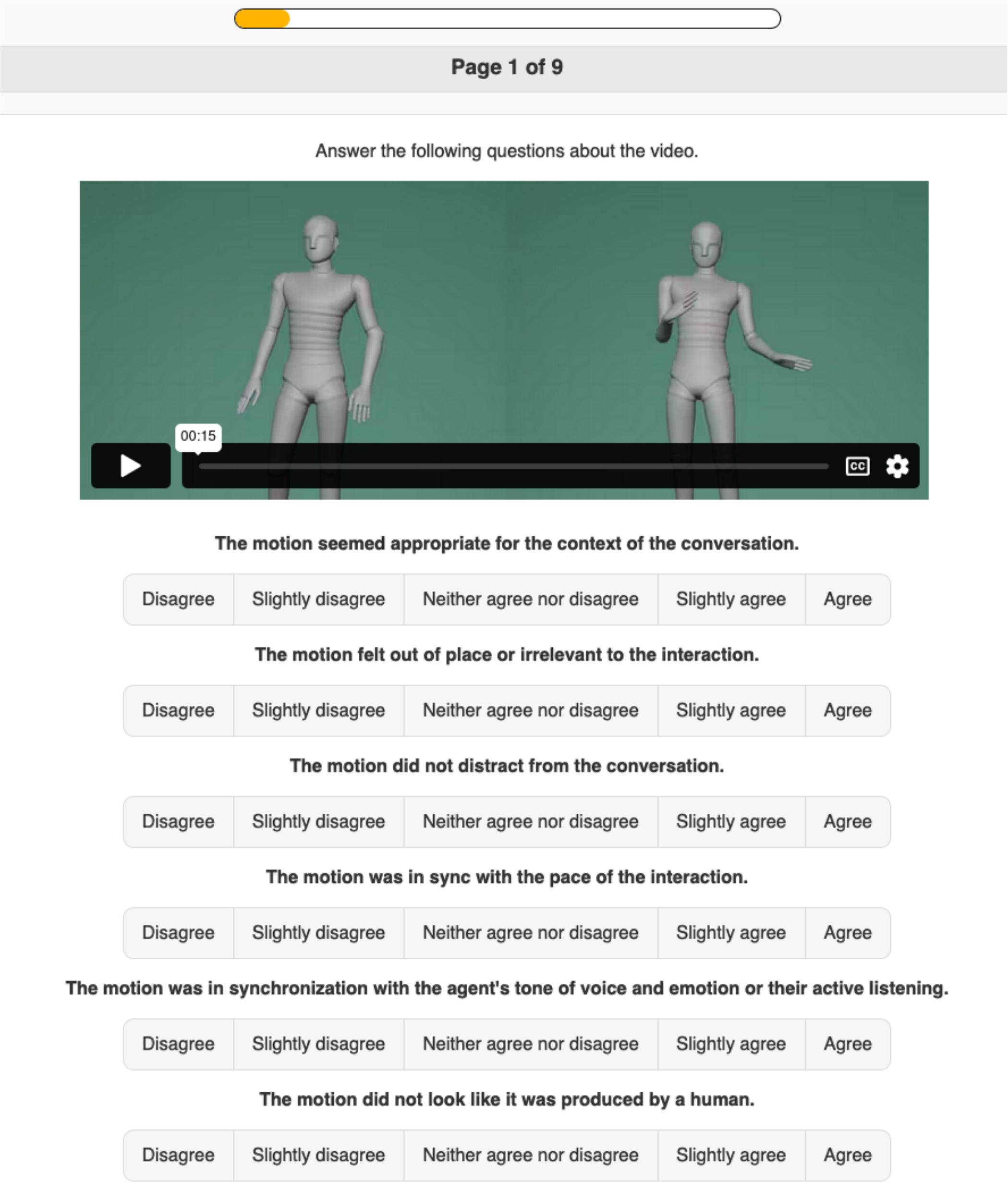
For this study, we examined the human-likeness of the generated listening behaviour and compared this to the baseline model and the ground truth motion. We recruited 22 participants. Thirty listening segments from our test set were selected, and the listening behaviour was synthesised with both models. The listening behaviour was then rendered on the avatar in a video. The audio was not included in the videos, as we wanted the participants to solely focus on the motion itself. The participants were asked the following question: 'How human-like does the listening motion appear?' Three videos were placed on one screen using the HEMVIP framework for evaluating the stimuli, similar to study 1 [18] (see Figure 2 for a screenshot of the interface). The order of the videos on the screen was randomised, as well as the order in which the screens were presented to the participant. Each participant encountered two attention checks inserted randomly during the experiment. Both attention checks would ask the participant to rate the video with a certain score. The text for the attention check would only appear halfway through the video (so that it would take out both participants that did not watch the complete video or that were not paying attention to the video at all). The participants were then asked to rate the human-likeness on a scale from 0 to 100. Each participant rated 14 screens with 3 stimuli per screen, totalling 42 ratings per participant and 308 ratings per condition.

4.4.5. Study 4: Appropriateness for Listening

For this study, we aimed to investigate the appropriateness of the listening motion for the conversation generated by the model. We recruited 27 participants. The set-up of this study followed the setup for study 2, but instead of selecting speaking segments, we selected segments where the main speaker was listening to the interlocutor. We took 30 segments and 30 additional segments as mismatching stimuli. These segments were then paired with the other speaker, resulting in two avatars being visible side-by-side in each video. See Figure 3 for a screenshot of the interface. Then, the listener, for which the motion was synthesised, was placed on the left. Both videos were placed on the same page, and the participants were asked to indicate in which of the two clips the character on the left moved appropriately for the speech. To establish an appropriateness baseline, we included matched and mismatched videos from the ground truth. We hypothesised that the participants would be able to identify the correct segments for full-body listening behaviour. Throughout the experiment, each participant was presented two attention checks inserted into random places during the experiment. One attention check was text-based, and the other one was audio-based. Halfway through the video, text would appear or an audio message could be heard, asking the participant to select a specific option in the interface. We used Barnard's test for identifying statistically significant differences between conditions at the level of $\alpha = 0.05$. Additionally, we applied the Holm–Bonferroni method to correct for multiple comparisons.

4.4.6. Studies 5 and 6: Questionnaire

We made use of the questionnaire as described in the section on questionnaire design for rating stimuli displaying both gesticulation and listening behaviour. Study 5 covered stimuli displaying generated gesture motion. Study 6 evaluated generated listening motion. We recruited 46 participants for study 5 and 48 participants for study 6. The participants were asked to rate their level of agreement with each statement on a scale with five answer options: disagree, slightly disagree, neutral, slightly agree, and agree. Each participant was presented 8 videos, with 1 per screen, and 1 attention check. The attention check consisted of one statement of the 15 on a page that asked the participant to select one answer option. Each video was accompanied by 15 statements on the page, for which the participant had to indicate his or her level of agreement. For this, we adapted the HEMVIP interface to display one video with 15 statements. Only upon answering all statements would the 'next' button be activated, and the interface is shown in Figure 4.



Page 1 of 9

Answer the following questions about the video.

The motion seemed appropriate for the context of the conversation.

Disagree | Slightly disagree | Neither agree nor disagree | Slightly agree | Agree

The motion felt out of place or irrelevant to the interaction.

Disagree | Slightly disagree | Neither agree nor disagree | Slightly agree | Agree

The motion did not distract from the conversation.

Disagree | Slightly disagree | Neither agree nor disagree | Slightly agree | Agree

The motion was in sync with the pace of the interaction.

Disagree | Slightly disagree | Neither agree nor disagree | Slightly agree | Agree

The motion was in synchronization with the agent's tone of voice and emotion or their active listening.

Disagree | Slightly disagree | Neither agree nor disagree | Slightly agree | Agree

The motion did not look like it was produced by a human.

Disagree | Slightly disagree | Neither agree nor disagree | Slightly agree | Agree

Figure 4. A screenshot of the interface used for the questionnaire. Participants were instructed that they were evaluating the motion for the left video. Each video was accompanied by 15 questions (not all visible in the image).

5. Results

5.1. User Studies

5.1.1. Study 1: 'Human-Likeness for Gesticulation'

In the study on the human-likeness of gesturing, the participants were asked to rate the stimuli for human-likeness on a scale from 1 to 100. The scores are visualized in Figure 5. Details on the demographics can be found in Table 2. All participants passed the attention checks.

We conducted Wilcoxon signed-rank tests to compare the similarity ratings between the SG, baseline, and ground truth conditions. The test results revealed that there was a significant difference in the similarity ratings between the SG and ground truth conditions (W = 6116.0, $p < 0.001$) as well as between the baseline and ground truth conditions (W = 6865.5, $p < 0.001$). However, there was no significant difference in the similarity ratings between the SG and baseline conditions (W = 20631.0, $p = 0.097$). These findings suggest that the SG system was able to produce gestures that were comparable to those produced in the baseline condition but not as similar as those produced in the ground truth condition.

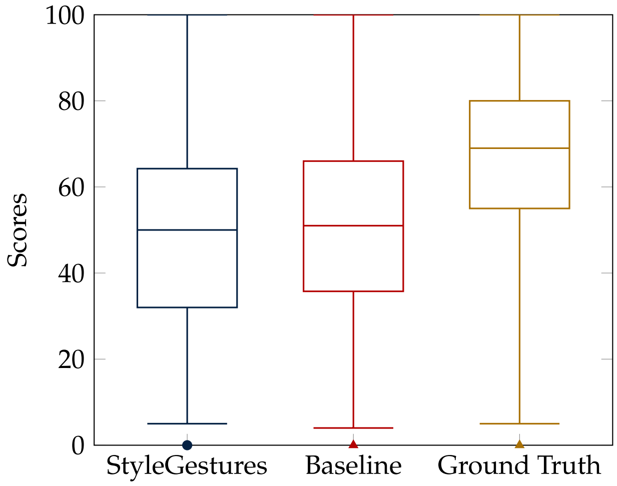


Figure 5. Boxplots of human-likeness scores on gesturing for StyleGestures (SG), baseline (BL), and ground truth (GT) conditions.

Table 2. Participant demographics for each study.

Study	N	Mean Age (SD)	Male	Female	Nationality	Education
1	22	35.2 (12.4)	16	6	UK (20), USA (1), IE (1)	P: 0, HS: 0, BS: 12, M: 1, D: 1, O: 8
2	27	40 (12.54)	14	13	UK (21), CA (2), IE (3), AU (1)	P: 0, HS: 0, BS: 14, M: 3, D: 0, O: 10
3	22	41.8 (13.94)	13	9	UK (16), USA (1), IE (2), NZ (2), CAN (1)	P: 0, HS: 0, BS: 10, M: 1, D: 0, O: 8
4	26	39 (11.55)	10	16	UK (15), CA (5), IE (3), AU (3)	P: 0, HS: 0, BS: 9, M: 7, D: 0, O: 10
5	46	42 (14.6)	22	24	UK (40), USA (1), AUS (1), CA (4)	P: 0, HS: 0, BS: 19, M: 7, D: 0, O: 22
6	48	37 (13)	27	21	UK (39), IE (3), CA (3), AU (3)	P: 0, HS: 0, BS: 20, M: 11, D: 0, O: 15

Abbreviations: P = primary school, HS = high school, BS = bachelors, M = masters, D = doctorate, and O = other.

5.1.2. Study 2: Appropriateness for Gesticulation

The purpose of this study was to investigate the ability of the participants to select the correct matching segment belonging to a conversation. For this, we used the match/mismatch paradigm initially proposed in [20] and later also used in [13]. The participants were presented with pairs of matching and mismatching videos and asked to choose which one featured the correct gesturing motion. They also had the option to choose whether the videos were equal. Details on the demographics can be found in Table 2.

Figure 6 shows the percentage of votes for matched, equal, and mismatched per condition. For SG, 62 videos were reported as matching, 56 were mismatched, and 65 were equal. For the baseline condition, 74 were reported as matching, 69 were mismatching, and 40 were equal. For the ground truth, 120 were matched, 30 were mismatched, and 24 were reported as equal.

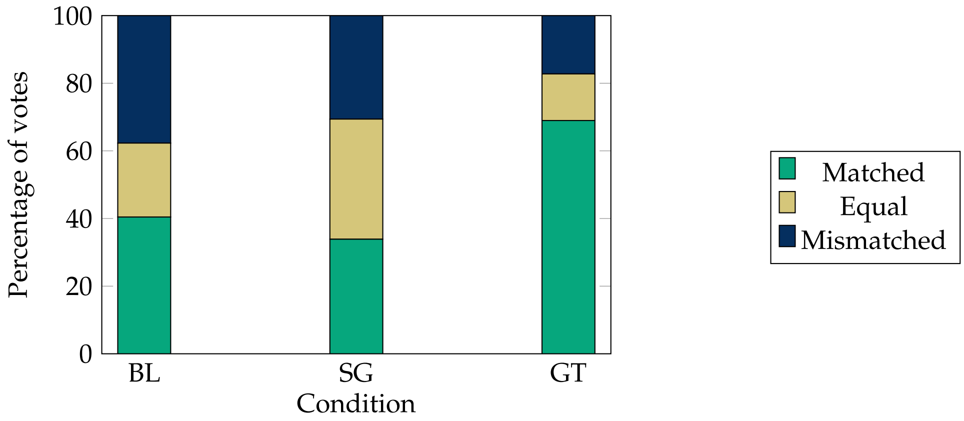


Figure 6. Stacked bar charts showing the percentage of votes on gesturing for StyleGestures (SG), baseline (BL), and ground truth conditions (GT) in study 2.

To analyse these results, chi-squared tests were conducted with Holm–Bonferroni correction applied for multiple comparisons. For SG, matching differed significantly from mismatching ($\chi^2 = 179, p < 0.0001$). For baseline, matching differed significantly from mismatching $\chi^2 = 179, p < 0.0001$, as well as the ground truth ($\chi^2 = 155, p < 0.0001$).

Lastly, we tested for differences between the conditions, in which ties were split equally over matching and mismatching. For SG versus the ground truth, there was a significant difference ($\chi^2 = 21.99, p < 0.0001$. For baseline versus the ground truth, there was a significant difference $\chi^2 = 21.99, p < 0.0001$. SG and baseline did not differ significantly.

5.1.3. Study 3: 'Human-Likeness for Listening'

The user study examined the human-likeness for listening behaviour of SG compared with a baseline and the ground truth (GT). The scores for each condition are visualised in Figure 7. Details on the demographics are provided in Table 2. All participants passed the attention checks.

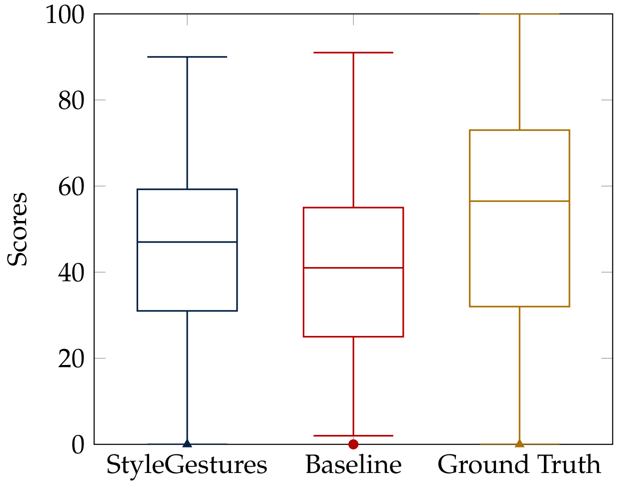


Figure 7. Boxplots of human-likeness scores for listening behaviour.

To analyse the data, Wilcoxon signed-rank tests were conducted between the SG and baseline, SG and GT, and baseline and GT conditions. The results showed that there was a statistically significant difference in the human-likeness perception between the SG and baseline conditions (Z = 16,265.5, $p < 0.0001$). This suggests that the SG system was perceived to be more human-like than the baseline system in relation to listening behaviour.

The results also showed a significant difference between the SG and GT conditions (Z = 16,506.0, $p < 0.0001$). This indicates that the participants perceived the SG system to be less human-like compared with the GT system, although the effect size was relatively small.

Lastly, there was a significant difference between the baseline and GT conditions (Z = 11,646.5, $p < 0.0001$). The participants perceived the baseline system to be less human-like than the GT one.

Overall, these results suggest that the SG system was perceived to be more human-like than the baseline system although less human-like compared with the ground truth.

5.1.4. Study 4: Appropriateness for Listening

The purpose of this study was to investigate the ability of the participants to select the correct, matching listening segment belonging to a conversation. For this, we used the match/mismatch paradigm initially proposed in [20] and later also used in [13]. The participants were presented with pairs of matching and mismatching videos and asked to choose which one featured the correct listening motion. They also had the option to choose whether the videos were equal. Details on the demographics can be found in Table 2. One participant was excluded as they did not pass the attention checks.

Figure 8 shows the percentage of votes for matched, equal, and mismatched per condition. For SG, 60 videos were reported as matching, 66 were mismatched, and 49 were equal. For the baseline condition, 73 were reported as matching, 44 were mismatching, and 58 were equal. For the ground truth, 86 were matched, 71 were mismatched, and 22 were reported as equal.

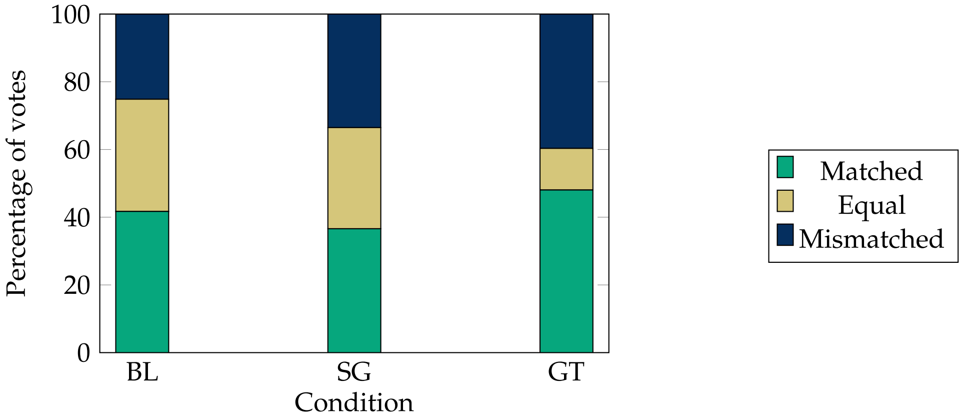


Figure 8. Stacked bar charts showing the percentage of votes on listening for baseline (BL), StyleGestures (SG), and ground truth (GT) in study 4.

For SG, matching differed significantly from mismatching ($\chi^2 = 149, p < 0.0001$). For the baseline, matching differed significantly from mismatching $\chi^2 = 170, p < 0.0001$, and the same held for the ground truth ($\chi^2 = 174, p < 0.0001$).

Lastly, we tested for differences between the conditions, where ties were split equally between matching and mismatching. There were no significant differences between the three conditions.

5.1.5. Study 5: Questionnaire for Gesturing

We examined the perceived quality of the synthesised gestures across three dimensions: appropriateness, human-likeness, and intelligibility. To assess the internal consistency of the rating scales used for these dimensions, we calculated the Cronbach's alpha coefficients. Details on the demographics can be found in Table 2. Two participants were excluded from the analysis as they did not pass the attention check.

For the appropriateness dimension, the Cronbach's alpha coefficient was 0.90 (95% CI [0.89, 0.92]), suggesting good internal consistency among the items assessing appropriateness. The human-likeness dimension had a Cronbach's alpha coefficient of 0.92 (95% CI

[0.91, 0.94]), indicating high internal consistency among the items measuring human-likeness. Furthermore, the intelligibility dimension exhibited excellent internal consistency, as indicated by a Cronbach's alpha coefficient of 0.97 (95% CI [0.97, 0.98]). This suggests a high degree of reliability among the items measuring intelligibility.

We performed a Mann–Whitney U test for each construct between each condition (StyleGestures vs. baseline, StyleGestures vs. ground truth, and baseline vs. ground truth). There were no significant differences between the scores for each construct. The mean scores are visualised in Figure 9.

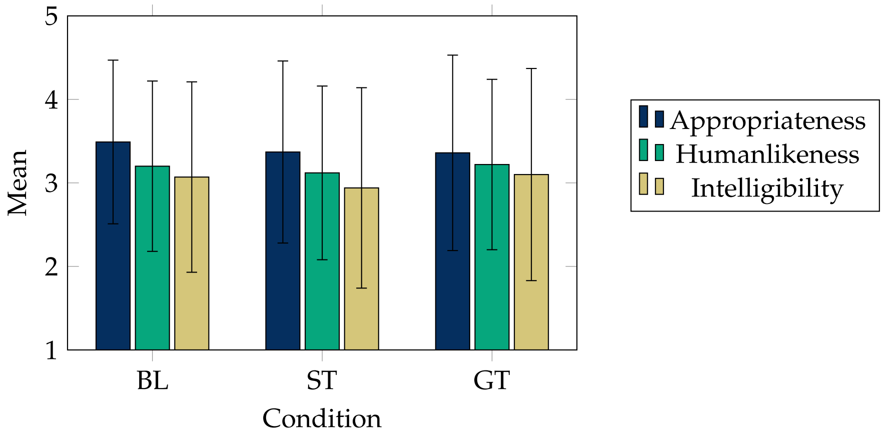


Figure 9. Mean and error bars for baseline (BL), StyleGestures (SG), and ground truth (GT) in study 5.

5.1.6. Study 6: Questionnaire for Listening

We examined the perceived quality of synthesised listening motion across three dimensions: appropriateness, human-likeness, and intelligibility. To assess the internal consistency of the rating scales used for these dimensions, we calculated the Cronbach's alpha coefficients. Details on the demographics can be found in Table 2. One participant was excluded as they did not pass the attention checks.

For the appropriateness dimension, the Cronbach's alpha coefficient was found to be 0.90 (95% CI [0.88, 0.91]), indicating good internal consistency and agreement among the items assessing appropriateness. For the human-likeness, the Cronbach's alpha coefficient was 0.93 (95% CI [0.92, 0.94]). For the intelligibility construct, the Cronbach's alpha coefficient was found to be 0.98 (95% CI [0.98, 0.98]).

We performed a Mann–Whitney U test for each construct between each condition (StyleGestures vs. baseline, StyleGestures vs. ground truth, and baseline vs. ground truth). There were no significant differences between the scores for each construct. The mean scores are visualised in Figure 10.

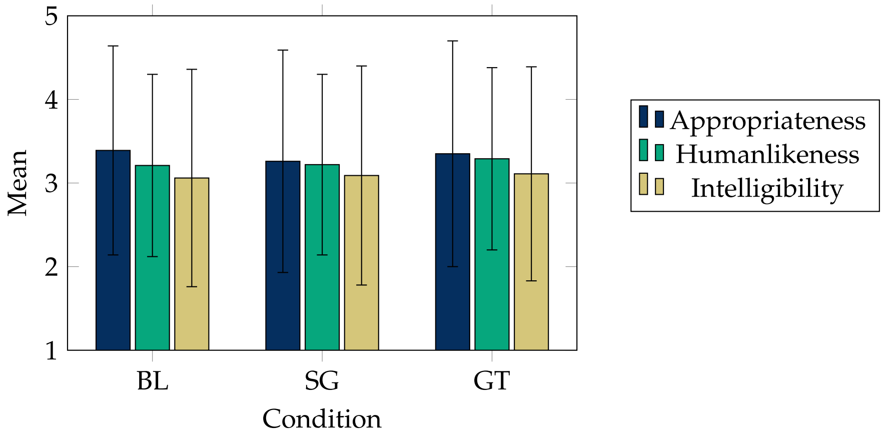


Figure 10. Mean and error bars for baseline (BL), StyleGestures (SG), and ground truth (GT) in study 6.

5.1.7. Completion Times

To assess the efficiency of test takers using different evaluation methodologies, we conducted an analysis of the test completion times and report on the mean completion time per page per study for the participants. Study 1 on human-likeness took 60 s (SD = 80), with 20 s per stimulus. Study 2 on appropriateness took 44 s (SD = 107), with 22 s per stimulus. Study 3 took 60 s (SD = 29), with 20 s per stimulus. Study 4 took 30 s (SD = 45), with 15 s per stimulus. The last two studies featured one stimulus per page. Study 5 with the questionnaire took 52 s to complete on average (SD = 34). For study 6, the average was 55 s (SD = 40).

6. Discussion

For this paper, we trained two models on a data set containing dyadic conversations. We used the synthesised stimuli in six subjective studies. Two of the three evaluation paradigms are existing evaluation strategies from the field of co-speech gesture generation. We compared these two to a newly made questionnaire based on the usage of Likert scales and constructs in a previous work on gesture generation [1].

In study one, the human-likeness for gesturing was evaluated. No significant difference was found between StyleGestures (SG) and the baseline. However, the ground truth scored significantly higher on average, and both SG and the baseline differed significantly from the ground truth. The finding that SG and the baseline did not differ significantly is interesting since the baseline model incorporates semantic information. Human-likeness evaluations focus on motion quality rather than appropriateness of gestures with speech audio. Previous works on gesture generation evaluation yielded more robust differences using this methodology [13]. Interestingly, the ground truth condition scored 70 on average, even though the stimuli in this condition were based on human recorded motion. This has been found previously in other works as well [13,17] and is most likely because of the different embodiment of the avatar in relation to that of humans, as well as the lack of other human-like features (such as a face).

Another angle of evaluating the appropriateness of gesticulation can be conducted through the use of matching and mismatching videos [20]. In study 2, we took the same 30 segments and combined them into one video with the interlocutor. In one of the two videos presented to the participant, the gesture sequence of the avatar on the left was not related to that part of the conversation. We found significant differences for both the baseline and SG with the ground truth condition but not for the baseline versus SG. When we look at Figure 6, we see that more videos were identified as ‘equal’ for the SG condition. As expected, the ground truth videos were identified as matching more than 70% of the time. Since the baseline model also had access to the text beside the speech audio, one would expect this model to generate more appropriate (and even semantic-related) gestures, but the results from the appropriateness study do not seem to confirm this.

We wanted to know whether these two paradigms of human-likeness and appropriateness testing could be used with more subtle forms of human nonverbal behaviour, such as listening behaviour. In the third study, we evaluated the human-likeness of the generated listening behaviour and compared it to the ground truth. It is important to mention that for the human-likeness evaluation, we excluded the audio to only assess the quality of the motion. We found significant differences between all conditions, with SG scoring past the ground truth. However, the overall rating for each condition was not particularly high. We think that human-likeness testing for motion for listening behaviour is difficult since appropriate listening behaviour is really dependent on the conversation. Omitting the audio could also have led to the participants not being able to see that this motion was supposed to be part of a conversation. Another reason could be that in terms of motion quality, it was all similar and therefore scored the same because of the lack of context.

Appropriateness testing of listening behaviour could help figure out whether it actually matters what listening behaviour is tied to a conversation and whether participants can spot differences in generated listening behaviour. Here, we cannot report any significant difference between the three conditions. It appears that the participants had difficulty identifying the right listening behaviour. One reason for that could be that listening behaviour takes place more with facial expressions than with body language, and that the body pose alone is not enough to say that someone is attending a conversation. Another reason could be that the avatar visualisation was too far from human-likeness, and therefore the participants had a harder time believing that it was a human that was moving. Listening behaviour is not only dependent on full-body motion, but it is also often combined with verbal feedback [38].

For our last two studies, we designed a questionnaire with appropriateness, human-likeness, and intelligibility as constructs. We based our choice of statements and constructs on earlier work, reported in [1]. The internal consistency, measured through the Cronbach's alpha coefficient, was high. For each construct for both studies, the Cronbach's alpha coefficient was equal to or greater than 0.9. This provides an indication that together, the statements measured the intended construct. However, there were no significant differences in scores between the systems for either the gesticulation or listening behaviour evaluations. We can see small differences when we look at the figures, but these differences are not statistically significant. The evaluation of nonverbal behaviour poses a challenging task, as evidenced by the multitude of diverse evaluation paradigms employed over the years [1]. Unfortunately, there is a lack of a standardised and unified approach to measuring nonverbal behaviour, further complicating the evaluation process. In a recent paper by He et al. [14], multiple measuring methods were applied to test the gesticulation of an avatar in an interaction, and only the behavioural method (through gaze tracking) yielded significant differences. Here, a construct from the Godspeed questionnaire [15] was also included. Because of the high internal consistency, one could argue that multiple statements were measuring the same thing, reducing the resolution of the questionnaire, or that the Godspeed questionnaire is not a good questionnaire for evaluating human-like motion. On top of that, we only provided five answer options, which is a common way of applying Likert rating scales [49]. A possible explanation for the non-significant results is that the five answer options were not enough, and it is not sufficient for picking up these small differences in generated nonverbal behaviour. The number of statements the participants had to answer per video could also have led to fatigue in the participants, even with the low number of videos presented to the participants.

We ran a small analysis to look at the completion times for evaluating stimuli per evaluation method. The time per stimulus in the test appeared to be the lowest for appropriateness testing. However, in earlier research, it has been shown that pairwise comparisons might be faster but scale worse when comparing multiple conditions [42]. The questionnaire method, with 15 statements per video, took the longest to complete. Here, the participants needed almost one minute per video. In comparison with the human-likeness methodology, this took three times longer. We therefore can conclude that when looking at time efficiency, the HEMVIP methodology is the most efficient method for gathering ratings per stimulus.

The current uptick in the application of human-likeness testing using direct question and appropriateness testing, with matching and mismatching stimuli for generated nonverbal behaviour, seems to yield interesting results and makes comparisons easy when researchers include at least one condition with motion from previous work. Thanks to researchers working on the GENEA Challenge [13,17], more and more code has become available for running objective and subjective evaluations and comparative studies. However, a standardised method of measuring intelligibility, which was the only construct not compared to previous evaluation methodologies, is missing in the field at the moment. This could be asked through a direct question or by presenting the statements we came up with. As previous research mentioned before, subjective evaluations through an interaction could be performed to study intelligibility [1,12,31], but this has to be carried out carefully and does not always lead to statistically significant differences [14].

7. Conclusions

In this work, we compared three different methods for subjectively evaluating computer-generated gestures, all of which are ideal for studies run through crowd-sourcing platforms. Two methods rely on direct rating of the motion quality, and these were compared to a new questionnaire. We found that for gesticulation behaviour, both human-likeness and appropriateness testing yielded results that made it possible to compare and rank the quality of generative models. However, when it came to listening behaviour, the differences were less clear between different systems and the ground truth motion. It is our advice to stick to direct evaluations of human-likeness, preferably using the HEMVIP framework. Regarding the appropriateness, we think that the current methodology of matching stimuli yields good results, but we are wary of its scaling when assessing multiple conditions. Our work offers valuable insights for researchers in gesture generation and the evaluation of generated behaviour. We encourage researchers to leverage existing methodologies and incorporate previously evaluated systems or generated motion. This will facilitate meaningful comparisons among works from various researchers. For future work, we suggest evaluating the appropriateness testing methodology against a direct approach to measure appropriateness rather than relying solely on matching and mismatching stimuli. Additionally, it could be interesting to explore different ways of incorporating questionnaire constructs.

Author Contributions: Conceptualisation, P.W., G.E.H. and T.B.; investigation, P.W.; methodology, P.W.; software, P.W.; supervision, G.E.H. and T.B.; validation, P.W.; visualisation, P.W.; writing—original draft, P.W. and G.E.H.; writing—review and editing, P.W., G.E.H. and T.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Flemish Research Foundation grant no. 1S95020N.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of the Faculty of Psychology and Pedagogical Sciences of Ghent University (reference 2020/92).

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

Data Availability Statement: The data presented in this study are openly available in Zenodo at https://zenodo.org/doi/10.5281/zenodo.10641146.

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

References

1. Wolfert, P.; Robinson, N.; Belpaeme, T. A Review of Evaluation Practices of Gesture Generation in Embodied Conversational Agents. *IEEE Trans. Hum.-Mach. Syst.* **2022**, *52*, 379–389. [CrossRef]
2. Knapp, M.L.; Hall, J.A.; Horgan, T.G. *Nonverbal Communication in Human Interaction*; Cengage Learning: Boston, MA, USA, 2013.
3. McNeill, D. *Hand and Mind: What Gestures Reveal about Thought*; University of Chicago Press: Chicago, IL, USA, 1992; Volume 351.
4. Holler, J.; Kendrick, K.H.; Levinson, S.C. Processing language in face-to-face conversation: Questions with gestures get faster responses. *Psychon. Bull. Rev.* **2018**, *25*, 1900–1908. [CrossRef] [PubMed]
5. Chidambaram, V.; Chiang, Y.H.; Mutlu, B. Designing persuasive robots: How robots might persuade people using vocal and nonverbal cues. In Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction, Boston, MA, USA, 5–8 March 2012; pp. 293–300.
6. Ham, J.; Cuijpers, R.H.; Cabibihan, J.J. Combining robotic persuasive strategies: The persuasive power of a storytelling robot that uses gazing and gestures. *Int. J. Soc. Robot.* **2015**, *7*, 479–487. [CrossRef]
7. Salem, M.; Eyssel, F.; Rohlfing, K.; Kopp, S.; Joublin, F. To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *Int. J. Soc. Robot.* **2013**, *5*, 313–323. [CrossRef]
8. Alexanderson, S.; Henter, G.E.; Kucherenko, T.; Beskow, J. Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows. In *Proceedings of the Computer Graphics Forum*; Wiley Online Library: Hoboken, NJ, USA, 2020; Volume 39, pp. 487–496.
9. Ahuja, C.; Lee, D.W.; Morency, L.P. Low-resource adaptation for personalized co-speech gesture generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 20566–20576.
10. Alexanderson, S.; Nagy, R.; Beskow, J.; Henter, G.E. Listen, Denoise, Action! Audio-Driven Motion Synthesis with Diffusion Models. *ACM Trans. Graph.* **2023**, *42*, 44. [CrossRef]
11. Osorio, P.; Sagawa, R.; Abe, N.; Venture, G. A Generative Model to Embed Human Expressivity into Robot Motions. *Sensors* **2024**, *24*, 569. [CrossRef] [PubMed]

12. Huang, C.M.; Mutlu, B. Modeling and Evaluating Narrative Gestures for Humanlike Robots. In *Proceedings of the Robotics: Science and Systems*; Citeseer: State College, PA, USA, 2013; Volume 2.
13. Yoon, Y.; Wolfert, P.; Kucherenko, T.; Viegas, C.; Nikolov, T.; Tsakov, M.; Henter, G.E. The GENEA Challenge 2022: A large evaluation of data-driven co-speech gesture generation. In Proceedings of the 2022 International Conference on Multimodal Interaction, Bengaluru, India, 7–11 November 2022; pp. 736–747.
14. He, Y.; Pereira, A.; Kucherenko, T. Evaluating data-driven co-speech gestures of embodied conversational agents through real-time interaction. In Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents, Faro, Portugal, 6–9 September 2022; pp. 1–8.
15. Bartneck, C.; Kulić, D.; Croft, E.; Zoghbi, S. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int. J. Soc. Robot.* **2009**, *1*, 71–81. [CrossRef]
16. Wolfert, P.; Henter, G.E.; Belpaeme, T. "Am I listening?", Evaluating the Quality of Generated Data-driven Listening Motion. In Proceedings of the Companion Publication of the 25th International Conference on Multimodal Interaction, Paris, France, 9–13 October 2023; pp. 6–10.
17. Kucherenko, T.; Jonell, P.; Yoon, Y.; Wolfert, P.; Henter, G.E. A large, crowdsourced evaluation of gesture generation systems on common data: The GENEA Challenge 2020. In Proceedings of the 26th International Conference on Intelligent User Interfaces, College Station, TX, USA, 13–17 April 2021; pp. 11–21.
18. Jonell, P.; Yoon, Y.; Wolfert, P.; Kucherenko, T.; Henter, G.E. HEMVIP: Human Evaluation of Multiple Videos in Parallel. In Proceedings of the 2021 International Conference on Multimodal Interaction, Montreal, QC, Canada, 18–22 October 2021; pp. 707–711.
19. Jonell, P.; Kucherenko, T.; Henter, G.E.; Beskow, J. Let's face it: Probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings. In Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents, Online, 20–22 October 2020; pp. 1–8.
20. Rebol, M.; Güti, C.; Pietroszek, K. Passing a non-verbal turing test: Evaluating gesture animations generated from speech. In Proceedings of the 2021 IEEE Virtual Reality and 3D User Interfaces (VR), Lisboa, Portugal, 27 March–1 April 2021; pp. 573–581.
21. Kucherenko, T.; Nagy, R.; Yoon, Y.; Woo, J.; Nikolov, T.; Tsakov, M.; Henter, G.E. The GENEA Challenge 2023: A large-scale evaluation of gesture generation models in monadic and dyadic settings. In Proceedings of the 25th International Conference on Multimodal Interaction, Paris, France, 9–13 October 2023; pp. 792–801.
22. Kucherenko, T.; Hasegawa, D.; Henter, G.E.; Kaneko, N.; Kjellström, H. Analyzing input and output representations for speech-driven gesture generation. In Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, Paris, France, 2–5 July 2019; pp. 97–104.
23. Yoon, Y.; Ko, W.R.; Jang, M.; Lee, J.; Kim, J.; Lee, G. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 4303–4309.
24. Yoon, Y.; Cha, B.; Lee, J.H.; Jang, M.; Lee, J.; Kim, J.; Lee, G. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Trans. Graph. (TOG)* **2020**, *39*, 222. [CrossRef]
25. Kucherenko, T.; Jonell, P.; Van Waveren, S.; Henter, G.E.; Alexandersson, S.; Leite, I.; Kjellström, H. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In Proceedings of the 2020 International Conference on Multimodal Interaction, Online, 25–29 October 2020; pp. 242–250.
26. Ahuja, C.; Lee, D.W.; Ishii, R.; Morency, L.P. No gestures left behind: Learning relationships between spoken language and freeform gestures. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 16–20 November 2020; pp. 1884–1895.
27. Ahuja, C.; Ma, S.; Morency, L.P.; Sheikh, Y. To react or not to react: End-to-end visual pose forecasting for personalized avatar during dyadic conversations. In Proceedings of the 2019 International Conference on Multimodal Interaction, Suzhou, China, 14–18 October 2019; pp. 74–84.
28. Tuyen, N.T.V.; Celiktutan, O. Agree or Disagree Generating Body Gestures from Affective Contextual Cues during Dyadic Interactions. In Proceedings of the 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), Napoli, Italy, 29 August–2 September 2022; pp. 1542–1547.
29. Ao, T.; Zhang, Z.; Liu, L. GestureDiffuCLIP: Gesture diffusion model with CLIP latents. *arXiv* **2023**, arXiv:2303.14613.
30. Mehta, S.; Wang, S.; Alexanderson, S.; Beskow, J.; Székely, É.; Henter, G.E. Diff-TTSG: Denoising probabilistic integrated speech and gesture synthesis. *arXiv* **2023**, arXiv:2306.09417.
31. Nyatsanga, S.; Kucherenko, T.; Ahuja, C.; Henter, G.E.; Neff, M. A Comprehensive Review of Data-Driven Co-Speech Gesture Generation. *arXiv* **2023**, arXiv:2301.05339.
32. Heylen, D.; Bevacqua, E.; Pelachaud, C.; Poggi, I.; Gratch, J.; Schröder, M. Generating listening behaviour. In *Emotion-Oriented Systems: The Humaine Handbook*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 321–347.
33. Buschmeier, H.; Kopp, S. Communicative listener feedback in human-agent interaction: Artificial speakers need to be attentive and adaptive. In Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems, Stockholm, Sweden, 10–15 July 2018; pp. 1213–1221.

34. Maatman, R.; Gratch, J.; Marsella, S. Natural behavior of a listening agent. In Proceedings of the Intelligent Virtual Agents: 5th International Working Conference, IVA 2005, Kos, Greece, 12–14 September 2005; Proceedings 5; Springer: Berlin/Heidelberg, Germany, 2005; pp. 25–36.
35. Gillies, M.; Pan, X.; Slater, M.; Shawe-Taylor, J. Responsive Listening Behavior. *Comput. Animat. Virtual Worlds* **2008**, *19*, 579–589. [CrossRef]
36. Mlakar, I.; Kačič, Z.; Rojc, M. Describing and animating complex communicative verbal and nonverbal behavior using Eva-framework. *Appl. Artif. Intell.* **2014**, *28*, 470–503. [CrossRef]
37. Poppe, R.; Truong, K.P.; Reidsma, D.; Heylen, D. Backchannel strategies for artificial listeners. In Proceedings of the Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, 20–22 September 2010; Proceedings 10; Springer: Berlin/Heidelberg, Germany, 2010; pp. 146–158.
38. Antonio Gomez Jauregui, D.; Giraud, T.; Isableu, B.; Martin, J.C. Design and evaluation of postural interactions between users and a listening virtual agent during a simulated job interview. *Comput. Animat. Virtual Worlds* **2021**, *32*, e2029. [CrossRef]
39. Weiss, A.; Bartneck, C. Meta analysis of the usage of the godspeed questionnaire series. In Proceedings of the 2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), Kobe, Japan, 31 August–4 September 2015; pp. 381–388.
40. Fitrianie, S.; Bruijnes, M.; Richards, D.; Abdulrahman, A.; Brinkman, W.P. 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, Paris, France, 2–5 July 2019; pp. 159–161.
41. Fitrianie, S.; Bruijnes, M.; Richards, D.; Bönsch, A.; Brinkman, W.P. 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, Online, 20–22 October 2020; pp. 1–8.
42. Wolfert, P.; Girard, J.M.; Kucherenko, T.; Belpaeme, T. To rate or not to rate: Investigating evaluation methods for generated co-speech gestures. In Proceedings of the 2021 International Conference on Multimodal Interaction, Montreal, QC, Canada, 18–22 October 2021; pp. 494–502.
43. Grassia, F.S. Practical parameterization of rotations using the exponential map. *J. Graph. Tools* **1998**, *3*, 29–48. [CrossRef]
44. Papamakarios, G.; Nalisnick, E.; Rezende, D.J.; Mohamed, S.; Lakshminarayanan, B. Normalizing flows for probabilistic modeling and inference. *J. Mach. Learn. Res.* **2021**, *22*, 2617–2680.
45. Henter, G.E.; Alexanderson, S.; Beskow, J. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Trans. Graph. (TOG)* **2020**, *39*, 236. [CrossRef]
46. Chang, C.J.; Zhang, S.; Kapadia, M. The IVI Lab entry to the GENEA Challenge 2022–A Tacotron2 based method for co-speech gesture generation with locality-constraint attention mechanism. In Proceedings of the 2022 International Conference on Multimodal Interaction, Bengaluru, India, 7–11 November 2022; pp. 784–789.
47. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In Proceedings of the 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4779–4783.
48. Schoeffler, M.; Bartoschek, S.; Stöter, F.R.; Roess, M.; Westphal, S.; Edler, B.; Herre, J. webMUSHRA—A comprehensive framework for web-based listening tests. *J. Open Res. Softw.* **2018**, *6*, 8. [CrossRef]
49. Schrum, M.L.; Johnson, M.; Ghuy, M.; Gombolay, M.C. Four years in review: Statistical practices of likert scales in human-robot interaction studies. In Proceedings of the Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, Cambridge, UK, 23–26 March 2020; pp. 43–52.

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