"Am I listening?", Evaluating the Quality of Generated Data-driven Listening Motion

Pieter Wolfert
pieter.wolfert@ugent.be
IDLab-AIRO - Ghent University
Ghent, Belgium

Gustav Eje Henter
Speech, Music and Hearing KTH
Royal Insitute of Technology
Stockholm, Sweden
ghe@kth.se

Tony Belpaeme
tony.belpaeme@ugent.be
IDLab-AIRO - Ghent University
Ghent, Belgium

ABSTRACT

This paper asks if recent models for generating co-speech gesticulation also may learn to exhibit listening behaviour as well. We consider two models from recent gesture-generation challenges and train them on a dataset of audio and 3D motion capture from dyadic conversations. One model is driven by information from both sides of the conversation, whereas the other only uses the character's own speech. Several user studies are performed to assess the motion generated when the character is speaking actively, versus when the character is the listener in the conversation. We find that participants are reliably able to discern motion associated with listening, whether from motion capture or generated by the models. Both models are thus able to produce distinctive listening behaviour, even though only one model is truly a listener, in the sense that it has access to information from the other party in the conversation. Additional experiments on both natural and model-generated motion finds motion associated with listening to be rated as less human-like than motion associated with active speaking.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**.

KEYWORDS

listening behaviour, embodied conversational agents

ACM Reference Format:
Pieter Wolfert, Gustav Eje Henter, and Tony Belpaeme. 2023. "Am I listening?", Evaluating the Quality of Generated Data-driven Listening Motion. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23 Companion), October 9–13, 2023, Paris, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3610661.3617160

1 INTRODUCTION

During face-to-face natural language interaction, there is always one speaker but there is also at least one listener. Data-driven methods for generating speech behaviour have received significant attention, while comparatively less emphasis has been placed on data-driven methods for generating listening behaviour. Active listening, in which the listener signals attention, engagement and a willingness to be part of the interaction, is equally important to a successful conversation [6]. It is important that both the speaker and the listener are on the same wavelength, and understand each others (non)verbal cues, since this serves as 'social glue' [17]. It has been known for a while that nonverbal synchrony and mimicry play a key role in one-one interactions [13]. Some efforts have been made to incorporate nonverbal facial back channelling in virtual avatars, to improve human-agent interactions. One example of this is the work by Jonell et al. [11], where facial nonverbal feedback is generated based on data from human dyadic interactions. However, fewer attempts have focused on generating listening behaviour using a data driven approach, and evaluating that aspect of nonverbal dyadic interaction.

In the current study, we want to focus on the abilities of motion generating models to generate listening motion. For this, we adapted an existing generative model named StyleGestures, to deal with dyadic conversational data [2]. We were also interested in the performance of an existing model on listening behaviour generation. For this we picked a baseline model that was submitted to the GENEA Challenge 2022, and received the best reproducability award. The baseline model has already been compared to other submissions in the challenge, that focused on co-speech gesture generation.

2 RELATED WORK

2.1 Gesturing

Many recent studies have focused on generating speech motion for embodied conversational agents (ECAs). For instance, Kucherenko et al. [14] leveraged representation learning to map audio to motion, while Yoon et al. [30] used input text to generate motion while ignoring the audio input channel. Subsequently, other researchers have combined both audio and text input, along with speaker identity, in their gesture generation models, such as [15, 29]. However, since the goal of gesture generation for ECAs is to help facilitate effective human-agent interaction, some researchers have explored generating nonverbal behaviour while considering the interlocutor [1, 26]. A more in depth review on the field of gesture generation, especially considering deep learning, can be found in [21]. Despite these contributions, comparing different models is challenging, as highlighted by Wolfert et al. [28]. Recently, the GENEA Challenge [31] has been proposed to address this issue, allowing multiple teams to submit their model's motion for a shared evaluation.

ICMI '23 Companion, October 9–13, 2023, Paris, France

ACM ISBN 979-8-4007-0321-8/23/10…$15.00
https://doi.org/10.1145/3610661.3617160

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 [10]. For instance, Buschmeier et al. [4] 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 [19] proposed a model that generates listening behaviour based on available features during a conversation. Their system extracts audio and body posture features to drive the listening behaviour. Another approach by Gillies et al. [7] utilised input audio from the speaker to generate listening behaviour through motion graphs, where existing motion clips are combined to match new audio input. Mlakar [20] introduced a framework and scripting method to synthesise both verbal and nonverbal motion, that entails both gestures and listening. Poppe et al. [22] 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 [3]. They used a multi-modal corpus of interviews to generate listening behaviour in a virtual agent conducting interviews. 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. [11]. 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.

3 METHODS

3.1 Dataset and Preprocessing

To ensure that the StyleGestures (SG) model is applicable to a wider range of conversational interactions, we opted to train it on a data set that includes human dyadic interactions, rather than just a single speaker. "Talking With Hands 16.2," provides a rich source of dyadic conversational data [18]. This data set includes both motion capture and audio, totalling 50 hours of recorded interactions. For the baseline model we made use of annotations provided by the GENEA Challenge 2022 [31]. We opted for only including 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. This resulted in subset with 10 hours of interactions. 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 [2]. The resulting features were down-sampled 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 prevents discontinuities [8], and full-body motion was used excluding finger and facial information.

3.2 Models

For this work, we make use of two models that were originally designed to learn gesture behaviour from human data. SG is taken for its generative capabilities, and we adapted it to work with dyadic conversational data. As we aim for a fair comparison in relation to the ground-truth data, we also trained another model named 'baseline'.

3.2.1 StyleGestures. StyleGestures is a probabilistic generative sequential model based on MoGlow, which uses normalizing flows [2]. The model was modified to accept dyadic input, with the input being a concatenation of two audio streams (speaker and interlocutor), a one hot encoding of the speaker identity, and the motion stream of the interlocutor. The output of the model is joint rotations for the speaker. The modified SG model was trained using the standard parameters from the SG paper [2], with batch size 120, noam_learning_rate_decay with 3000 warm up steps, and a minimum of 0.00015. The optimiser used was Adam, with a learning rate of 0.0015. Due to the size of the input data, the model was trained for 160k steps before test motion was generated. We applied post-processing to the motion data to improve the quality of our generated listening behaviour. Specifically, we used a Butter worth low pass filter to smooth the rotation data. The cutoff frequency was set to 3.0 Hz and the filter order was set to 4.

3.2.2 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 "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 [5]. The baseline model is based on the Tacotron2 architecture from speech synthesis with a locality constraint attention mechanism, and takes text and speech audio as input to generate motion data [25]. It was trained on only the text and speech input data from the speaker whose motion we are predicting, namely speaker 1 (in contrast to our SG model that was trained on full dyadic data). For the training parameters we relied on the values used by [5].

3.3 Visualisation

We rendered the generated motion on a faceless avatar (see figure 1), that was provided by the GENEA Challenge 2022. The hands were fixed for all positions since we did not learn the finger positions.

3.4 User Studies

For the first study we relied on pairwise comparisons, where we mismatched listening behaviour segments with speech behaviour segments. This mismatching paradigm has previously been used by [11, 23, 31]. For the human-likeness evaluations of the listening and speech behaviour, we used an existing evaluation methodology named HEMVIP [12] that is based on the ITU MUSHRA standard for audio quality evaluation [24]. HEMVIP utilises a sliding scale with 100 steps and five anchors (bad, poor, fair, good, excellent), and has been used by various researchers for evaluating generated behaviour [9, 16, 27, 31]. All participants were recruited on Prolific[1].

[1] https://prolific.co

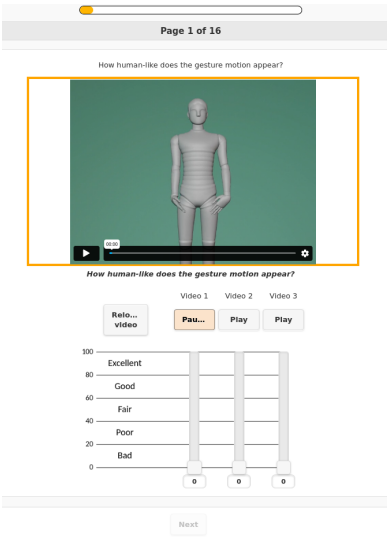

Figure 1: The HEMVIP system with the human-likeness evaluation showing the avatar we utilised [12].

3.4.1 Study 1: "Does it listen?" The purpose of this study was to investigate the ability of participants to identify generated listening segments when presented with unrelated speech motion fragments. We recruited 32 participants who were required to be native English speakers. Listening segments were generated using either the baseline model or the SG model. To determine whether participants were able to distinguish matching listening motion versus mismatching speech motion, speech motion segments were obtained from the ground-truth. Each matching or mismatching segment was then added to a video containing a speaker, who was positioned on the left of the video with the listener on the right. Audio for each conversation was added to the video. We selected 30 listening segments per condition, totalling 60 segments. The videos containing the conversations (matching versus mismatching) were presented side by side in a random order, and the order of presentation was also randomised. Participants were asked the question: "Please indicate in which of the two clips the character on the right moves like a listening person." Each participant was presented two attention checks, inserted at random points during the experiment. One check was text based and the other one audio based, halfway the video it would ask the 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 $\alpha = 0.05$. Additionaly, the Holm-Bonferroni method was applied to correct for multiple comparisons.

3.4.2 Study 2: "Human-likeness for listening". This study investigated the human-likeness of the generated listening behaviour. For this, we compared it to the baseline and ground-truth motion. We recruited 22 participants who were required to be native English speakers. From the test set, 30 listening segments were selected, and listening motion was synthesised from SG and the baseline, or taken from the ground-truth. The videos did not feature audio, as we wanted participants to specifically focus on the motion. Participants were asked the following question: *"How human-like does the listening motion appear?"*, and had to rate the videos on a scale from 0 to 100. Three videos were placed on one screen, using the HEMVIP framework for evaluating the stimuli [12](see figure 1). 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 checks, inserted at random places 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 the video. Each participant rated 14 screens with 3 stimuli per screen, totalling 42 ratings per participant and 308 ratings per condition.

3.4.3 Study 3: "Human-likeness for gesticulation". For this study we followed the approach of study 2. In this study we investigated the human-likeness of the model generated speaking behaviour. For this, we compared it to the baseline and the ground-truth motion. We recruited 22 participants who were required to be native English speakers. As in study 2, 30 segments were selected where the avatar was talking. Participants were asked the following question: *"How human-like does the gesture motion appear?"* Each participant rated 14 screens with 3 stimuli per screen, totalling 42 ratings per participant and 308 ratings per condition.

3.5 Objective Analysis

As pointed out often before, there is no single objective metric that can capture the quality of the generated motion. Therefore, we rely on commonly used and reported objective metrics in the field such as the acceleration and jerk.

4 RESULTS

4.1 User Studies

4.1.1 Study 1: "Does it listen?" In this study, we looked at matching versus mismatching for listening behaviour, where the mismatched video used unrelated speech motion. [2] [3] Participants were presented with pairs of matching/mismatching videos and asked to choose which one featured the listening motion. They also had the option to choose that the videos were equal. 32 participants were recruited, of which 30 passed the attention checks. Of these, the mean age was 40.6 years (SD=11.5). 15 identified as female and the other 15 identified as male. 28 participants were from the UK, 1 from the USA, and 1 from New Zealand.

For SG 46 (16%) videos were reported as equal, 178 (61%) as matching and 69 (23%) as mismatched. For the baseline this was 25 (8%) reported as equal, 215 (72%) as matching and 57 (20%) as mismatching. We further performed Barnard's test with Holm-Bonferroni correction to analyse the data. In the SG condition, we found a significant difference between matched and mismatched videos (Chi2 stat: 69.0, p-value: < 0.001). Similarly, in the baseline condition, there was a significant difference between matched and mismatched videos (Chi2 stat: 57.0, p-value < 0.001). These results suggest that participants were able to perceive which video of a pair featured the listening behaviour.

[2] SG listening on the right: https://player.vimeo.com/video/820607290?h=586ce22531
[3] BL listening on the right: https://player.vimeo.com/video/820607372?h=0e0fc479a9

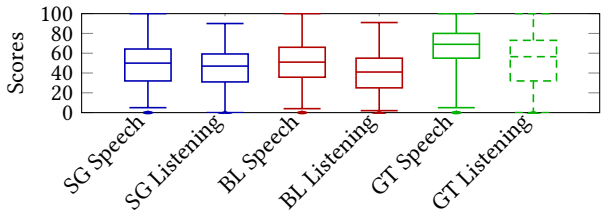

Figure 2: Boxplots of human-likeness scores for StyleGestures (SG), baseline (BL), and ground truth conditions (GT), with conditions grouped by speech or listening.

Condition	Mean Jerk	Mean Acceleration
Ground-truth S	38660.78 (SD=830	1101.37 (SD=287.00)
Ground-truth L	23980.68 (SD=4494.39)	524.74 (SD=148.00)
Baseline S	10318.88 (SD=2741.81)	422.34 (SD=120.20)
Baseline L	4633.45 (SD=1981.94)	182.61 (SD=84.49)
StyleGestures S	3392.73 (SD=6620.53)	235.97 (SD=306.05)
StyleGestures L	3395.10 (SD=4340.59)	215.46 (SD=174.55)

Table 1: Mean Jerk and mean Acceleration for the generated speech (S) and listening (L) behaviour.

4.1.2 Study 2: "Human-likeness for listening". 22 participants participated and passed the attention checks. The mean age was 41.8 years (SD=13.94). 9 identified as female and 13 identified as male. 16 participants were from the UK, 1 from the USA, 2 from Ireland, and 2 from New Zealand.

The median score for SG was 47 (95% CI[45.00,49.00]), for the baseline 41(95% CI[40.00, 44.00] and for the ground truth 56.5(95% CI[53.00, 60.00]. Paired Wilcoxon signed-rank tests were conducted between SG and baseline, SG and GT, and baseline and GT conditions for listening. The results showed a statistically significant difference in the human-likeness perception between the SG and baseline conditions ($Z = 16265.5$, $p < 0.0001$). The results also showed a significant difference between the SG and GT conditions ($Z = 16506.0$, $p < 0.0001$). Lastly, there was a significant difference between the baseline and GT conditions ($Z = 11646.5$, $p < 0.0001$).

4.1.3 Study 3: "Human-likeness for gesticulation". 22 persons participated and passed the attention checks. The mean age was 35.2 years (SD=12.4). 6 identified as female and the other 16 identified as male. 20 participants were UK nationals, 1 identified as a USA national and 1 participant resided in Ireland.

The median score for SG was 47 (95% CI[45.00,49.00]), for the baseline 41(95% CI[40.00, 44.00] and for the ground truth 56.5(95% CI[53.00, 60.00]. We conducted paired Wilcoxon signed-rank tests to the SG, baseline and ground-truth conditions. It revealed 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).

4.2 Objective Analysis

We calculated the mean jerk and mean acceleration.The result for the listening and speech motion can be found in table 1.

5 DISCUSSION

We conducted three user studies to evaluate the quality of our model on generating listening motion. We found that our adaptation of StyleGestures (SG) under performs in comparison to the baseline (BL) and the ground-truth (GT) for the mismatching study. Even though 60% are correctly identified as matching stimuli, more stimuli are identified as "they're equal", than for the BL (where only 72% was correctly identified. It shows that for quite some situations participants found it hard to identify the correct segment.

The second study looked at human-likeness for listening behaviour. Here we found significant differences between the three conditions. GT scored the highest, followed by SG and BL. Since this evaluation excludes audio, participants are more focused on evaluating the motion aspect.

In the third study we evaluated human-likeness for speaking. We found a significant difference for the two conditions with GT, but no significant difference between SG and BL, which is an interesting finding since the baseline model incorporates semantic information in relation to its gestures. However, the notion of semantic related gestures is not something that we can catch with human-likeness evaluations, since these revolve around motion quality and not appropriateness of gestures with speech audio.

When we look at the results of the objective metrics, we can observe that for BL the jerk and acceleration is much higher for the speech behaviour than the listening behaviour. For SG, there is not a large difference in mean jerk and mean acceleration between speech and listening (although the standard deviation is). One possible explanation for this phenomenon is the significant disparity between listening and speech behaviours within the data.

Since the main aim of this work was to compare SG to BL and the GT for generated listening behaviour, the results from study 1 and 2 give an indication that we can use generative models, originally used for co-speech gesture generation, for generating listening motion.

Future work should focus on the appropriateness of generated listening behaviour in relation to (speech) audio, and the evaluation of data driven listening behaviour in embodied conversational agents to see whether such an approach could lead to proper nonverbal feedback. Additionally, these full body motion models should be compared with and against models that integrate facial expressions [11], as traditionally much attention has been paid to nonverbal facial feedback channels as well as the fact that a lot of human to human interaction revolves around face to face communication.

6 CONCLUSION

We assessed a generative model's efficacy in creating listening motion. Comparing three conditions, our approach closely approximated ground-truth human-likeness. It is an initial step toward automating nonverbal feedback integration, needing more research for real-life scenarios.

ACKNOWLEDGEMENT

This research was funded by Flemish Research Foundation grant no. 1S95020N and no. V410522N. We would like to thank Anouk Neerincx for proof reading, and Taras Kucherenko for insightful discussions during the main author's stay in Stockholm, Sweden.

REFERENCES

[1] Chaitanya Ahuja, Shugao Ma, Louis-Philippe Morency, and Yaser Sheikh. 2019. To react or not to react: End-to-end visual pose forecasting for personalized avatar during dyadic conversations. In *2019 International conference on multimodal interaction*. 74–84.

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

[3] David Antonio Gomez Jauregui, Tom Giraud, Brice Isableu, and Jean-Claude Martin. 2021. Design and evaluation of postural interactions between users and a listening virtual agent during a simulated job interview. *Computer Animation and Virtual Worlds* 32, 6 (2021), e2029.

[4] Hendrik Buschmeier and Stefan Kopp. 2018. 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*. 1213–1221.

[5] Che-Jui Chang, Sen Zhang, and Mubbasir Kapadia. 2022. 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*. 784–789.

[6] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, et al. 2019. What makes a good conversation? Challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[7] Marco Gillies, Xueni Pan, Mel Slater, and John Shawe-Taylor. 2008. Responsive Listening Behavior. *Computer Animation and Virtual Worlds* 19, 5 (2008), 579–589.

[8] F Sebastian Grassia. 1998. Practical parameterization of rotations using the exponential map. *Journal of graphics tools* 3, 3 (1998), 29–48.

[9] Yuan He, André Pereira, and Taras Kucherenko. 2022. 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*. 1–8.

[10] Dirk Heylen, Elisabetta Bevacqua, Catherine Pelachaud, Isabella Poggi, Jonathan Gratch, and Marc Schröder. 2011. Generating listening behaviour. *Emotion-oriented systems: The Humaine handbook* (2011), 321–347.

[11] Patrik Jonell, Taras Kucherenko, Gustav Eje Henter, and Jonas Beskow. 2020. 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*. 1–8.

[12] Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, and Gustav Eje Henter. 2021. HEMVIP: Human Evaluation of Multiple Videos in Parallel. In *Proceedings of the 2021 International Conference on Multimodal Interaction* (Montréal, QC, Canada) *(ICMI '21)*. Association for Computing Machinery, New York, NY, USA, 707–711. https://doi.org/10.1145/3462244.3479957

[13] Adam Kendon. 1970. Movement coordination in social interaction: Some examples described. *Acta psychologica* 32 (1970), 101–125.

[14] Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. 2019. Analyzing input and output representations for speech-driven gesture generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. 97–104.

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

[16] Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. 2021. A large, crowdsourced evaluation of gesture generation systems on common data: The GENEA Challenge 2020. In *26th international conference on intelligent user interfaces*. 11–21.

[17] Jessica L Lakin and Tanya L Chartrand. 2003. Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychological science* 14, 4 (2003), 334–339.

[18] Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha S Srinivasa, and Yaser Sheikh. 2019. Talking with hands 16.2 m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 763–772.

[19] RM Maatman, Jonathan Gratch, and Stacy Marsella. 2005. Natural behavior of a listening agent. In *Intelligent Virtual Agents: 5th International Working Conference, IVA 2005, Kos, Greece, September 12-14, 2005. Proceedings 5*. Springer, 25–36.

[20] Izidor Mlakar, Zdravko Kačič, and Matej Rojc. 2014. Describing and animating complex communicative verbal and nonverbal behavior using Eva-framework. *Applied artificial intelligence* 28, 5 (2014), 470–503.

[21] Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. 2023. A Comprehensive Review of Data-Driven Co-Speech Gesture Generation. *arXiv preprint arXiv:2301.05339* (2023).

[22] Ronald Poppe, Khiet P Truong, Dennis Reidsma, and Dirk Heylen. 2010. Backchannel strategies for artificial listeners. In *Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings 10*. Springer, 146–158.

[23] Manuel Rebol, Christian Güti, and Krzysztof Pietroszek. 2021. Passing a non-verbal turing test: Evaluating gesture animations generated from speech. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, 573–581.

[24] B Series. 2014. Method for the subjective assessment of intermediate quality level of audio systems. *International Telecommunication Union Radiocommunication Assembly* (2014).

[25] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 4779–4783.

[26] Nguyen Tan Viet Tuyen and Oya Celiktutan. 2022. Agree or Disagree Generating Body Gestures from Affective Contextual Cues during Dyadic Interactions. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1542–1547.

[27] Pieter Wolfert, Jeffrey M Girard, Taras Kucherenko, and Tony Belpaeme. 2021. To rate or not to rate: Investigating evaluation methods for generated co-speech gestures. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 494–502.

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

[29] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–16.

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

[31] Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2022. The GENEA Challenge 2022: A large evaluation of data-driven co-speech gesture generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 736–747.