The GENEA Challenge 2022: A large evaluation of data-driven co-speech gesture generation

Youngwoo Yoon*
youngwoo@etri.re.kr
ETRI
Daejeon, Republic of Korea

Pieter Wolfert*
pieter.wolfert@ugent.be
IDLab, Ghent University – imec
Ghent, Belgium

Taras Kucherenko*
tkucherenko@ea.com
SEED – Electronic Arts (EA)
Stockholm, Sweden

Carla Viegas
cviegas@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, USA
NOVA University Lisbon
Lisbon, Portugal

Teodor Nikolov
tnikolov@hotmail.com
Umeå University
Umeå, Sweden

Mihail Tsakov
tsakovm@gmail.com
Umeå University
Umeå, Sweden

Gustav Eje Henter
ghe@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

ABSTRACT

This paper reports on the second GENEA Challenge to benchmark data-driven automatic co-speech gesture generation. Participating teams used the same speech and motion dataset to build gesture-generation systems. Motion generated by all these systems was rendered to video using a standardised visualisation pipeline and evaluated in several large, crowdsourced user studies. Unlike when comparing different research papers, differences in results are here only due to differences between methods, enabling direct comparison between systems. This year's dataset was based on 18 hours of full-body motion capture, including fingers, of different persons engaging in dyadic conversation. Ten teams participated in the challenge across two tiers: full-body and upper-body gesticulation. For each tier we evaluated both the human-likeness of the gesture motion and its appropriateness for the specific speech signal. Our evaluations decouple human-likeness from gesture appropriateness, which previously was a major challenge in the field.

The evaluation results are a revolution, and a revelation. Some synthetic conditions are rated as significantly more human-like than human motion capture. To the best of our knowledge, this has never been shown before on a high-fidelity avatar. On the other hand, all synthetic motion is found to be vastly less appropriate for the speech than the original motion-capture recordings.

*Equal contribution and joint first authors.

ICMI '22, November 7–11, 2022, Bengaluru, India

ACM ISBN 978-1-4503-9390-4/22/11.
https://doi.org/10.1145/3536221.3558058

CCS CONCEPTS

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

KEYWORDS

gesture generation, embodied conversational agents, evaluation paradigms

ACM Reference Format:
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 *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3536221.3558058

1 INTRODUCTION

This paper is concerned with systems for automatic generation of nonverbal behaviour, and how these can be compared in a fair and systematic way in order to advance the state-of-the-art. This is of importance as nonverbal behaviour plays a key role in conveying a message in human communication [40]. A large part of nonverbal behaviour consists of so called co-speech gestures, spontaneous hand and body gestures that relate closely to the content of the speech [4], and that have been shown to improve understanding [20]. Embodied conversational agents (ECAs) benefit from gesticulation, as it improves interaction with social robots [45] and willingness to cooperate with an ECA [44].

Synthetic gestures used to be based on rule-based systems, e.g., [10, 46]; see [52] for a review. These are gradually being supplanted by data-driven approaches, e.g., [5, 13, 33, 36], with recent work [1, 31, 59, 60] showing improvements in gesticulation production for ECAs. However, results from different gesture-generation studies are typically not directly comparable [56]. Studies usually rely on different data sources to train their models. The visualisations of

their generated gestures often have different avatars and production values, which can affect the perception of the gestures. On top of this, studies make use of a variety of different methodologies to evaluate the gestures. All these differences are, however, external to the actual methods that drive the gesture generation.

In this paper, we report on the GENEA Challenge 2022, the second joint gesture-generation challenge. (GENEA stands for "Generation and Evaluation of Non-verbal Behaviour for Embodied Agents".) The aim of the challenge is not to select the best team – it is not a contest, nor a competition – but to be able to directly compare different approaches and outcomes. By providing a common dataset for building gesture-generation systems, along with common evaluation standards and a shared visualisation procedure, we control for all other sources of variation except the system building itself. This makes it possible to assess and advance the state of the art in gesture generation, and to measure the gap between it and natural co-speech gestures. Comparing different methods and their performance also helps identify what matters most in gesture generation, and where the bottlenecks are. In particular, this year's results make it abundantly clear that natural-looking data-driven gesture motion is achievable today, but that synthetic gestures are much less appropriate for the accompanying speech than the ground-truth motion is. Our concrete contributions are:

(1) Four large-scale user studies that jointly evaluate a large number of gesture-generation models on a common dataset using a common 3D model and rendering method.
(2) Demonstrating a new method for subjective assessment of gesture appropriateness for speech, that successfully controls for the human-likeness of the motion.
(3) To the best of our knowledge, the first results that identify synthetic gesture motion that surpasses the human-likeness of good motion capture data on a high-fidelity avatar.
(4) The first clear evidence that synthetic gestures are much less appropriate for the specific speech than natural motion is, even when controlling for the human-likeness of the motion.
(5) Providing open code and high-quality, to facilitate reproducibility and enable future research to compare and benchmark against systems from the challenge.
(6) Bringing researchers together in order to advance the state-of-the-art in gesture generation.

The remainder of this paper first briefly discusses current gesture-evaluation practices and how challenges can help. We then describe this year's challenge data, setup, evaluation, and results, as well as implications of our findings. Additional material is available via the project website at youngwoo-yoon.github.io/GENEAchallenge2022/.

2 RELATED WORK

Most previous work proposing new gesture-generation methods incorporates an evaluation to support the merits of their method. Human gesture perception is highly subjective, and there are currently no widely accepted objective measures of gesture perception, so many publications have conducted human assessments instead. However, previous subjective evaluations have several drawbacks, as reviewed in [56]. Some major issues are the coverage of systems being compared and the scale of the studies. This creates an insular landscape where particular model families only are applied to particular datasets, and never contrasted against one another. Evaluations also sometimes fail to anchor system performance against natural ("ground truth") motion from test data held out from training. Studies also differ in how the motion is visualised, where some prior work displays motion through stick figures, or applies it to a physical agent. Neither of these may allow the same expressiveness or range of motion as a high-quality 3D-rendered avatar.

Other fields have done well using challenges to standardise evaluation techniques, establish benchmarks, and track and evolve the state of the art. For example, the Blizzard Challenges have since their inception in 2005 (see [7]) helped advance our sister field of text-to-speech (TTS) technology and identified important trends in the specific strengths and weaknesses in different speech-synthesis paradigms [27]. Data, evaluation stimuli, and subjective ratings remain available after these challenges, and have been widely used both for benchmarking subsequent TTS systems, e.g., [12, 49], and in research on the perception of natural and artificial speech, e.g., [17, 41, 42, 47, 62].

In 2020 we organised the first gesture-generation challenge, the GENEA Challenge 2020 [32]. In addition to being an exercise in benchmarking both new [29, 38, 50] and previously-published [1, 30, 60] gesture-generation methods, the results of that challenge have since helped improve gesture-generation benchmarking in other ways as well. Researchers have, for example, used the 2020 visualisation [53], and the objective [6] and subjective [61] evaluation methodologies, as a basis for future research. The data has also been used to benchmark subsequent gesture-generation models [15, 58], and even for automatic quality assessment [19]. In this paper, we follow up on the 2020 challenge by reporting on the second gesture-generation challenge, the GENEA Challenge 2022.

3 TASK AND DATA

The GENEA Challenge 2022 focused on data-driven automatic co-speech gesture generation. Specifically, given a sequence $\boldsymbol{s}$ of input features that describe human speech – which could involve any combination of an audio waveform, a time-aligned text transcription, and a speaker ID – the task is to generate a corresponding sequence $\hat{\boldsymbol{g}}$ of 3D poses describing gesture motion that an agent might perform while uttering this speech (facial expression is not considered). This is the same basic task as in the 2020 challenge, while at the same time we changed the dataset (as described below) and refined the evaluation (as detailed in Section 5).

Compared to 2020, we wanted to expand the dataset to include finger motion, lower-body motion, and material from multiple speakers in dyadic interactions. We therefore based our new challenge on the Talking With Hands 16.2M gesture dataset [35], which comprises 50 hours of audio (close-talking microphones) and motion-capture recordings of several pairs of people having a conversation freely on a variety of topics, recorded in distinct takes each about 10 minutes long. This is one of the largest datasets of parallel speech and 3D motion (in joint-angle space) publicly available in the English language. We removed parts of the dataset (46 out of 116 takes) that lacked audio or had low motion-capture quality, especially for the fingers. Note that despite the dataset being dyadic by design, this year's challenge focused on generating one side of the conversation, without awareness of the interaction partner.

Speech data was shared with participants as WAV audio with no additional processing beyond the anonymisation applied by [35], which replaced many proper nouns with silence. We also provided text transcriptions of the speech, in tab-separated value (TSV) files, and a metadata file with unique anonymous labels for each speaker. The TSV files were created by first applying Google Cloud automatic speech recognition, followed by thorough manual review to correct recognition errors and add punctuation for all parts of the dataset (training, validation, and test).

Motion data was downsampled to 30 frames per second and further transformed in two ways. Firstly, we updated the default skeletal definition relative to which all motion data is defined, from what appeared to be a contorted and arbitrary definition, to a standard "T-pose". The data was recomputed to match this pose using motion re-targeting inside MotionBuilder, retaining as much of the original visual quality as possible, whilst ensuring that the data had no discontinuities (e.g., at rotations near 180°). We found that this transformation substantially improved the output of the baseline system UBA in Section 4.2. Secondly, we standardised the position and orientation of speakers in all takes. Originally, each take would have the two speakers occupy two locations and face each other. We standardised this on a per-take basis such that all speakers, on average, face the same direction, and occupy the same location. This change was made to streamline data visualisation and to remove potential confusion due to different positions and orientations across different takes. Motion data was shared with participants in the Biovision hierarchy (BVH) format.

The challenge data was split into a training set (18 h), a validation set (40 min), and a test set (40 min), with only the training and validation sets initially shared with the teams. All these data subsets are publicly available via the Zenodo data release at doi.org/10.5281/zenodo.6998230. The validation and test data each comprised 40 *chunks* (contiguous excerpts approximately one minute long), to promote generation methods that are stable over long segments of speech, and was restricted to recordings ("takes" in the nomenclature of [35]) with finger motion tracking for the chosen speaker. The validation data was intended for internal benchmarking during development, so participants were allowed to train their final submitted models on both training and validation data if they wished.

Teams were allowed to only train on a subset of the data and were allowed to enhance the data they trained on however they liked. They were also allowed to make use of additional speech data (audio and text) from other sources, and models derived from such data, e.g., BERT [14] and Wav2Vec [2]. However, it was not permitted to use any other motion data, nor any pre-trained motion models, other than what we provided for the challenge.

4 SETUP AND PARTICIPATION

The challenge began on May 16, 2022, when speech-motion training data was released to participating teams. Test inputs (WAV, TSV, and speaker ID, but no motion output) were released to the teams on June 20, with teams required to submit BVH files with their generated gesture motion for these inputs by June 27. Manual tweaking of test inputs or the output motion was not allowed, since the idea was to evaluate synthesis performance in an unattended setting. As a precondition for participating in the evaluation, teams agreed to submit a companion paper describing their system for review and possible publication at ACM ICMI.

4.1 Tiers

This year's challenge evaluation was divided into two tiers, one for full-body motion and one for upper-body motion only. Teams could enter motion into either tier, or into both, but could only make one submission per tier. Teams that entered into both tiers were allowed to submit different motion (BVH files) to each tier, if they wished. Both tiers used the same training data but differed in which parts of the avatar that were allowed to move, and in the camera angle used for the video stimuli in the evaluation, as follows:

Full-body tier In this tier, the entire virtual character was free to move, including moving around in space relative to the fixed camera. Motion was visualised from an angle facing the character that showed most of the legs, but not where the feet touched the ground. This perspective was chosen to show as much as possible of the character, whilst obscuring foot penetration or foot sliding artefacts from view, since these artefacts arguably do not relate to co-speech gestures. For an example of this camera perspective, see Figure 1a.

Upper-body tier In this tier, the virtual character used a fixed position and a fixed pose from the hips down, with only the upper body free to move. Motion was visualised from a camera angle facing the character, cropped slightly below the hips, such that the hands always should remain in view. Any motion of the lower-body joints in submitted BVH files was ignored by the visualisation. This camera perspective is shown in Figure 1b.

4.2 Baselines and participating teams

The challenge evaluation featured three types of motion sources: natural motion capture from the speakers in the database, baseline systems based on open code, and submissions by teams participating in the challenge. We call each source of motion in a tier a *condition* (not a "system", since not all conditions represent motion synthesised by an artificial system). Each condition was assigned a unique three-letter *label* or *condition ID*, where the first character signifies the tier, with F for the full-body tier and U for the upper-body tier.

Natural motion was labelled **FNA** in the full-body tier and **UNA** in the upper-body tier (NA for "natural"). These conditions can be seen as a top line, and surpassing their performance essentially means outperforming the dataset itself, subject to limitations due to the motion capture and visualisation.

The natural top line can be contrasted against the two baseline systems included in the challenge, which represent previously published gesture-generation approaches with free and open code, adapted to run on the 2022 challenge training data. These two baselines were:

Text-based baseline (FBT/UBT) This motion was generated by the gesture-synthesis approach from [60] (which takes text transcriptions with word-level timestamps as the input) but adapted to joint rotations as described in [32]. Motion from this baseline used a fixed lower body but was included in

Table 1: Conditions participating in the evaluation. Teams are ordered alphabetically. The following non-standard abbreviations were used: *AR* for "Auto-regression", *SA* for "Neural self-attention" (e.g., Transformers), *GANs* for "Generative adversarial networks or adversarial loss terms", and *MM* for "Motion matching", *Frame-wise* for "Generating output frame-by-frame", *Stoch. output* for "Stochastic output", and *Smoothed* for "Smoothing was applied".

Baseline or team name	Inputs used			Techniques used					Frame-wise	Stoch. output	Smoothed
	Aud.	Text	Sp. ID	AR	RNNs	SA	VAEs	Other			
Audio-only baseline [30]	✓				✓				✓		✓
Text-only baseline [60]		✓		✓	✓				✓		✓
DeepMotion [39]	✓	✓		✓		✓	✓	CNNs	✓	✓	
DSI [43]	✓			✓	✓	✓					
FineMotion [28]	✓	✓		✓	✓				✓		✓
Forgerons [16]	✓			✓	✓		✓		✓	✓	
GestureMaster [63]	✓	✓	✓					Hand-crafted rules, MM			✓
IVI Lab [11]	✓	✓	✓	✓	✓				✓	✓	✓
Murple AI lab	No paper submitted										
ReprGesture [57]	✓	✓	✓	✓	✓	✓	✓	CNNs, GANs			✓
TransGesture [26]	✓			✓	✓				✓		✓
UEA Digital Humans [54]	✓	✓	✓		✓				✓		

both tiers, as conditions **FBT** and **UBT** (B for "baseline" and T for "text"). The code is available at github.com/youngwoo-yoon/Co-Speech_Gesture_Generation/.

Audio-based baseline (UBA) This motion was generated by the Audio2Repr2Pose motion-synthesis approach [30], which only takes speech audio into account when generating output, adapted to joint rotations as described in [32]. Motion from this baseline was only included in the upper-body tier, as condition **UBA** (A for "audio"). Code is available in the challenge GitHub repository at github.com/genea-workshop.

These are the same baselines as in the GENEA Challenge 2020. They were included to track the progress of the field and to provide continuity between different years of the challenge.

Separate from top lines and baselines, a total of 10 teams participated in the GENEA evaluation, with 8 *entries* (a.k.a. *submissions*) to the full-body tier and 8 entries to the upper-body tier. Submissions were labelled with the prefix FS and US (S for "submission") depending on the tier, followed by a single character to distinguish between different submissions in the same tier. In particular, challenge entries to the full-body tier were labelled **FSA**–**FSI**, and entries to the upper-body tier were labelled **USJ**–**USQ**. Condition FSE was withdrawn before the evaluation. These labels are anonymous and have no relationship to team names or identities, but teams are free to reveal their label(s) if they wish.

Table 1 lists the baselines and participating teams, with basic information about their approach and references to their system-description papers. One team lacks information, since they did not submit a paper for review.

5 EVALUATION

We conducted a large-scale, crowdsourced, joint evaluation of gesture motion from the 10 full-body conditions and 11 upper-body conditions using a within-subject design (i.e., every rater evaluated all conditions in each tier). For each tier, two orthogonal aspects of the generated gestures were evaluated:

Human-likeness Whether the motion of the virtual character looks like the motion of a real human, controlling for the effect of the speech. We sometimes use "motion quality" as a synonym for this.

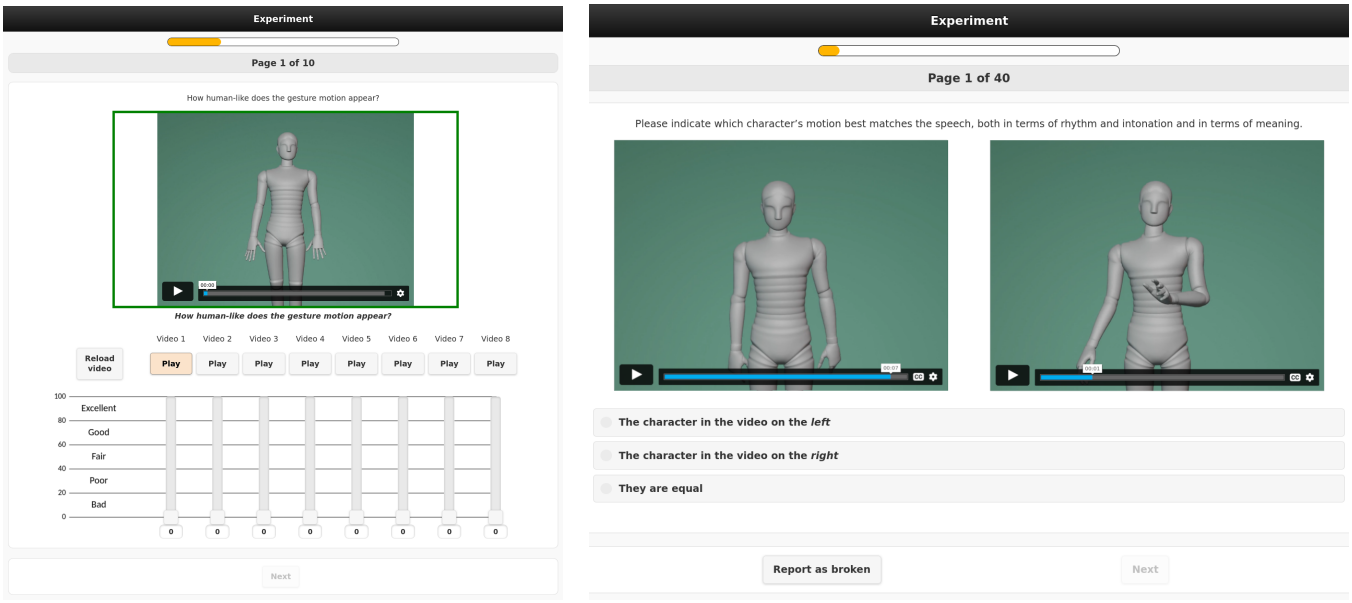
Appropriateness (a.k.a. "specificity") Whether the motion of the virtual character is appropriate for the given speech, controlling for the human-likeness of the motion.

5.1 Stimuli

From the 40 test-set chunks we selected 48 short *segments* of test speech and corresponding test motion to be used in the subjective evaluations, based on the following criteria: i) Segments should be around 8 to 10 seconds long. ii) The character should only be speaking, not passively listening, in the segments. (No turn-taking, but backchannels from the interlocutor were OK.) iii) Segments should not contain any parts where Lee et al. had replaced the speech by silence for anonymisation. iv) Segments should be more or less complete phrases, starting at the start of a word and ending at the end of a word, and not end on a "cliffhanger". v) Finally, recorded motion capture in the segments (i.e., the FNA motion) should not contain any significant artefacts such as whole-body vibration or hands flicking open and closed due to poor finger tracking. This does not imply that the motion capture is perfect or completely natural, just that the level of finger-tracking quality in the stimuli was consistent with the better parts of the source data.

The 48 segments selected in this way were between 5.6 and 12.1 seconds in duration and on average 9.5 seconds long. Audio was loudness normalised following EBU R128 [51] to achieve a consistent listening volume in the user studies.

We used the same virtual avatar for all all videos rendered during the challenge and the evaluation. The avatar can be seen in Figure 1. The avatar originally had 56 joints (full body including fingers) and was designed to be gender neutral and omit eyes or mouth, to help evaluators focus on the rest of the body instead. All teams had access to the official visualisation and rendering pipeline during the system-building phase, in the form of code, a portable Docker



(a) Human-likeness interface (HEMVIP) and full-body video

(b) Appropriateness interface and upper-body videos

Figure 1: Screenshots of the evaluation interfaces used in the studies, also showing the camera perspectives used by the tiers.

container, as well as a webserver to which BVH files could be submitted to be rendered as video. The visualisation server code is provided at github.com/TeoNikolov/genea_visualizer/ and the rendered stimulus videos at doi.org/10.5281/zenodo.6997925.

5.2 Human-likeness evaluation

The human-likeness evaluation closely followed the human-likeness evaluation in the GENEA Challenge 2020 [32], by presenting multiple motion examples in parallel and asking the subject to provide a rating for each one. All stimulus videos on the same page (a.k.a. screen) of the evaluation corresponded to the same speech segment but different conditions. The advantage of this method, called HEMVIP (Human Evaluation of Multiple Videos in Parallel) [25], is that differences in rating between the different conditions can be analysed using pairwise statistical tests, which helps control for variation between different subjects and different input speech segments; see [25]. The videos used in this evaluation had the audio removed, since it has been found that speech and gesture perception influence each other [8] and can confound motion evaluations [23]. Code is provided at github.com/jonepatr/hemvip/tree/genea2022/.

Each evaluation page asked participants "How human-like does the gesture motion appear?" and presented eight video stimuli to be rated on a scale from 0 (worst) to 100 (best) by adjusting an individual GUI slider for each video. An example of the evaluation interface can be seen in Figure 1a. Like in [25, 32], the 100-point rating scale was anchored by dividing it into successive 20-point intervals with labels (from best to worst) "Excellent", "Good", "Fair", "Poor", and "Bad", from the Mean Opinion Score ITU standard [22].

After reading the instructions, each subject completed one training page (to familiarise them with the task) followed by 10 pages of ratings for the evaluation. Responses given on the training page were not included in the analysis. The evaluation was balanced in exactly the same way as in [32]. Condition FNA/UNA was included on every page to help calibrate evaluators' ratings and keep them consistent throughout. Since motion-capture data projected onto a virtual character may not necessarily look perfectly natural, there was no requirement to rate the best motion as 100.

5.3 Appropriateness evaluation

The appropriateness evaluation was designed to assess the link between the motion and the input speech, separate from the intrinsic human-likeness of the motion. In the previous GENEA Challenge, appropriateness was evaluated using a HEMVIP-based rating study very similar to that for human-likeness, except that speech audio was included in the videos. Test takers were asked to ignore the motion quality and only rate the appropriateness of the motion for the speech [32]. Unfortunately, that evaluation was not altogether successful, since the *mismatched* condition M – which paired natural motion segments with unrelated speech segments, intended as a bottom line – attained the second-highest appropriateness rating, above all synthetic systems. This suggests a significant dependence between the human-likeness of a motion segment and its perceived appropriateness for speech, confounding the evaluation.

For the GENEA Challenge 2022, we decided to evaluate motion appropriateness for speech in a different way. Our design goal was to assess appropriateness whilst controlling for the human-likeness of the motion in an effective way. To do so, we took the idea of mismatching and used it within every condition: On each page, subjects were presented with a pair of videos containing the same speech audio. Both videos contained motion from the same condition and thus had the same overall motion quality, but one was matched to the speech audio and the other mismatched,

belonging to unrelated speech. Whether the left or the right video was mismatched was randomised. Subjects were then asked to "Please indicate which character's motion best matches the speech, both in terms of rhythm and intonation and in terms of meaning." In response, they could choose the character on the left, on the right, or indicate that the two were equally well matched ("They are equal", also referred to as *equal* or a *tie*). We asked for preferences rather than ratings since there is evidence [55] that this is more efficient in pairwise comparisons like these. A screenshot of the interface used for the appropriateness studies is presented in Figure 1b.

The extent to which test-takers prefer the character with the matched motion reveals how specific the gesture motion is to the given speech: Random motion will result in a 50–50 split, whereas conditions whose motion is more specifically appropriate to the input speech are expected to elicit a higher relative preference for the matched motion. In this type of evaluation, condition M (the mismatched condition) from the 2020 challenge will perform at chance rate, rather than being tied for second highest as in 2020. This approach to control for motion quality was first piloted in [23].

Concretely, we created the mismatched stimuli by taking the 48 existing speech and motion segments from the evaluation, and permuted the motion in between them such that no motion segment ever remained in its original place. As the 48 different segments did not all have the same length, a longer or shorter segment of motion generally had to be excerpted from the motion chunks (original or generated), so as to match the new speech duration. The starting point of the motion video was always the same as in the respective matched stimulus video (i.e., corresponding to the start of a phrase).

After an instruction page and a training page, each subject evaluated 40 pages with one pair of videos each. This means that subjects watched 80 videos total in each study, the same number of videos as was evaluated in the human-likeness studies (ignoring the training pages in all cases). Each study was balanced such that each speech segment, condition, and order of the two videos appeared approximately equally many times.

5.4 Test takers and attention checks

It has recently been found that crowdsourced evaluations are not significantly different from in-lab evaluations in terms of results and consistency [24]. The challenge therefore adopted an entirely crowdsourced approach. Test takers (a.k.a. subjects) were recruited through the crowdsourcing platform Prolific. We used Prolific's built-in pre-screening tools to restrict the pool of test-takers in two ways: i) subjects were required to reside in any of six English-speaking countries, namely UK, IE, USA, CAN, AUS, and NZ, and ii) subjects were required to have English as their first language.

We conducted four user studies, two for human-likeness and two for appropriateness. A subject could take one or more studies, but could only participate in each study at most once, and could not use a phone or tablet to take the test.

Each study incorporated four attention checks per person, to make sure that subjects were paying attention to the task and remove insincere test-takers. For the human-likeness studies, these attention checks took the form of a text message "Attention! You must rate this video NN" superimposed on the video. "NN" would be a number from 5 to 95, and the subject had to set the corresponding slider to the requested value, plus or minus 3, to pass that attention check. For the appropriateness studies, the attention checks either displayed a brief text message over the gesticulating character, reading "Attention! Please report this video as broken", or they temporarily replaced the audio with a synthetic voice speaking the same message. Subjects were exposed to two attention checks of each kind. To pass the attention check, participants had to click a button marked "Report as broken" seen in Figure 1b, forwarding them to the next pair of videos in the evaluation. In all studies, the attention-check messages did not appear until a few seconds into each attention-check video, so that participants who only would watch the first seconds would be unlikely to pass the checks.

Subjects who failed two or more attention checks were removed from the respective study without being paid, since Prolific's policies do not allow rejecting a subject on the basis of a single failed attention check. Right before submitting their results, subjects also filled in a short questionnaire to gather broad, anonymous demographic information about the population taking the test.

A design goal of the human-likeness studies was that every combination of two distinct conditions should appear on the pages approximately equally often, and at least 600 times (not counting FNA/UNA, which appeared on every page). To meet this goal, we recruited 121 test takers that successfully passed the attention checks and completed the full-body study, and 150 test takers that successfully passed the attention checks and completed the upper-body study. Of the 121 test takers in the full-body study, 60 identified as female, 60 as male, and 1 did not want to disclose their gender. The same numbers for the 150 upper-body test takers were 74, 75, and 1, respectively. For the full-body test takers, 2 resided in AU, 2 in CAN, 3 in IE, 110 in the UK, and 4 in the USA. The upper-body study had 1 AU, 4 IE, 134 UK, and 11 USA.

For the appropriateness studies, our design goal was for each condition to receive as many responses per condition as the number of ratings that each condition (aside from FNA/UNA) received in the corresponding human-likeness evaluation. This works out to 880 responses per condition in the full-body studies and 990 responses per condition in the upper-body studies. Because a subject in these studies provided half as many responses as in a human-likeness study (40 vs. 80), the appropriateness studies needed to recruit approximately twice as many test takers. In the end, 247 test takers successfully passed the attention checks in the full-body study, while 304 passed the attention checks in the upper-body study. Of the 247 subjects in the full-body study, 137 identified as female, 107 as male, and 3 did not want to disclose their gender. The same numbers for the 304 upper-body test takers were 127, 173, and 4, respectively. For the full-body test takers, 3 resided in AU, 13 in CAN, 10 in IE, 2 in NZ, 211 in the UK, and 8 in the USA. The upper-body study had 2 AU, 10 CAN, 1 IE, 256 UK, and 35 USA.

Test takers were remunerated 6 GBP for each successfully completed human-likeness study. Since the median completion time was 28 minutes each, this corresponds to a median compensation just above 12 GBP per hour. Similarly, the appropriateness studies took a median of 24 or 25 minutes to complete, and earned a reward of 5.5 GBP each, amounting to around 13 GBP per hour. These compensation levels all exceed the UK national living wage.

Response data from the evaluation and statistical analysis code is provided at doi.org/10.5281/zenodo.6939888.

Table 2: Summary statistics of responses from all user studies, with 95% confidence intervals. "M." stands for "matched" and "Mism." for "mismatched". "Percent matched" identifies how often subjects preferred matched over mismatched motion.

(a) Full-body

ID	Median human-likeness	Appropriateness: Num. responses M.	Tie	Mism.	Percent matched (splitting ties)
FNA	70 ∈ [69, 71]	590	138	163	74.0 ∈ [70.9, 76.9]
FBT	27.5 ∈ [25, 30]	278	362	250	51.6 ∈ [48.2, 55.0]
FSA	71 ∈ [70, 73]	393	216	269	57.1 ∈ [53.7, 60.4]
FSB	30 ∈ [28, 31]	397	163	330	53.8 ∈ [50.4, 57.1]
FSC	53 ∈ [51, 55]	347	237	295	53.0 ∈ [49.5, 56.3]
FSD	34 ∈ [32, 36]	329	256	302	51.5 ∈ [48.1, 54.9]
FSF	38 ∈ [35, 40]	388	130	359	51.7 ∈ [48.2, 55.1]
FSG	38 ∈ [35, 40]	406	184	319	54.8 ∈ [51.4, 58.1]
FSH	36 ∈ [33, 38]	445	166	262	60.5 ∈ [57.1, 63.8]
FSI	46 ∈ [45, 48]	403	178	312	55.1 ∈ [51.7, 58.4]

(b) Upper-body

ID	Median human-likeness	Appropriateness: Num. responses M.	Tie	Mism.	Percent matched (splitting ties)
UNA	63 ∈ [61, 65]	691	107	189	75.4 ∈ [72.5, 78.1]
UBA	33 ∈ [31, 34]	424	264	303	56.1 ∈ [52.9, 59.3]
UBT	36 ∈ [34, 39]	341	367	287	52.7 ∈ [49.5, 55.9]
USJ	53 ∈ [52, 55]	461	164	365	54.8 ∈ [51.6, 58.0]
USK	41 ∈ [40, 44]	454	185	353	55.1 ∈ [51.9, 58.3]
USL	22 ∈ [20, 25]	282	548	159	56.2 ∈ [53.0, 59.4]
USM	41 ∈ [40, 42]	503	175	328	58.7 ∈ [55.5, 61.8]
USN	44 ∈ [41, 45]	443	190	352	54.6 ∈ [51.4, 57.8]
USO	48 ∈ [47, 50]	439	209	335	55.3 ∈ [52.1, 58.5]
USP	29.5 ∈ [28, 31]	440	180	376	53.2 ∈ [50.0, 56.4]
USQ	69 ∈ [68, 70]	504	182	310	59.7 ∈ [56.6, 62.9]

6 RESULTS AND DISCUSSION

6.1 Results of human-likeness studies

Each test taker in the human-likeness studies contributed 76 ratings to the analyses after removing attention checks, giving a total of 9,196 ratings for the full-body study and 11,400 ratings for the upper-body study. The results are visualised in Figure 2, with summary statistics for the ratings of all conditions given in the first half of Table 2, together with 95% confidence intervals for the true median. These confidence intervals were computed using order statistics, leveraging the binomial distribution cdf; see [18].

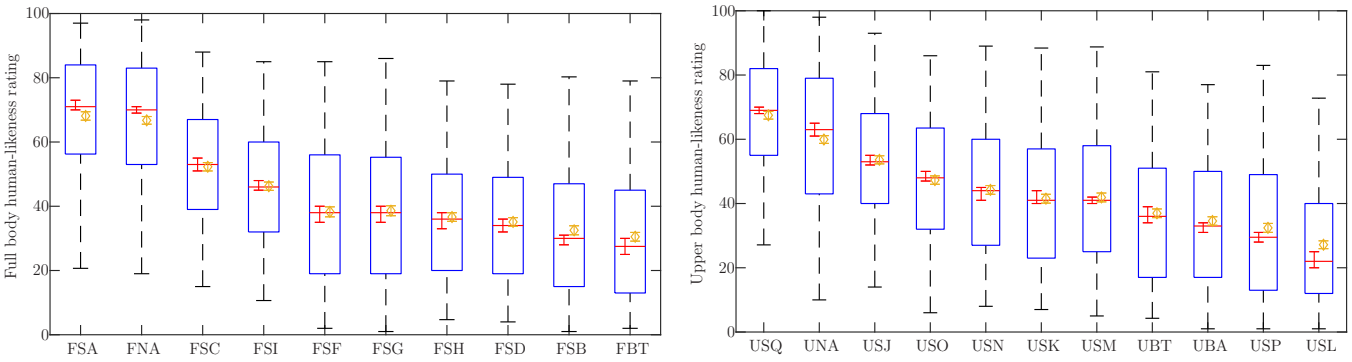
The distributions are seen to be quite broad. This is common in evaluations like HEMVIP [25], since the range of the responses not only reflects differences between conditions, but also extraneous variation, e.g., between stimuli, in individual preferences, and in how critical different raters are in their judgments. In contrast, the plotted confidence intervals are seen to be quite narrow, since the statistical analysis can mitigate the effects of much of this variation.

To analyse the significance of differences in median rating between different conditions, we applied two-sided pairwise Wilcoxon signed-rank tests to all unordered pairs of distinct conditions in each study. (This is the same methodology as in the GENEA Challenge 2020 [32].) Unlike Student's t-test, which assumes that rating differences follow a Gaussian distribution, this analysis is valid also for ordinal response scales, like those we have here. For each condition pair, only cases where both conditions appeared on the same page were included in the analysis of significant differences. Because this analysis is based on pairwise statistical tests, it can potentially resolve differences between conditions that are smaller than the width of the confidence intervals for the median in Figure 2, since those confidence intervals are inflated by variation that the statistical test controls for. The p-values computed in the significance tests were adjusted for multiple comparisons on a per-study basis using the Holm-Bonferroni method [21].

Our statistical analysis found all but 5 out of 45 condition pairs to be significantly different in the full-body study and all but 2 out of 55 condition pairs to be significantly different in the upper-body study, all at the level $\alpha = 0.05$ after Holm-Bonferroni correction. The significant differences we identified are visualised in Figure 3.

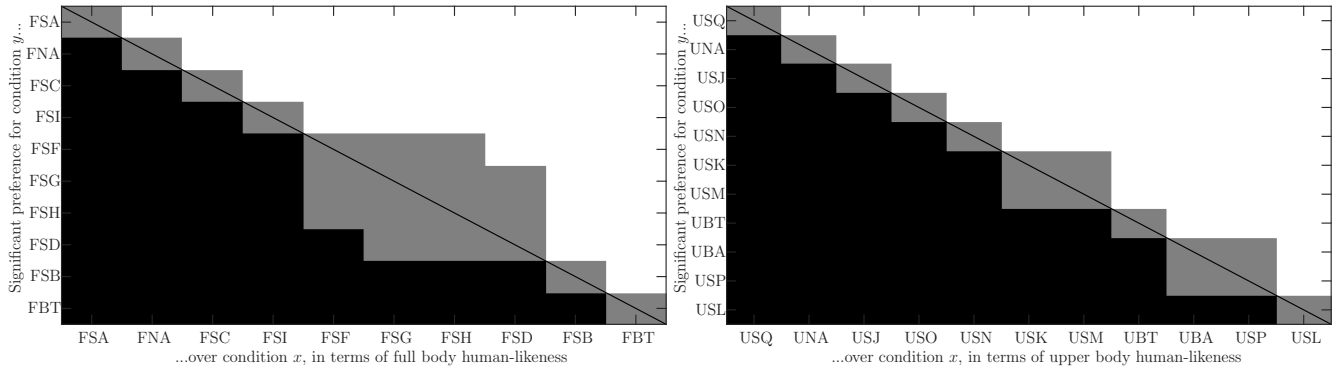
6.2 Discussion of human-likeness results

Generating convincingly human-like gestures is a difficult problem, and nearly all conditions rated significantly below natural motion capture. However, each tier contains an entry which is rated significantly above the motion from the motion-capture recordings in terms of human-likeness. This is a leap forwards from GENEA 2020, and we believe it represents a motion quality not before seen in large-scale evaluations. That said, we caution that this does not mean that the motion is completely human-like – indeed, the median rating is much below 100, which would constitute "completely human-like" as per our explicit instructions to test takers. What it does mean is that the motion was perceived as more human-like (in terms of median) than the motion-capture in the database, specifically than the motion-capture data used for FNA/UNA in the subjective evaluation. In making this distinction, it is important to keep in mind that our human-likeness evaluation is constrained by several factors: For example, the nominally natural motion is constrained by our ability to accurately capture and visualise human motion. Finger motion capture is especially problematic, and the finger motion could not be chosen so as to look completely natural in all segments evaluated, potentially degrading the ratings of FNA/UNA as a result. Moreover, the use of a deliberately neutral 3D avatar lacking potentially distracting human features such as gaze and lip motion significantly reduces the bandwidth of the communication channel to the user, which lowers the threshold for what needs to be achieved in order to match human motion ratings in the evaluation. In addition, the greater interquartile range of ratings of UNA compared to FNA could mean that the process of imposing full-body motion from a walking and talking human onto an avatar with fixed lower body may not always yield completely natural results. An artificial system might have its training data cleaned of problematic instances, so as to prevent it from generating such motion, giving it an edge over UNA. Future GENEA Challenges intend to only consider full-body motion.



(a) Full-body **(b) Upper-body**

Figure 2: Box plots visualising the ratings distribution in the human-likeness studies. Red bars are medians and yellow diamonds are means, each with a 0.05 confidence interval and a Gaussian assumption for the means. Box edges are at 25 and 75 percentiles, while whiskers cover 95% of all ratings for each condition. Conditions are ordered descending by sample median for each tier.



(a) Full-body **(b) Upper-body**

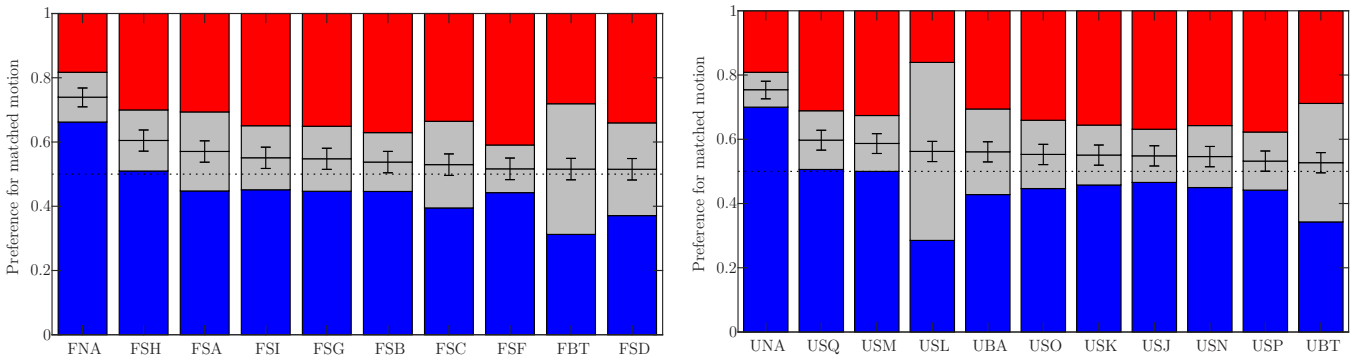
Figure 3: Significant differences in human-likeness. White means the condition listed on the y-axis rated significantly above the condition on the x-axis, black means the opposite (y rated below x), and grey means no statistically significant difference at level $\alpha = 0.05$ after Holm-Bonferroni correction. Conditions use the same order as the corresponding subfigure in Figure 2.

We found fewer significant differences in the full-body study, perhaps meaning that full-body motion is more difficult to rate consistently. For example, it contains more behavioural variation, as the character now is moving their legs and changing position, perhaps in response to the conversation partner. Future challenges intend to include information about both conversation parties in the evaluation, so that test takers can be interlocutor-aware.

6.3 Results of appropriateness studies

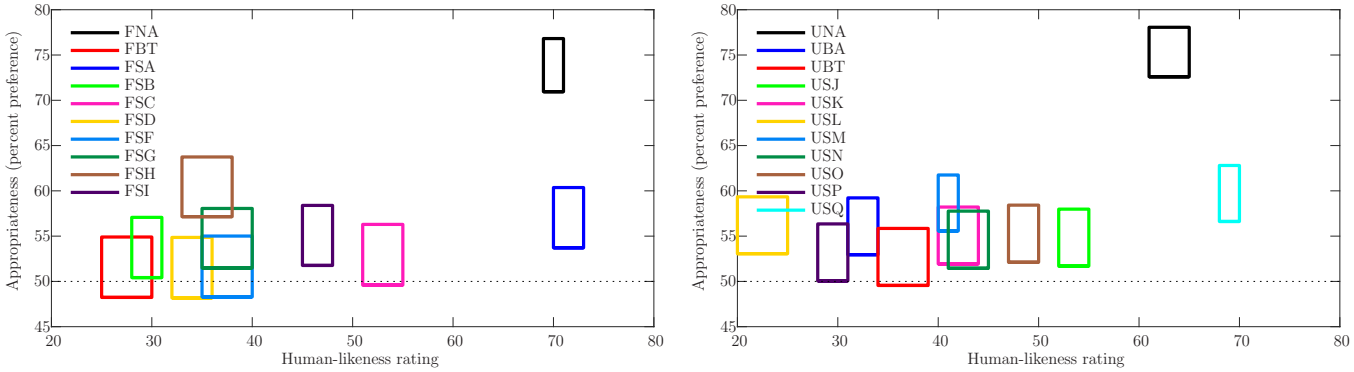
We gathered a total of 8,867 responses for the full-body study and 10,910 responses from the upper-body study that were included in the analysis. Raw response statistics for all conditions in each of the two studies are shown in the second half of Table 2, together with 95% Clopper-Pearson confidence intervals for the fraction of time that the matched video was preferred over the mismatched, after dividing ties equally between the two groups (rounding up in case of non-integer counts). The quoted confidence intervals were rounded outward to ensure sufficient coverage.

The response distributions in the two studies are further visualised through bar plots in Figure 4, while Figure 5 visualises the results of the entire challenge in a single coordinate system per tier. Overall, the distribution of the three different responses across the different conditions is consistent with the mismatching study reported in [23]. No system has a relative preference for matched motion below 50%, which is the theoretical bottom line, attained by a system whose motion has no relation to the speech. (Here and forthwith, we only consider the relative preference in the sample after dividing ties equally.) The greatest relative preference, a 75% preference for matched motion, is observed for natural motion capture, i.e., FNA/UNA. This should be considered a good result, since previous studies that have incorporated mismatched stimuli, e.g., [23, 32], have found that they sometimes are difficult for participants to distinguish from matched ones, especially if they – like here – both correspond to segments where the character is speaking. Furthermore, both matched and mismatched motion stimuli have their starting points aligned to the start of a phrase in the speech, meaning that the motion in the stimulus videos might initially be



(a) Full-body **(b) Upper-body**

Figure 4: Bar plots visualising the response distribution in the appropriateness studies. The blue bar (bottom) represents responses where subjects preferred the matched motion, the light grey bar (middle) represents tied ("They are equal") responses, and the red bar (top) represents responses preferring mismatched motion, with the height of each bar being proportional to the fraction of responses in each category. The black horizontal lines bisecting the light grey bars represent the proportion of matched responses after splitting ties, each with a 0.05 confidence interval. The dotted black line indicates chance-level performance. Conditions are ordered by descending preference for matched motion after splitting ties.



(a) Full-body **(b) Upper-body**

Figure 5: Joint visualisation of the evaluation results for each tier. Box widths show 95% confidence intervals for the median human-likeness rating and box heights show 95% confidence intervals for the preference for matched motion in percent, indicating appropriateness.

more similar to each other than if the mismatched motion had been excerpted completely at random and not aligned to the start of phrase boundaries.

Unlike the human-likeness studies, the responses in the appropriateness studies are restricted to three categories and do not necessarily come in pairs for statistical testing in the same way as for the parallel sliders in HEMVIP. A different method for identifying significant differences therefore needs to be adopted. We used Barnard's test [3] to identify statistically significant differences at the level $\alpha = 0.05$ between all pairs of distinct conditions, applying the Holm-Bonferroni method [21] to correct for multiple comparisons as before. This analysis found 13 of 45 condition pairs to be significantly different in the full-body study and 10 out of 55 condition pairs to be significantly different in the upper-body study. Specifically, FNA/UNA were significantly more appropriate for the specific speech signal compared to all other, synthetic conditions. In addition, FSH was significantly more appropriate than FBT, FSC, FSD, and FSF in the full-body study. No other differences were statistically significant in either study.

Instead of comparing the appropriateness of different synthesis approaches against one another, one can compare against a random baseline (50/50 performance), and test if the observed effect size is statistically significantly different from zero. We can assess this at the 0.05 level by checking whether or not the confidence interval on the effect size overlaps with chance performance. From this perspective, FSA, FSB, FSG, FSH, FSI are significantly more appropriate than chance in the full-body study, and all systems except UBT are more appropriate than chance in the upper-body study. Unlike other significance tests in this text, these do not include a correction for multiple comparisons.

6.4 Discussion of appropriateness results

We find the results of the appropriateness evaluation thought-provoking, and revealing about the state of the field. It is clear that generating meaningful and appropriate gestures is still far from being a solved problem.

We see fewer statistical differences compared to the appropriateness study in GENEA 2020, which asked participants to rate the appropriateness of the stimuli on an absolute scale using HEMVIP [32]. However, that study was strongly biased towards conditions with high human-likeness, as discussed in Section 5.3. In effect, we have traded the high-resolution, high-bias method from GENEA 2020 for a reduced-resolution, low-bias method. We think this is a step forward, since most prior evaluations of gesture appropriateness for speech have been highly confounded by motion quality, whereas our new methodology is not. The fact that some synthetic conditions that distinguished themselves the most in terms of appropriateness, namely FSH and USM, exhibited middle-of-the-pack human-likeness, highlights success in disentangling motion appropriateness from motion quality.

7 CONCLUSIONS AND IMPLICATIONS

We have hosted the GENEA Challenge 2022, to directly compare many different gesture-generation methods and assess the state of the art in data-driven co-speech gesture generation. Our evaluation results show that, with the right method, synthetic motion can attain a human-likeness equal or better than the underlying motion-capture data. This is a big step forwards compared to the 2020 challenge. However, using a new evaluation paradigm, we find that synthetic gestures are much less appropriate for the speech than human gestures, also when controlling for differences in human-likeness.

We believe the challenge adds value to the research community in many ways. A lot can doubtlessly be learnt from the system-description papers by the participating teams. The materials we release from the challenge (e.g., time-aligned splits of audio, text, and gesture data; visualisation; code; and evaluation stimuli and responses) can have broad use for future benchmarking and research in gesture generation, similar to what happened after the 2020 challenge. In particular, the new methodology we demonstrate for assessing motion appropriateness for speech is much more accurate at controlling for the effect of motion quality and does not involve subjects making any direct comparisons between videos generated by different conditions. We believe this may enable direct comparison between different studies on the same data, *without* having to include the various other synthetic baseline conditions in the new user study.

Based on the fact that one condition in each tier managed to achieve excellent human-likeness, we expect that, in the medium-term future, gesture-generation systems should be able to advance to more consistently match motion capture in terms of human-likeness. This is similar to recent developments in verbal behaviour generation, where neural language models [9] and speech synthesisers [37, 48] trained on large datasets are approaching the text and speech produced by humans in terms of surface quality (but not necessarily appropriateness). As that evolution runs its course, we believe that research into appropriate rather than human-like motion is poised to become the new frontier in gesture generation. There is already evidence that existing deep-learning methods in principle can predict even the hard case of semantically motivated, communicative gestures from speech [33, 34].

We think that future challenges should study more difficult scenarios that are farther from being solved, for example full-body motion in dyadic interaction. That can also provide interesting opportunities for exploring other types of appropriateness, e.g., with respect to the interlocutor stance and behaviour, as studied in [23]. In general, challenges like the one described here can play an important part in identifying key factors for generating convincing co-speech gestures in practice, and help drive and validate future progress toward the goal of endowing embodied agents with natural and appropriate gesture motion.

ACKNOWLEDGMENTS

The authors wish to thank Meta Research for the data; Carolyn Saund, Axel Johansson, Christianne Sandstig, Leonhard Grosse, Natalia Kalyva, and Natasha Greenwood for the transcriptions; Esther Ericsson for the 3D character; Zerrin Yumak for input; Judith Bütepage, Minsu Jang, and Tony Belpaeme for informal review.

This research was partially supported by IITP grant no. 2017-0-00162 (Development of Human-care Robot Technology for Aging Society) funded by the Korean government (MSIT), by the Flemish Research Foundation (FWO) grant no. 1S95020N, by the Portuguese Foundation for Science and Technology grant no. SFRH/BD/127842/2016, and by the Knut and Alice Wallenberg Foundation, both through Wallenberg Research Arena (WARA) Media and Language – with in-kind contribution from the Electronic Arts (EA) R&D department, SEED – and through the Wallenberg AI, Autonomous Systems and Software Program (WASP).

REFERENCES

[1] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-controllable speech-driven gesture synthesis using normalising flows. *Comput. Graph. Forum* 39, 2 (2020), 487–496. https://doi.org/10.1111/cgf.13946

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS '20, Vol. 33)*. 12449–12460. https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html

[3] George Alfred Barnard. 1945. A new test for 2×2 tables. *Nature* 156, 3954 (1945), 177. https://doi.org/10.1038/156783b0

[4] Kirsten Bergmann, Volkan Aksu, and Stefan Kopp. 2011. The relation of speech and gestures: Temporal synchrony follows semantic synchrony. In *Proceedings of the Workshop on Gesture and Speech in Interaction (GESPIN '11)*.

[5] Kirsten Bergmann and Stefan Kopp. 2009. GNetIc – Using Bayesian decision networks for iconic gesture generation. In *Proceedings of the International Conference on Intelligent Virtual Agents (IVA '09)*. Springer, 76–89. https://doi.org/10.1007/978-3-642-04380-2_12

[6] Uttaran Bhattacharya, Elizabeth Childs, Nicholas Rewkowski, and Dinesh Manocha. 2021. Speech2AffectiveGestures: Synthesizing co-speech gestures with generative adversarial affective expression learning. In *Proceedings of the ACM International Conference on Multimedia (MM '21)*. ACM, New York, NY, USA. https://doi.org/10.1145/3474085.3475223

[7] Alan W. Black and Keiichi Tokuda. 2005. The Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech '05)*. ISCA, 77–80. https://doi.org/10.21437/Interspeech.2005-72

[8] Hans Rutger Bosker and David Peeters. 2021. Beat gestures influence which speech sounds you hear. *P. Roy. Soc. B* 288 (2021), 20202419. https://doi.org/10.1098/rspb.2020.2419

[9] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan,

Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS '20)*. 1877–1901. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[10] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2001. BEAT: The behavior expression animation toolkit. In *Proceedings of SIGGRAPH*. ACM, 477–486. https://doi.org/10.1007/978-3-662-08373-4_8

[11] 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 ACM International Conference on Multimodal Interaction (ICMI '22)*. ACM.

[12] Marcela Charfuelan and Ingmar Steiner. 2013. Expressive speech synthesis in MARY TTS using audiobook data and EmotionML. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech '13)*. ISCA, 1564–1568. https://doi.org/10.21437/Interspeech.2013-395

[13] Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella. 2015. Predicting co-verbal gestures: A deep and temporal modeling approach. In *Proceedings of the International Conference on Intelligent Virtual Agents (IVA '15)*. Springer, 152–166. https://doi.org/10.1007/978-3-319-21996-7_17

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '18)*. ACL. https://doi.org/10.18653/v1/N19-1423

[15] Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2021. ExpressGesture: Expressive gesture generation from speech through database matching. *Comput. Animat. Virt. W.* 32, 3-4 (2021), e2016. https://doi.org/10.1002/cav.2016

[16] Saeed Ghorbani, Ylva Ferstl, and Marc-André Carbonneau. 2022. Exemplar-based stylized gesture generation from speech: An entry to the GENEA Challenge 2022. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI '22)*. ACM.

[17] Avashna Govender, Anita E. Wagner, and Simon King. 2019. Using pupil dilation to measure cognitive load when listening to text-to-speech in quiet and in noise. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech '19, Vol. 20)*. ISCA, 1551–1555. https://doi.org/10.21437/Interspeech.2019-1783

[18] Gerald J. Hahn and William Q. Meeker. 1991. *Statistical Intervals: A Guide for Practitioners*. Vol. 92. John Wiley & Sons. https://books.google.be/books?id=y3o0DgAAQBAJ

[19] Zhiyuan He. 2022. Automatic quality assessment of speech-driven synthesized gestures. *Int. J. Comput. Games. Tech.* 2022 (2022). https://doi.org/10.1155/2022/1828293

[20] Judith Holler, Kobin H. Kendrick, and Stephen C. Levinson. 2018. Processing language in face-to-face conversation: Questions with gestures get faster responses. *Psychon. B. Rev.* 25, 5 (2018), 1900–1908. https://doi.org/10.3758/s13423-017-1363-z

[21] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6, 2 (1979), 65–70. https://www.jstor.org/stable/4615733

[22] International Telecommunication Union, Telecommunication Standardisation Sector. 1996. *Methods for subjective determination of transmission quality*. Recommendation ITU-T P.800. https://www.itu.int/rec/T-REC-P.800-199608-I

[23] 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 ACM International Conference on Intelligent Virtual Agents (IVA '20)*. ACM, Article 31, 8 pages. https://doi.org/10.1145/3383652.3423911

[24] Patrik Jonell, Taras Kucherenko, Ilaria Torre, and Jonas Beskow. 2020. Can we trust online crowdworkers? Comparing online and offline participants in a preference test of virtual agents. In *Proceedings of the ACM International Conference on Intelligent Virtual Agents (IVA '20)*. ACM, Article 30, 8 pages. https://doi.org/10.1145/3383652.3423860

[25] 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 ACM International Conference on Multimodal Interaction (ICMI '21)*. ACM, 707–711. https://doi.org/10.1145/3462244.3479957

[26] Naoshi Kaneko, Yuna Mitsubayashi, and Geng Mu. 2022. TransGesture: Autoregressive gesture generation with RNN-transducer. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI '22)*. ACM.

[27] Simon King. 2014. Measuring a decade of progress in text-to-speech. *Loquens* 1, 1 (2014), e006. https://doi.org/10.3989/loquens.2014.006

[28] Vladislav Korzun, Anna Beloborodva, and Arkady Ilin. 2022. ReCell: replicating recurrent cell for auto-regressive pose generation. In *Companion publication of the 2021 ACM International Conference on Multimodal Interaction (ICMI '22)*. ACM.

[29] Vladislav Korzun, Ilya Dimov, and Andrey Zharkov. 2021. Audio and text-driven approach for conversational gestures generation. In *Proceedings of Computational Linguistics and Intellectual Technologies (DIALOGUE '21)*. http://www.dialog-21.ru/media/5526/korzunvaplusdimovinpluszharkovaa031.pdf

[30] 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 ACM International Conference on Intelligent Virtual Agents (IVA '19)*. ACM, 97–104. https://doi.org/10.1145/3308532.3329472

[31] Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexanderson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI '20)*. ACM, 242–250. https://doi.org/10.1145/3382507.3418815

[32] 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 *Proceedings of the ACM Annual Conference on Intelligent User Interfaces (IUI '21)*. ACM, 11–21. https://doi.org/10.1145/3397481.3450692

[33] Taras Kucherenko, Rajmund Nagy, Patrik Jonell, Michael Neff, Hedvig Kjellström, and Gustav Eje Henter. 2021. Speech2Properties2Gestures: Gesture-property prediction as a tool for generating representational gestures from speech. In *Proceedings of the ACM International Conference on Intelligent Virtual Agents (IVA '21)*. ACM, 145–147. https://doi.org/10.1145/3472306.3478333

[34] Taras Kucherenko, Rajmund Nagy, Michael Neff, Hedvig Kjellström, and Gustav Eje Henter. 2022. Multimodal analysis of the predictability of hand-hesture properties. In *Procceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS '22)*. IFAAMAS, 770–779. https://doi.org/10.5555/3535850.3535937

[35] Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha S. Srinivasa, and Yaser Sheikh. 2019. Talking With Hands 16.2M: 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 (ICCV '19)*. IEEE, 763–772. https://doi.org/10.1109/ICCV.2019.00085

[36] Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun. 2010. Gesture controllers. *ACM Trans. Graph.* 29, 4, Article 124 (2010), 11 pages. https://doi.org/10.1145/1778765.1778861

[37] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. 2019. Neural speech synthesis with transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI '19, Vol. 33)*. 6706–6713. https://doi.org/10.1609/aaai.v33i01.33016706

[38] JinHong Lu, TianHang Liu, ShuZhuang Xu, and Hiroshi Shimodaira. 2021. Double-DCCCAE: Estimation of body gestures from speech waveform. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '21)*. IEEE, 900–904. https://doi.org/10.1109/ICASSP39728.2021.9414660

[39] Shuhong Lu and Andrew Feng. 2022. The DeepMotion entry to the GENEA Challenge 2022. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI '22)*. ACM.

[40] David McNeill. 1992. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press. https://doi.org/10.1177/002383099403700208

[41] Gabriel Mittag and Sebastian Möller. 2020. Deep learning based assessment of synthetic speech naturalness. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech '20)*. ISCA, 1748–1752. https://doi.org/10.21437/Interspeech.2020-2382

[42] Sebastian Möller, Florian Hinterleitner, Tiago H. Falk, and Tim Polzehl. 2010. Comparison of approaches for instrumentally predicting the quality of text-to-speech systems. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech '10)*. ISCA, 1325–1328. https://doi.org/10.21437/Interspeech.2010-413

[43] Khaled Saleh. 2022. Hybrid seq2seq architecture for 3D co-speech gesture generation. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI '22)*. ACM.

[44] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2013. To err is human(-like): Effects of robot gesture on perceived anthropomorphism and likability. *Int. J. Soc. Robot.* 5, 3 (2013), 313–323. https://doi.org/10.1007/s12369-013-0196-9

[45] Maha Salem, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2011. A friendly gesture: Investigating the effect of multimodal robot behavior in human-robot interaction. In *Proccedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN '11)*. IEEE, 247–252. https://doi.org/10.1109/ROMAN.2011.6005285

[46] Giampiero Salvi, Jonas Beskow, Samer Al Moubayed, and Björn Granström. 2009. SynFace—Speech-driven facial animation for virtual speech-reading support. *EURASIP J. Audio Spee.*, Article 191940 (2009), 10 pages. https://doi.org/10.1155/2009/191940

[47] Ibon Saratxaga, Jon Sanchez, Zhizheng Wu, Inma Hernaez, and Eva Navas. 2016. Synthetic speech detection using phase information. *Speech Commun.* 81 (2016), 30–41. https://doi.org/10.1016/j.specom.2016.04.001

[48] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A.

Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2018. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In *Proceedings of the IEEE international conference on Acoustics, Speech and Signal Processing (ICASSP '18)*. IEEE, 4799–4783. https://doi.org/10.1109/ICASSP.2018.8461368

[49] Éva Székely, João P. Cabral, Mohamed Abou-Zleikha, Peter Cahill, and Julie Carson-Berndsen. 2012. Evaluating expressive speech synthesis from audiobooks in conversational phrases. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC '12)*. ELRA, 3335–3339. https://aclanthology.org/L12-1513/

[50] Ausdang Thangthai, Kwanchiva Thangthai, Arnon Namsanit, Sumonmas Thatphithakkul, and Sittipong Saychum. 2021. Speech gesture generation from acoustic and textual information using LSTMs. In *Proceedings of the International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON '21)*. IEEE, 718–723. https://doi.org/10.1109/ECTI-CON51831.2021.9454931

[51] European Broadcasting Union. 2020. Loudness normalisation and permitted maximum level of audio signals. EBU Recommendation EBU R 128v4. https://tech.ebu.ch/docs/r/r128.pdf

[52] Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview. *Speech Commun.* 57 (2014), 209–232. https://doi.org/10.1016/j.specom.2013.09.008

[53] Siyang Wang, Simon Alexanderson, Joakim Gustafson, Jonas Beskow, Gustav Eje Henter, and Éva Székely. 2021. Integrated speech and gesture synthesis. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI '21)*. ACM, 177–185. https://doi.org/10.1145/3462244.3479914

[54] Jonathan Windle, David Greenwood, and Sarah Taylor. 2022. UEA Digital Humans entry to the GENEA Challenge 2022. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI '22)*. ACM.

[55] 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 ACM International Conference on Multimodal Interaction (ICMI '21)*. ACM, 494–502. https://doi.org/10.1145/3462244.3479889

[56] 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* 52, 3 (2022), 379–389. https://doi.org/10.1109/THMS.2022.3149173

[57] Sicheng Yang, Zhiyong Wu, Minglei Li, Mengchen Zhao, Jiuxin Lin, Liyang Chen, and Weihong Bao. 2022. The ReprGesture entry to the GENEA Challenge 2022. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI '22)*. ACM.

[58] Payam Jome Yazdian, Mo Chen, and Angelica Lim. 2021. Gesture2Vec: Clustering gestures using representation learning methods for co-speech gesture generation. https://openreview.net/forum?id=0Kj5mhn6sw

[59] 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 Trans. Graph.* 39, 6, Article 222 (2020), 16 pages. https://doi.org/10.1145/3414685.3417838

[60] 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 *Proc. ICRA (ICRA '19)*. IEEE, 4303–4309. https://doi.org/10.1109/ICRA.2019.8793720

[61] Youngwoo Yoon, Keunwoo Park, Minsu Jang, Jaehong Kim, and Geehyuk Lee. 2021. SGToolkit: An interactive gesture authoring toolkit for embodied conversational agents. In *Proccedings of the Annual ACM Symposium on User Interface Software and Technology (UIST '21)*. ACM, 826–840. https://doi.org/10.1145/3472749.3474789

[62] Takenori Yoshimura, Gustav Eje Henter, Oliver Watts, Mirjam Wester, Junichi Yamagishi, and Keiichi Tokuda. 2016. A hierarchical predictor of synthetic speech naturalness using neural networks. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech '16)*. ISCA, 342–346. https://doi.org/10.21437/Interspeech.2016-847

[63] Chi Zhou, Tengyue Bian, and Kang Chen. 2022. GestureMaster: Graph-based speech-driven gesture generation. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI '22)*. ACM.