Evaluating Gesture Generation in a Large-scale Open Challenge: The GENEA Challenge 2022

TARAS KUCHERENKO*, SEED, Electronic Arts Inc, Stockholm, Sweden
PIETER WOLFERT*, Donders Institute for Brain, Cognition & Behaviour, Radboud Universiteit, Nijmegen, Netherlands and IDLab, Ghent University, Gent, Belgium
YOUNGWOO YOON*, ETRI, Daejeon, Korea (the Republic of)
CARLA VIEGAS, Carnegie Mellon University, Pittsburgh, United States and Nova University of Lisbon, Lisboa, Portugal
TEODOR NIKOLOV, Department of Computing Science, Umeå Universitet, Umeå, Sweden and Motorica AB, Stockholm, Sweden
MIHAIL TSAKOV, Department of Computing Science, Umeå Universitet, Umeå, Sweden
GUSTAV EJE HENTER, Division of Speech, Music and Hearing, KTH Royal Institute of Technology, Stockholm, Sweden and Motorica AB, Stockholm, Sweden

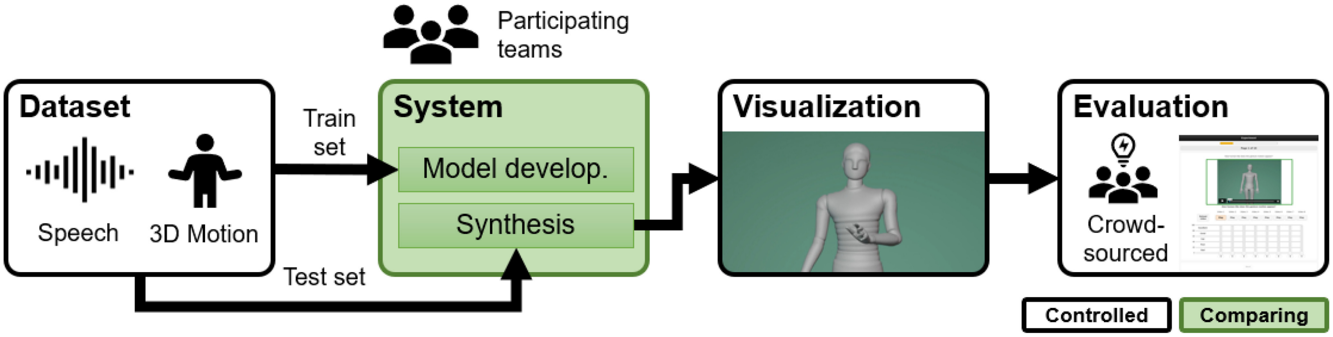


Fig. 1. Overview of the GENEA Challenge. We controlled the dataset, visualisation, and evaluation in order to compare different gesture-generation approaches in a fair and systematic way. The dataset includes speech audio, time-aligned speech text transcription, and speaker identity as input modalities and 3D body motion as the output modality. For the synthesised motion from the participating teams, video stimuli were rendered by a shared visualisation pipeline and evaluated jointly in crowdsourced user studies.

This article 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 articles, differences in results are here only due to differences between methods, enabling direct comparison between systems. The dataset was based on 18 hours of full-body motion capture, including fingers, of different persons engaging in a dyadic conversation. Ten teams participated in the challenge across two tiers: full-body and upper-body gesticulation. For each tier, we evaluated

*Taras Kucherenko, Pieter Wolfert, and Youngwoo Yoon contributed equally and are joint first authors.

This research was partially supported by the Industrial Fundamental Technology Development Program (no. 20023495) funded by MOTIE, Korea, 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 R&D department, SEED—and through the Wallenberg AI, Autonomous Systems and Software Program (WASP).

Authors' Contact Information: Taras Kucherenko, SEED, Electronic Arts Inc, Stockholm, Sweden; e-mail: tkucherenko@ea.com; Pieter Wolfert, Donders Institute for Brain, Cognition & Behaviour, Radboud Universiteit, Nijmegen, Gelderland, Netherlands and IDLab, Ghent University, Gent, Belgium; e-mail: pieter.wolfert@donders.ru.nl; Youngwoo Yoon, ETRI, Daejeon, Daejeon, Korea (the Republic of); email: youngwoo@etri.re.kr; Carla Viegas, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States and Nova University of Lisbon, Lisboa, Lisboa, Portugal; e-mail: cviegas@andrew.cmu.edu; Teodor Nikolov, Department of Computing Science, Umeå Universitet, Umeå, Sweden and Motorica AB, Stockholm, Sweden; e-mail: tnikolov@hotmail.com; Mihail Tsakov, Department of Computing Science, Umeå Universitet, Umeå, Sweden; e-mail: tsakovm@gmail.com; Gustav Eje Henter, Division of Speech, Music and Hearing, KTH Royal Institute of Technology, Stockholm, Stockholm, Sweden and Motorica AB, Stockholm, Sweden; e-mail: ghe@kth.se.

ACM 0730-0301/2024/06-ART32
https://doi.org/10.1145/3656374

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 has been a difficult problem in the field.

The evaluation results show some synthetic gesture conditions being rated as significantly more human-like than 3D human motion capture. To the best of our knowledge, this has not been demonstrated before. On the other hand, all synthetic motion is found to be vastly less appropriate for the speech than the original motion-capture recordings. We also find that conventional objective metrics do not correlate well with subjective human-likeness ratings in this large evaluation. The one exception is the Fréchet gesture distance (FGD), which achieves a Kendall's tau rank correlation of around -0.5. Based on the challenge results we formulate numerous recommendations for system building and evaluation.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → *Animation;*

Additional Key Words and Phrases: Animation, gesture generation, embodied conversational agents, evaluation paradigms

ACM Reference Format:
Taras Kucherenko, Pieter Wolfert, Youngwoo Yoon, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2024. Evaluating Gesture Generation in a Large-scale Open Challenge: The GENEA Challenge 2022. *ACM Trans. Graph.* 43, 3, Article 32 (June 2024), 28 pages. https://doi.org/10.1145/3656374

1 Introduction

This article 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 [McNeill 1992]. 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 [Bergmann et al. 2011], and that have been shown to improve understanding [Holler et al. 2018]. **Embodied conversational agents (ECAs)** benefit from gesticulation, as it improves interaction with social robots [Salem et al. 2011] and willingness to cooperate with an ECA [Salem et al. 2013]. Knowledge of how and when to gesture is also needed. This can for example be learnt from interaction data; see, e.g., Jonell et al. [2020a].

Synthetic gestures used to be based on rule-based systems, e.g., Cassell et al. [2001] and Salvi et al. [2009]; see Wagner et al. [2014] for a review. These are now being supplanted by data-driven approaches, e.g., Bergmann and Kopp [2009], Chiu et al. [2015], and Levine et al. [2010], with recent work [Alexanderson et al. 2020; Kucherenko et al. 2020; Yoon et al. 2020, 2019] showing improvements in gesticulation production for ECAs. For more in-depth reviews of recent data-driven approaches see Liu et al. [2021] and Nyatsanga et al. [2023].

Unfortunately, results from different gesture-generation studies are typically not directly comparable [Wolfert et al. 2022]. 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 article, we report on the GENEA[1] Challenge 2022. 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, as illustrated in Figure 1. Our setup is unique to the field of gesture generation, making it possible to reliably assess and advance the state-of-the-art, 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, the 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 natural motion-capture data is. The results also show that most objective metrics are not informative about the perceived human-likeness of the generated gestures.

Challenge participants benefit by working on the same problem together with researchers interested in the same topic, strengthening the research community, and get an opportunity to compare their systems to other competitive systems in a large and carefully-executed joint evaluation. They also work on and contribute towards a standardised evaluation setup which encourages future benchmarking and reproduction of results. Participants are required to write down their methods, results and experience in a system paper to be presented in conjunction with the challenge itself, giving them a chance to publish their work at ACM ICMI, a leading conference in the field. 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. A subjective evaluation that successfully disentangles motion human-likeness from its appropriateness for the associated speech.
3. To the best of our knowledge, the first results that identify synthetic gesture motion that surpasses the human-likeness of good 3D motion-capture data.
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. A validation study of many objective metrics for predicting motion human-likeness, finding that all metrics except the **Fréchet gesture distance (FGD)** are statistically indistinguishable from chance prediction.
6. Providing open code and high-quality data in the spirit of open source and reproducible research. This includes pre-processed multimodal training, validation, and test datasets; the standardised visualisation; submitted motion and video stimuli; a large number of subjective responses from the studies; and evaluation and analysis code.
7. Bringing researchers together in order to advance the state-of-the-art in gesture generation, and enabling future

[1] For "Generation and Evaluation of Non-verbal Behaviour for Embodied Agents".

research to compare and benchmark against systems and stimuli from the challenge.

Materials derived from the challenge are publicly available at youngwoo-yoon.github.io/GENEAchallenge2022.

This article is an extension of a previously published conference article on the challenge [Yoon et al. 2022], adding more comprehensive information and analyses, experiments on objective metrics, and a more in-depth discussion of challenge submissions, findings, recommendations, and limitations. The remainder of this article first (in Section 2) briefly discusses current gesture-evaluation practices and how challenges can help overcome their shortcomings. We then describe the challenge task and dataset in Section 3, followed by a breakdown of the challenge tiers and participating teams in Section 4. Section 5 then reviews the approaches taken by different challenge entries. In Section 6, we describe the design of the challenge evaluation, with results of the various evaluations recounted in Section 7 and discussed in Section 8. Each of these three sections detail both the core subjective evaluation as well as the objective metrics we computed, in that order. We round off by discussing challenge limitations (in Section 9) and summarising its conclusions and implications (in Section 10).

2 Related Work

2.1.1 Issues with Prior Evaluations and Evaluation Practices. Most works that propose new gesture-generation methods incorporate 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. Instead, human assessment via careful user studies is the gold standard in the field. However, previous subjective evaluations have several drawbacks, as reviewed in Wolfert et al. [2022]. Some major issues are the coverage of systems being compared and the scale of the studies. Like in Alexanderson et al. [2020], Kucherenko et al. [2021a, 2020], and Sadoughi and Busso [2019], proposed models are at most compared to one or two prior approaches (often a highly similar baseline) or possibly only to ablated versions of the same model. A large number of studies do not compare their outcomes with other methods at all, let alone other systems trained on the same data. This creates an insular landscape where particular model families are only applied to particular datasets and never contrasted against one another.

As for scale, large evaluations are expensive, and studies may not be able to recruit enough participants, thus leaving the differences between many pairs of studied systems unresolved and not statistically significant (cf. Yoon et al. [2020, 2019]). Questionnaires, which are one popular evaluation methodology (cf. Bergmann et al. [2010], Ishi et al. [2018], Ishii et al. [2018], Salem et al. [2012], Shimazu et al. [2018], and Yoon et al. [2019]) demand a lot of time and cognitive effort from test participants and may be infeasible to scale up to bigger studies. In addition, items used in questionnaires differ across studies and the set of questions used is often not standardised. Evaluations sometimes also fail to anchor system performance against natural ("ground truth") motion from test data held out from training (e.g., Ishii et al. [2018], Le and Pelachaud [2012], and Salem et al. [2012]).

Studies also differ in the dataset they train on (e.g., Ishii et al. [2018], Le and Pelachaud [2012], and Salem et al. [2012]) and in how the motion is visualised. For the latter, some prior work displays motion through stick figures (e.g., Kucherenko et al. [2019] and Wolfert et al. [2019]), or applies it to a physical agent (e.g., Ishi et al. [2018] and Salem et al. [2012]). Neither of these may allow the same expressiveness or range of motion as a 3D-rendered humanoid mesh avatar as seen in, e.g., Alexanderson et al. [2020] and Kucherenko et al. [2020].

2.1.2 Benefits of Challenges in Other Fields. 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 Black and Tokuda [2005]) helped advance our sister field of **text-to-speech** (**TTS**) technology and identified important trends in the specific strengths and weaknesses of different speech-synthesis paradigms [King 2014]. These challenges are defined by the use of common data and evaluation and their open participation in both academia and industry. More specifically, participants are provided a common dataset of speech audio and associated text transcriptions, which they use to build a system that generates synthetic speech audio. After the participants submit their systems, the resulting generated speech is subsequently evaluated in a large, joint evaluation, the results of which are provided to the teams. Submitted entries are identified by anonymised labels in official Blizzard Challenge results, but in practice, the vast majority of teams identify which label represents their entry in their article at the Blizzard Challenge Workshop describing the system that they submitted. Data, evaluation stimuli, and subjective ratings remain available after these challenges and have been widely used both for benchmarking subsequent TTS systems, e.g., Charfuelan and Steiner [2013] and Székely et al. [2012], and in research on the perception of natural and artificial speech, e.g., Govender et al. [2019], Huang et al. [2022], Mittag and Möller [2020], Möller et al. [2010], Saratxaga et al. [2016], and Yoshimura et al. [2016]. This has led to the development of new and novel methods, driven by past results, and since participants had access to the same data, significant advances have been made.

Challenges are also actively used in the computer-vision community, for instance for benchmarking purposes. Recent NTIRE [Zhang et al. 2020] and CLIC [Toderici et al. 2020] challenges, for example, compared systems for image compression and super-resolution respectively, also incorporating subjective human assessments similar to the challenge described in this article (although NTIRE used a MOS-like setup, which has been found to be less efficient than the side-by-side evaluation methodology we employ [Ribeiro et al. 2015]). This addresses the over-reliance on objective metrics in computer-vision evaluation, which, just like in speech quality and gesture generation, do not always align with human perception. The GENEA Challenge is inspired by these successes of challenges in other fields of study.

In 2020, we organised the first gesture-generation challenge, the GENEA Challenge 2020 [Kucherenko et al. 2021b]. In addition to being an exercise in benchmarking both new [Korzun et al. 2021; Lu et al. 2021; Thangthai et al. 2021] and previously-published [Alexanderson et al. 2020; Kucherenko et al. 2019; Yoon et al. 2019] 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 [Teshima et al. 2022; Wang et al. 2021; Zhang et al. 2023], and the objective [Bhattacharya et al. 2021] and subjective [Yoon et al. 2021] evaluation methodologies, as a basis for future research. The data has also been used to benchmark subsequent gesture-generation models [Ferstl et al. 2021; Yazdian et al. 2022], and even for automatic quality assessment [He 2022]. In this article, we follow up on the 2020 challenge by reporting on the second gesture-generation challenge, the GENEA Challenge 2022. This challenge expands the dataset, the range of behaviour considered, and the number of participating teams, and also improves the visualisation and the evaluation practises, in addition to considering objective metrics together with a large subjective evaluation.

2.1.3 Objective Metrics. Given that subjective evaluations are labour intensive, time-consuming, and costly, a large number of different objective metrics have been proposed as automated indicators of gesture-generation performance. Some of these, such as the commonly used **average position error** (**APE**) and **mean-squared position error** (**MSE**) [Nyatsanga et al. 2023; Wolfert et al. 2022], as well as the "**probability of correct keypoints**" (**PCK**) and its extension to **semantic relevance gesture recall** (**SRGR**) [Liu et al. 2022b], are used to score the similarity of generated poses to a corresponding recording of human motion. Alternatively, **canonical correlation analysis** (**CCA**) can be used to quantify the linear (Pearson) correlations between generated and reference poses [Bozkurt et al. 2015; Lu et al. 2021; Sadoughi and Busso 2019]. These methods are likely to struggle with the stochastic, one-to-many nature of human gestures (there is no single "correct" way to move), leading to high variance.

To accommodate the stochastic nature of motion, statistics such as the average magnitude of motion acceleration and jerk, and distances between motion speed histograms have been used to quantify how similar generated motion is to the distribution of human motion [Kucherenko et al. 2021a]. More recent developments have built on the **Fréchet inception distance** (**FID**) from image generation [Heusel et al. 2017] to propose new methods for comparing gesture-motion distributions [Ahuja et al. 2020; Yoon et al. 2020]. These methods were later used by, e.g, Ahuja et al. [2022], Ao et al. [2022], and Liu et al. [2022b, 2022a]. Beat consistency, which was first proposed for dance motion [Li et al. 2021], has also been used to assess gesture generation [Liu et al. 2022a]. However, few of these works experimentally validate their metrics. In this article, we use the many conditions and ratings gathered in our user studies to compute and validate five of the above objective metrics for gesture generation.

3 Task and Data

The GENEA Challenge 2022 focused on data-driven automatic co-speech gesture generation. Specifically, given a sequence 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 $\hat{g}$ of 3D poses describing gesture motion that an agent might perform while uttering this speech (facial expression is not considered). To enable direct comparison of different data-driven gesture-generation methods, all methods evaluated in the challenge were trained on the same gesture-speech dataset and their motion was visualised using the same virtual avatar and rendering pipeline. This is the same task as in the 2020 challenge, whilst at the same time we changed the dataset (as described below) and refined the evaluation (see Section 6).

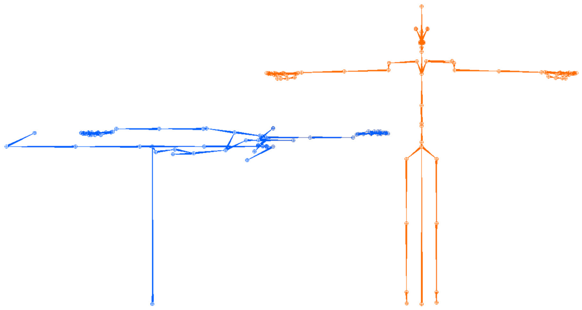


Fig. 2. Visualisations of the default skeletal pose of the data before and after processing. On the left (blue) is the original skeletal pose, as found in the Talking With Hands 16.2 M dataset shared by Lee et al. [2019]. On the right (orange) is the transformed skeletal pose (i.e., T-pose) used for the GENEA Challenge 2022.

3.1 Data

Compared to 2020, we wanted to expand the dataset to include finger motion, lower-body motion, and material from multiple speakers in dyadic interactions. The latter may provide more natural and interesting gestures than the Trinity Speech-Gesture Dataset [Ferstl and McDonnell 2018] used in 2020. We based our challenge on the Talking With Hands 16.2 M gesture dataset [Lee et al. 2019], which comprises 50 hours of audio (captured by close-talking directional 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. At the time of the challenge, this was likely the largest dataset 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, the challenge focused on generating one side of the conversation at a time, without awareness of the interaction partner. The data from the interaction partner in each dyad was typically also included in the challenge material, but as a separate recording without providing links between the two. This was the case for both the gesture synthesis and for the subsequent evaluation.

3.1.1 Speech Audio and Text. Speech data was shared with participants as WAV audio with no additional processing beyond the anonymisation applied by Lee et al. [2019], 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 to transcribe the audio recordings, followed by a manual review to correct recognition errors and add punctuation for all parts of the dataset (training, validation, and test).

3.1.2 Motion Data. Motion data was downsampled from 120 to 30 frames per second and further transformed in two ways:

First, we updated the default skeleton relative to which all motion data is defined, away from what appeared to be a contorted and arbitrary definition to a standard "T-pose", as shown in Figure 2. The T-pose is an animation-industry standard wherein all joint rotation values are described in relation to a T-shaped skeleton. This standard is widely adopted by existing 3D digital content-creation software like Blender and Maya. In fact, it is often the required 3D-skeleton pose when transferring the motion of one character onto another during animation re-targeting. Furthermore, the T-pose is expected to have better mathematical properties due to its symmetry and shape. In particular, the pose more closely resembles the poses found in the motion-capture data. As a result of this, most of the joint rotation values are expected to be closely distributed around zero. Consequently, this would reduce the risk of phase wrapping and gimbal locking across the skeleton, lending itself to smoother behaviour and interpolation in the Euler-angle space. This in turn leads to data that is more numerically stable, making it more practical for training machine-learning models. The data was recomputed to match a T-pose using motion re-targeting inside MotionBuilder (code and documentation are available[2]), 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.

Second, 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. More technically, in a right-hand xyz Cartesian coordinate system (y-up, z-forward), each speaker is on average positioned at world origin ($[x = 0,\ y = 0,\ z = 0]$), and on average facing the positive z-axis (a directional vector $[x = 0, y = 0, z = 1]$). Averaging was done for each take separately after taking 250 equidistant samples of the hip position and orientation, and then using linear-algebra operations to correct for the difference between the original and the standardised values. This change was made to streamline data visualisation and to remove potential confusion due to different absolute positions and orientations across different takes. The transformed motion data was shared with participants in the **Biovision Hierarchy** (**BVH**) format.

3.1.3 Data Splits. 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 participating teams. 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 were restricted to takes with finger-motion tracking for the chosen speaker. Some recordings with finger-capture data were excluded from consideration due to poor motion-capture quality, based on visual inspection of a short sample from each recording. 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. After the challenge, all the data subsets were made publicly available at zenodo.org/doi/10.5281/zenodo.6998230, and had been downloaded over 500 times when this article went to press.

3.1.4 Usage Rules. 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, for instance by manual annotation or by post-processing the speech and the motion. 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 [Devlin et al. 2018] and wav2vec [Baevski et al. 2020]. However, it was not permitted to use any other motion data, nor any pre-trained motion models, other than what the organisers provided for the challenge. Otherwise, the result would be likely to strongly depend on how much data each team can get access to (as has been the case in many Blizzard Challenges in speech synthesis), which is not an interesting scientific conclusion to replicate.

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, in the same format as that used by the challenge training data. 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 article describing their system for review and possible publication at the conference where the challenge took place. Evaluations took place only after the generated motion was submitted by all teams.

4.1 Tiers

The challenge evaluation was divided into two tiers, one for full-body motion and one for upper-body motion only. Each tier had its own reasons for being included. On the one hand, the data comprises recorded full-body motion from conversational interactions. It can furthermore be argued that human-embodied conversation uses the full body. Also, generating full-body behaviour seems like a harder problem, since it represents a higher-dimensional probability distribution which is more difficult to learn from a statistical perspective. Therefore, if full-body generation is solved, restricted versions of the problem can be expected to be solved as well. On the other hand, it is debatable to what extent the motion of the lower body whilst speaking constitutes co-speech gestures that depend on the speech, over other aspects such as proxemics and stance in response to the other parties in a conversation (which is data that was not provided to challenge participants this time). As a result, including lower-body motion may add unnecessary complexity to the gesture generation problem, and act as a distraction

[2] github.com/TeoNikolov/genea_visualizer/tree/master/scripts

Table 1. Conditions Participating in the Evaluation

	Per-tier		Inputs used			Hands	Techniques used					Frame-	Stoch.	Smoo-
Baseline or team name	label		Aud.	Text	Sp. ID	fixed	AR	RNNs	SA	VAEs	Other	wise	output	thed
GestureMaster [Zhou et al.]	FSA	USQ	✓	✓	✓						Hand-crafted rules, MGs			✓
Forgerons [Ghorbani et al.]	FSC	USO	✓			✓	✓	✓		✓		✓	✓	
DeepMotion [Lu and Feng]	FSI	USJ	✓	✓		✓	✓		✓	✓	CNNs	✓	✓	
DSI [Saleh]	FSF		✓				✓	✓	✓					
UEA Digital Humans [Windle et al.]	FSG	USM	✓	✓	✓			✓				✓		
ReprGesture [Yang et al.]		USN	✓	✓	✓	✓	✓	✓	✓	✓	CNNs, GANs			✓
IVI Lab [Chang et al.]	FSH	USK	✓	✓	✓	✓	✓	✓				✓	✓	✓
FineMotion [Korzun et al.]	FSD		✓	✓			✓	✓				✓		✓
Murple AI lab (no article submitted)	Not revealed		✓				✓	✓			Normalising flows	✓	✓	
Text-only baseline	FBT	UBT		✓		✓	✓	✓				✓		✓
Audio-only baseline		UBA	✓					✓				✓		✓
TransGesture [Kaneko et al.]		USL	✓			✓	✓	✓				✓		✓

Conditions are ordered based on their median human-likeness scores from higher to lower (see Table 3); this is why the previously published baselines appear near the bottom of the list, but not at the very end. The following abbreviations were used: *AR* for "Auto-regression", *CNN* for "Convolutional Neural Network", *RNN* for "Recurrent Neural Network", *SA* for "Neural self-attention" (e.g., Transformers), *GANs* for "Generative adversarial networks or adversarial loss terms", *VAEs* for "Variational auto-encoders", *MGs* for "Motion graphs", *Frame-wise* for "Generating output frame-by-frame", *Stoch. output* for "Stochastic output" (different output possible even if the inputs are the same), and *Smoothed* for "Smoothing was applied".

when evaluating the quality of generated gestures. Focusing on the upper body also is more consistent with earlier evaluations of co-speech gesture generation, such as the GENEA Challenge 2020 [Kucherenko et al. 2021b]. Because it is not clear which perspective to apply, our evaluation included one tier each for full-body and upper-body motion. 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 motions (BVH files) to each tier if they wished. Both tiers used the same training data but differed in which parts of the avatar 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, yet they may influence ratings if visible. An example of this camera perspective can be seen in Figure 3(a).

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 3(b).

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 (discussed in Sections 8.1 and 9).

The natural top line can be contrasted against the two baseline systems included in the challenge, which represent previously published gesture-generation methods that have been adapted to run on the 2022 challenge training data. The systems were selected with the requirements to be (1) open-sourced and well-documented, and (2) their authors were available to adapt the methods to the new data and also help challenge organisers and participants, should any issues arise. Unfortunately, none of the top-performing models from the previous challenge satisfied both conditions, whereas the two 2020 baselines did. Adapting these baselines to the present challenge provides continuity with the previous iteration of the challenge and helps track the progress of the field. The two baselines were thus:

Text-based baseline (FBT/UBT). This motion was generated by the gesture-synthesis approach from Yoon et al. [2019] (which takes text transcriptions with word-level timestamps as the input) but adapted to joint rotations. A neural sequence-to-sequence architecture is used, where an encoder processes a sequence of speech words and a decoder outputs a sequence of human poses. Motion from this baseline used a fixed lower body but was included in both tiers, as conditions **FBT** and **UBT** (B for "baseline" and T for "text"). The code is publicly available online at github.com/youngwoo-yoon/Co-Speech_Gesture_Generation.

Audio-based baseline (UBA). This motion was generated by the Audio2Repr2Pose motion-synthesis approach of Kucherenko et al. [2019], which only takes speech audio into account when generating output, adapted to joint rotations. This model uses a chain of two neural networks: one maps from speech to pose representation and another decodes the representation to a pose, generating motion frame-by-frame by sliding a window over the speech input. Motion from this baseline was only included in the upper-body tier, as condition **UBA** (A for "audio"). The code is publicly available online at github.com/genea-workshop/Speech_driven_gesture_ generation_with_autoencoder.

Source code for replicating the two baselines was available to participating teams during the challenge.

Separate from top lines and baselines, a total of 10 teams participated in the GENEA evaluation, with eight *entries* (a.k.a. *submissions*) to the full-body tier and eight entries to the upper-body tier. Together with test-set mocap and the baselines, this makes a total of 10 conditions in the full-body tier and 11 in 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 identities, but teams were free to reveal their label(s) in articles describing their systems if they wished.

Table 1 lists the baselines and participating teams, with basic information about their respective approaches and references to their system-description articles. All teams but one published an article about their system, and all of the published articles chose to reveal the label(s) of their submitted systems. We have therefore included that label information in Table 1.

5 Methods used by Challenge Entries

Based on the publications referenced in Table 1, we now review the technical approaches taken by the different teams in this challenge, and (in Section 5.2) contrast them against the five 2020 challenge entries. Note that we do not discuss the Murple AI Lab submission since that team did not submit a system-description article.

5.1 Motion-generation Approaches in 2022

Most of the teams proposed neural network-based methods for generating pose sequences. The RNN and Transformer architectures were the most common choices, which effectively led to smaller architectural differences between the systems. The GestureMaster team was unique among the teams in utilising motion matching [Büttner and Clavet 2015], which involves extracting and combining snippets of motion based on the training dataset, for their approach. Using ChoreoMaster [Chen et al. 2021] for dance as a starting point, they present a motion graph-based matching method for optimally selecting and combining gesture motion clips into a sequence, based on three criteria: rhythm, style, and the transition between consecutive clips [Zhou et al. 2022]. To generate the target style embeddings, they fed speech audio into a trained StyleGestures [Alexanderson et al. 2020] system.

The other approaches that employed neural networks presented various variations on input-feature context encoding and output-feature decoding for gesture generation.

To begin with, several approaches made use of custom representation learning for the different modalities. In particular, both the DeepMotion [Lu and Feng 2022] and ReprGesture teams [Yang et al. 2022] created pre-trained modality representations for their submissions. DeepMotion used a VQ-VAE to map motion data into a discrete space, whilst the ReprGesture team performed both modality-invariant and modality-specific representation learning, combining the two types of features for gesture generation. In a related move, the Forgerons team [Ghorbani et al. 2022] introduced a style-encoding component to learn to encode the style of an input motion.

Two submissions used sequence-to-sequence models with variants of neural attention mechanisms. The TransGesture submission [Kaneko et al. 2022] employed RNN-Transducers that only make use of past information during synthesis, meaning that the approach can be applied to streaming audio with no algorithmic latency. The IVI Lab submission [Chang et al. 2022] was based on the Tacotron 2 TTS architecture [Shen et al. 2018], but modified to use locally constrained attention when synchronising the motion with the input speech audio.

The DSI submission [Saleh 2022], similar to earlier work in autoregressive gesture-generation [Kucherenko et al. 2020], employed curriculum learning to reduce the error accumulation inherent in autoregressive generation.

Finally, two teams focussed on modifying the decoder. The FineMotion team [Korzun et al. 2022] proposed a linear layer-based decoder utilising the previous frame and speech context as input, instead of an RNN-based decoder, to improve motion stability between frames. The UEA Digital Humans submission [Windle et al. 2022] made use of a combination of decoders, where each individual decoder would generate parts of the resulting pose (face, upper-body, and hands).

In the next few subsections, we delve deeper into the representation and processing of the input and output data performed by various teams, seeing that these aspects can have a major impact on the results produced by data-driven synthesis methods. In Section 8.4.2—after reporting on the challenge results—we draw lessons from the system performance in relation to these aspects.

5.1.1 Input Modalities and Their Representation. As shown in Table 1, all submissions from participating teams used speech audio input, with some additionally employing speech text. Strictly speaking, speech text is not independent from speech audio, since speech audio carries all the information provided by the speech text. (Indeed, the text was derived by transcribing the audio recordings.) However, speech audio exposes rhythmic and paralinguistic information, whereas speech text offers a more direct representation of speech lexical content. This makes these different representations suitable for different tasks relevant to gesture generation [Kucherenko et al. 2022]. Hence, it is reasonable for a submission to utilise both audio and text. Additionally, to enable systems to capture individual differences in gesturing behaviour, a few teams made use of the provided speaker ID information as input.

For speech audio, most teams (4 out of 9) relied on **mel-frequency cepstrum coefficient** (**MFCC**) features [Davis and Mermelstein 1980]. Some used pre-trained off-the-shelf foundation models to represent the input modality. In particular, the

ReprGesture submission used WavLM [Chen et al. 2022] and the UEA submission used PASE+ [Ravanelli et al. 2020] for encoding the audio input. Among the submissions that used speech text as an input, most employed FastText to provide word embeddings [Bojanowski et al. 2017]. Four teams used speaker ID as input. Some used one-hot vectors of speaker ID as input features [Chang et al. 2022; Windle et al. 2022], whilst the Forgerons submission [Ghorbani et al. 2022] implemented methods for style control based on a given motion exemplar.

5.1.2 Output Motion Representation. For the challenge evaluation, teams had to generate BVH files, in the same format as used to distribute the dataset. Poses in these BVH files are represented by root-node positions and Euler angles for joint rotations. Due to discontinuities in Euler angles representations [Zhou et al. 2019], no team trained their neural networks to output Euler angles directly. Instead, the IVI Lab [Chang et al. 2022] and the TransGesture [Kaneko et al. 2022] submissions used exponential map representations for the output motion [Grassia 1998], whereas the DeepMotion [Lu and Feng 2022] and the UEA [Windle et al. 2022] submissions used a 6-dimensional representation [Zhou et al. 2019]. The DSI [Saleh 2022] and the ReprGesture [Yang et al. 2022] teams utilised rotation matrices, whilst the FineMotion [Korzun et al. 2022] team used an axis-angle representation. The Forgerons team [Ghorbani et al. 2022] employed a 2-axis rotation matrix [Zhang et al. 2018] to represent joint rotations and used a mixture of joint position, rotation, positional velocity, and rotational velocity as the output data of the gesture synthesis model.

5.1.3 Pre- and Post-processing. One popular strategy for training data pre-processing was to exclude segments where the character was not speaking, which in theory would make some methods produce better co-speech gesture models, as argued by the DeepMotion [Lu and Feng 2022] and Forgerons [Ghorbani et al. 2022] submissions. The ReprGesture team [Yang et al. 2022] decided to use data from only one speaker in the training set due to the potential interference caused by style diversity among speakers.

The most common post-processing technique among the submissions was to apply smoothing to the raw output motion. For systems that generate poses frame by frame, applying a smoothing filter often helps reduce visual artefacts in case of discontinuous, jerky, or jittery motion in the original model output.

5.2 Motion-generation Approaches in 2020

For the first GENEA Challenge [Kucherenko et al. 2021b], in 2020, five teams (Alexanderson [2020], Korzun et al. [2021], Lu et al. [2021], Pang et al. [2020], and Thangthai et al. [2021]) submitted entries to the crowd-sourced evaluation. There are some key differences between the systems that were entered into the first challenge and the more recent challenge. Some of these differences were due to the challenges setup. All submissions to the first challenge only supported gesture generation for one individual, since the dataset of that challenge only was sourced from one single person. In addition, the first challenge only considered upper-body motion.

All submissions to the first challenge used speech audio as input represented using MFCCs, with some teams also making use of the provided text transcriptions. The three submissions that made use of text transcripts used learnt embeddings like BERT [Devlin et al. 2018] and GloVe [Pennington et al. 2014] to extract and represent information from the text. As for the output poses, three teams used 3D joint rotations as the output features for the model to learn, with two teams instead relying on an exponential map representation. Two approaches relied on autoregressive architectures [Alexanderson 2020; Pang et al. 2020], whilst the other either relied on RNNs [Korzun et al. 2021] or an autoencoder [Lu et al. 2021; Thangthai et al. 2021]. Importantly, unlike 2022, none of the 2020 entries used motion graphs for output generation.

We will return to the different approaches taken by participating teams in 2020 and 2022 in Section 8.4, and relate the approaches to the outcomes of the respective challenges, after having reported on the 2022 challenge evaluation and its results.

6 Evaluation

We conducted a large-scale, crowdsourced, joint evaluation of gesture motion from the 10 full-body conditions and 11 upper-body conditions (listed in Table 1) in parallel using a within-subject design (i.e., every rater was exposed to and evaluated all conditions in each tier). The systems were evaluated in terms of the human-likeness of the gesture motion itself, as well as the appropriateness (a.k.a. specificity) of the gestures for the given input speech. The central difference from other gesture-generation evaluations is that all systems in our evaluation used the same motion data, the same visualisation/embodiment, and were rated together using the same evaluation methodology; only the motion-generation systems differed between the different entries that were compared. This allows the performance of systems to be compared directly, and the design aspects that influence performance can be traced more efficiently than in most previous publications. The subjective evaluation used an entirely crowdsourced approach, with attention checks used to exclude participants that were not paying attention, as detailed in Section 6.5. The remainder of this section describes the experiments we performed. Results of the subjective evaluation are subsequently presented in Section 7 and discussed in Section 8.

Although the aim of the challenge is to quantify how natural and appropriate motion appears to human observers, we have also seized the opportunity to compute a number of objective metrics of motion quality on the motion materials in the evaluation. The design of that experiment is described in Section 6.6, with results reported in Section 7.4 and discussed in Section 8.3. We see this primarily as an evaluation of the metrics themselves, and not as an evaluation of the different conditions in the challenge.

6.1 Subjective Evaluation Design Philosophy

For each tier, two different aspects of the generated gestures were evaluated (with one study per aspect and tier):

Human-likeness Whether the motion of the virtual character visually 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.

Human-likeness is thus a unimodal and unconditional quality measure (it only depends on the output motion), whereas speech appropriateness is multimodal and conditional on the speech. The former assesses system output quality whilst the latter assesses how well the output of the system relates to its input, disregarding the intrinsic quality of the output as much as possible. A deeper motivation for separating conditional and unconditional evaluation follows, with more details about the two different evaluations provided in Sections 6.3 and 6.4 further below.

6.1.1 *Why Separate Conditional and Unconditional Performance Measures?*

The complementarity of conditional and unconditional performance measures has long been recognised in other fields, and our decision to perform both unconditional (unimodal) and conditional (multimodal) subjective evaluations reflects a widespread distinction seen in both objective as well as subjective evaluation of synthesis methods in general. In image generation from text prompts, the FID [Heusel et al. 2017] is a widely used unconditional metric of synthesis quality. It does not take the input text into account at all, and would not notice disconnects between the input and output modalities, such as if the prompt “elephant” were to generate an image of an ant and vice versa. In contrast, multimodal CLIP embeddings [Radford et al. 2021] can assess the extent to which a synthetic image matches its corresponding text prompt, regardless of visual quality.

Similarly, our sister field of TTS distinguishes between the concepts of quality (or naturalness) and intelligibility, which are closely related to our respective constructs of human-likeness and appropriateness. Speech quality is usually assessed in a unimodal fashion (only involving audio), as reflected by evaluation standards such as ITU-T P.800 [International Telecommunication Union, Telecommunication Standardisation Sector 1996], whereas the most rigorous intelligibility evaluations are multimodal, in that they require audio to be transcribed to text (cf. King [2014]). Speech-synthesis literature also provides good support for the meaningfulness of separating the two aspects: signal quality can be high even if speech is unintelligible, as demonstrated by the unconditional, “babbling” WaveNet system in van den Oord et al. [2016], whereas rule-based speech synthesisers can achieve ceiling intelligibility even if their naturalness is poor [Malisz et al. 2019; Winters and Pisoni 2004]. A review of ten years of TTS challenges found that different technical approaches on average performed better on different measures (conditional vs. unconditional) and worse on others [King 2014], even though these trends were not obvious from individual years due to the large variation among challenge submissions.

More generally, the roles of conditional and unconditional performance measures, and their interplay with each other, were formalised in a domain-agnostic context by Blau and Michaeli [2018] as the *perception-distortion tradeoff.*

6.1.2 *Motion Aspects Deliberately Not Evaluated.*

Although an interesting question for a multispeaker dataset, we did not attempt to evaluate the appropriateness/specificity of the gesture motion style to different individuals in the database, since the data is too imbalanced to allow such an evaluation. Additionally, even though the speech and motion in the challenge come from joint full-body motion capture of dyadic interactions with separate close-talking microphones for each speaker, the challenge only considered generating one side of the conversation, without awareness of the other party in the interaction (neither for the synthesis, nor for the evaluation), in order to reduce problem complexity.

6.2 Stimuli

6.2.1 *Speech-segment Selection.*

The test set was deliberately made large to make it difficult to overfit to specific speech being evaluated. Like the GENEA Challenge 2020 and the Blizzard Challenges, not all test-set motion was included in the subjective evaluation. 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:

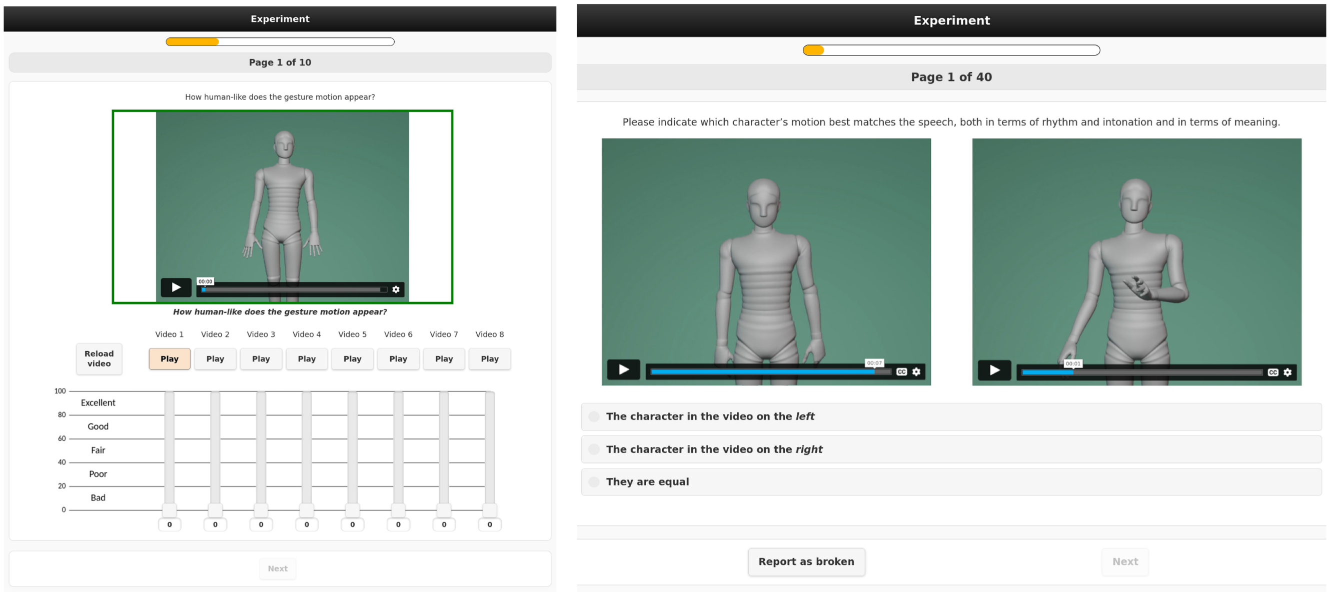
1. Segments should be around 8 to 10 seconds long, and ideally not shorter than 6 seconds.
2. The character should only be speaking, not passively listening, in the segments. (No turn-taking, but backchannels from the interlocutor were OK.)
3. Segments should not contain any parts where Lee et al. [2019] had replaced the speech with silence for anonymisation.
4. 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 ending on a “cliffhanger”.
5. The end of a segment should leave some margin until the chunk ends, to allow excerpting a longer segment if needed when creating mismatched stimuli as described in Section 6.4.2.
6. 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.

The last item does not imply that the motion capture was perfect or completely natural for all segments in the evaluation since the finger-tracking quality throughout the database does not allow our evaluations to reach that standard. It merely means that the level of finger-tracking quality in the stimuli was consistent with the better parts of the source material from Lee et al. [2019].

The 48 segments we selected were between 5.6 and 12.1 seconds in duration (average 9.5 seconds). Audio was loudness normalised to −23 dB LUFS following EBU R128 [European Broadcasting Union 2020] to maintain a consistent listening volume in the user studies.

6.2.2 *Visualisation.*

We used the same virtual avatar (shown in Figure 3) in all rendered videos during the challenge and the evaluation. The avatar had 56 joints (full body including fingers). Since the speech and motion presented to our test takers were sourced from multiple people, the avatar was designed to be a gender-neutral humanoid figure without the hallmarks of any particular individual. Instead of a fully realistic body shape and textures, a simplified design resembling a social robot was used. As the challenge did not encompass the generation of gaze information, lip motion, or facial expression, eyes and mouth were omitted from the avatar, to help evaluators instead focus on the motion of the rest of the body. The avatar is of sufficiently high polygon count so that its surface is shaded smoothly. All teams had



How human-like does the gesture motion appear?

How human-like does the gesture motion appear?

Please indicate which character's motion best matches the speech, both in terms of rhythm and intonation and in terms of meaning.

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

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

○ **They are equal**

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

(b) Appropriateness interface and upper-body videos

Fig. 3. Screenshots of the evaluation interfaces used in the studies, also showing the camera perspectives used by the two different tiers.

access to the official visualisation and rendering pipeline during the system-building phase, in the form of code, a portable Docker container, and a web server to which BVH files could be submitted to be rendered as video. Participants could send a 30-fps BVH file to the visualisation server, and these files would then be processed as quickly as possible into videos visualising the motion on the avatar. The visualisation server code has been open-sourced (see github.com/TeoNikolov/genea_visualizer). The code was available to the participants during the challenge, and they were free to use it to host their own servers if they wished. The final rendered stimuli used a resolution of 1440×1080.

6.3 Human-likeness Evaluation

The human-likeness evaluation of the GENEA Challenge 2022 closely followed the human-likeness evaluation in the GENEA Challenge 2020 [Kucherenko et al. 2021b]. Specifically, the evaluation was based on the **Human Evaluation of Multiple Videos in Parallel** (**HEMVIP**) methodology [Jonell et al. 2021], in which multiple motion examples (video stimuli) are presented on the same page (a.k.a. *screen*) and the subject is asked to provide a rating for each one before continuing. All stimulus videos on the same page correspond to the same speech segment but with different conditions. This property of the HEMVIP method brings two advantages, namely, (1) that any given study participant is always rating sets of stimuli for which the speech is the same and only the condition differs, which should make the numerical ratings and the relative condition ordering more consistent, and (2) 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 (as seen in the results in Section 7.1). For a detailed explanation of the evaluation interface, we refer the reader to Jonell et al. [2021], which introduced and validated the evaluation paradigm for gesture-motion stimuli. Code is provided at github.com/jonepatr/hemvip/tree/genea2022/.

6.3.1 *Evaluation Design.*

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 3(a). Note that by design only one video is visible at any given time; each play button corresponds to a distinct video stimulus, which is displayed when that button is clicked. Like in Jonell et al. [2021] and Kucherenko et al. [2021b], 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". These labels were based on those associated with the 5-point scale described in the Mean Opinion Score ITU-T P.800 standard [International Telecommunication Union, Telecommunication Standardisation Sector 1996] for audio quality evaluation.

6.3.2 *Motivation for the Use of Unimodal Stimuli.*

Although the videos on any given page in these human-likeness evaluations all corresponded to the same speech input and had the same length, the videos presented to participants were unimodal (motion-only), in that they were completely silent and did not include any audio. This ensures that ratings can only depend on the motion seen in the videos, and not on motion appropriateness for the speech since raters did not have any access to any speech information.

Minimising the influence of speech is important when rating the intrinsic human-likeness of gesture motion since speech and gesture perception are linked, to the extent that the same motion (possessing the same intrinsic human-likeness) may be perceived differently depending on what audio it is presented with. A classic example of this link is the McGurk effect [McGurk and MacDonald 1976], where the perceived identity of speech phonemes in a fixed audio stimulus changes if paired with video of specific facial articulation. This effect also extends to gesture perception [Bosker and Peeters 2021]. Conscious and unconscious human biases furthermore mean that raters may give lower or higher scores based on their perception of speaker traits such as likeability, gender, social status, and so on (e.g., Babel and Russell [2015] and Montgomery and Zhang [2018]), which would increase estimator variance and reduce statistical resolution. Removing speech content and only presenting the motion on a neutral avatar avoids these issues. More explicitly, Jonell et al. [2020a] found that including speech audio in a study of motion mimicry in annotated reference stimuli led to confounding, since subjects based their responses on speech semantic content rather than the relevant non-verbal interactions in the video modality. Consequently, our evaluation is unimodal and the task given to the raters is not conditional on the speech.

6.3.3 Evaluation Procedure. The test was preceded by a screen with instructions, which the participants would read. Then, each subject completed one training page showing a fixed set of videos with different motion, to familiarise participants with the task and what the stimuli would look like, before starting the study in earnest. The training phase was 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 such that each segment appeared on pages 1 through 10 with approximately equal frequency across all participants (segment order), and each condition was associated with each slider with approximately equal frequency across all pages (condition order). For any given participant and study, each of the 10 pages would use a different speech segment. Every page in the evaluation contained one stimulus video from condition FNA/UNA. This was used to help calibrate evaluators' ratings and keep them consistent throughout the test. Since motion-capture data projected onto a virtual character may not necessarily be perceived as perfectly natural, there was no requirement to rate the best motion as 100. After completing the rating pages, but before submitting the study, participants filled in a short questionnaire to gather broad, anonymous demographic information, the results of which are presented in Section 6.5.

6.4 Appropriateness Evaluation

The appropriateness evaluation was designed to assess the link between the gesture motion and the input speech, separate from the intrinsic human-likeness of the motion. It is thus inexorably multimodal, with user assessments of motion conditional on speech information provided to them.

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. In an attempt to control for the effect of motion quality in that evaluation, test takers were asked to ignore the motion quality and only rate the appropriateness of the motion for the speech [Kucherenko et al. 2021b]. Unfortunately, that evaluation was not altogether successful, since their *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 interaction between the perceived human-likeness of a motion segment and its perceived appropriateness for speech. That interaction acted as a confounder in their study, with the result that all submissions ranked below natural-looking motion unrelated to the speech, despite the latter being intended as a bottom line in terms of appropriateness.

6.4.1 Evaluation Design. For the GENEA Challenge 2022, we decided to evaluate motion appropriateness for speech in a different way. Our design goal for this challenge 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 motion mismatching like in Jonell et al. [2020a] (which studied facial motion rather than hand and body gestures) and used it within every condition, and not just for the recorded motion-capture data FNA/UNA.

On each page, subjects were presented with a pair of videos containing the same speech audio. Both videos contained motion from the same source—i.e., the same condition —and were thus expected to have similar motion quality and motion characteristics (at least on average), 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 [Wolfert et al. 2021] that this is more efficient in pairwise comparisons like these. A screenshot of the evaluation interface used for the appropriateness studies is presented in Figure 3(b).

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.

Rebol et al. [2021] used a similar methodology with preference tests to quantify the correlation (essentially, the appropriateness) between generated hand and body gestures and their associated speech, which we were not aware of until after conducting our challenge. However, they asked a different question of the users, did not quantify the appropriateness of real human motion, used data from monocular video rather than 3D motion capture (leading to noticeably lower data quality), and only used the approach to evaluate a single gesture-generation method.

Since speech audio has to be present in our appropriateness evaluation stimuli, test-taker perception may be subject to the human biases discussed in Section 6.3.2. However, the fact that the speech information (and the avatar used) is the exact same on both sides in every pairwise video presentation in the preference test should serve to control for the effect of these biases on user responses.

Table 2. Demographics of Test Takers from the Four User Studies

Study	Tier	Total number of test takers	Country of residence						Gender			Age (years)
			AUS	CAN	IE	NZ	UK	USA	F	M	X	mean ± std. dev.
Human-likeness	Full body	121	2	2	3	0	110	4	60	60	1	38 ± 12
	Upper body	150	1	0	4	0	134	11	74	75	1	40 ± 13
Appropriateness	Full body	247	3	13	10	2	211	8	137	107	3	38 ± 14
	Upper body	304	2	10	1	0	256	35	127	173	4	38 ± 13

Gender "F" stands for "female", "M" for "male", and "X" for "prefer not say".

6.4.2 Evaluation Procedure. Concretely, we created the mismatched stimuli by taking the 48 existing speech and motion segments selected for our evaluations, and permuting the motion in between them such that no motion segment ever remained in its original place. Each motion segment thus featured twice in the pairwise study: once with matched speech, and once with mismatched speech, where the corresponding matched stimulus would use another motion segment from the same source condition. 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. This was done by moving the endpoint of the mismatched motion segment such that the resulting motion duration exactly matched that of the new speech audio. The starting point of the mismatched motion video was never changed and was thus 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.

6.5 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 [Jonell et al. 2020b]. The challenge therefore adopted an entirely crowdsourced approach, as opposed to, for example, the Blizzard Challenge, which has used a mixed approach. Attention checks were used to exclude participants who were not paying attention. 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: (1) subjects were required to reside in any of six English-speaking countries, namely, Australia, Canada, Ireland, New Zealand, the United Kingdom, and the USA, and (2) 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. Which sliders on which pages that were used for attention checks was uniformly random, except that no page had more than one attention check, and the natural motion (condition FNA and UNA) was never replaced by an 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 3(b), forwarding them to the next pair of videos in the evaluation. Since reporting a video as broken avoids having to give a response, it can in theory be used to quickly skip through the test. To help prevent this, we implemented the button such that it becomes clickable after a 5-second delay after the page is loaded. However, as this does not fully prevent skipping through the test, subjects who used that button more than three times on pages without attention checks were also removed without pay. In all studies, the attention-check messages did not appear until a few seconds into each attention-check video, so that participants who only watched 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. Only the subjects who failed zero or one attention check for a study have been included in our analyses below. Responses to videos used for attention checks were not included in our analyses. 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 who successfully passed the attention checks and completed the full-body study and 150 test takers who did the same for the upper-body study. Of these, all passed all attention checks, except for one subject in the upper-body study, who failed one attention check. Since the upper-body study compared 11 conditions instead of only 10, it required more raters to reach the desired number of rating pairs. Table 2 provides demographic details of all subjects in the user studies.

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. All of these passed all attention checks, except for 10 participants in the full-body study and 14 participants in the upper-body study, who each failed one attention check. Demographic details are provided in Table 2. Each subject in a study contributed 36 ratings to the analyses after removing attention checks, unless they had to skip a page in the rare case of a video failing to load (which occurred approximately 1.6 times per 1,000 videos presented).

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 and also exceed the highest living wage quoted by the Living Wage Foundation in the UK at the time of the evaluation. All numbers are measured by Prolific, which uses the median rather than the mean for these calculations to prevent extreme completion times from skewing the data. Response data from the evaluation, together with statistical analysis code, is provided at doi.org/10.5281/zenodo.6939888.

6.6 Objective Metrics

The main goal of the GENEA challenge is to compare human subjective impressions of the outputs of different gesture-generation systems. We, therefore, discourage using the results of automated performance metrics as indicators of the perceptual impressions of different systems. However, since subjective evaluation is costly and time-consuming, it would be beneficial for the field to identify meaningful objective evaluation methods, especially for use during system development. As a step in this direction, we, therefore, considered five objective measures previously used to evaluate co-speech gestures, namely, average jerk, average acceleration, distance between gesture speed (i.e., absolute velocity) histograms, CCA, and the Fréchet distance between motion feature distributions. We computed these metrics for each condition in each tier using the complete test sequences, i.e., not only on the motion segments featured in the subjective evaluation. Details on each metric are provided below. The code for the numerical evaluations has been made publicly available to enhance reproducibility.[3]

To compare and validate these metrics against our subjective evaluation, we provide results on the rank correlations between subjective and objective metrics in Section 7.4.

6.6.1 Average Acceleration and Jerk. The third time derivative of the joint positions is called *jerk* and can be defined mathematically as $\text{jerk}(\boldsymbol{x}) = \boldsymbol{x}'''(t)$. The average value of the absolute magnitude of the jerk is commonly used to quantify motion smoothness [Kucherenko et al. 2019; Morasso 1981; Uno et al. 1989]. We report average values of absolute jerk (defined using finite differences) averaged across all test motion segments. A perfectly natural system should have an average jerk very similar to natural motion.

We also evaluated the same measure, but computed using the absolute value of the acceleration $\text{acc.}(x) = x''(t)$ instead of the jerk. Again, we expect natural-looking motion to have a similar average acceleration as in the reference data.

6.6.2 Comparing Speed Histograms. The distance between speed histograms has also been used to evaluate gesture quality [Kucherenko et al. 2019, 2020], since well-trained models should produce motion with similar properties to that of the actor it was trained on. In particular, it should have a similar motion-speed profile for any given joint. This metric is based on the assumption that synthesised motion should follow a speed distribution similar to the motion-capture data. To evaluate this similarity we calculate speed-distribution histograms for all systems and compare them to the speed distribution of natural motion (condition N) by computing the Hellinger distance [Nikulin 2001],

$$H(\boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)}) = \sqrt{1 - \sum_i \sqrt{h_i^{(1)} \cdot h_i^{(2)}}}, \quad (1)$$

between the histograms $\boldsymbol{h}^{(1)}$ and $\boldsymbol{h}^{(2)}$. Lower distance is better.

For both of the objective evaluations above the motion was first converted from joint angles to 3D coordinates.

6.6.3 CCA. CCA [Thompson 1984] is a form of linear subspace analysis and involves the projection of two sets of vectors (here the generated poses and those from FNA/UNA) onto a joint subspace. CCA has been used to evaluate gesture-generation models in previous work [Bozkurt et al. 2015; Lu et al. 2021; Sadoughi and Busso 2019].

The goal of CCA is to find a sequence of linear transformations of each variable set, such that the Pearson correlation between the transformed variables is maximised. This correlation is what we use as a similarity measure, and we report it as global CCA values in our results. A high value is considered better.

6.6.4 FGD. Recent work by Yoon et al. [2020] proposed the FGD to quantify the quality of generated gestures. This metric is based on the FID metric used in image-generation studies [Heusel et al. 2017] and can be written

$$\text{FGD}(X, \hat{X}) = ||\boldsymbol{\mu}_r - \boldsymbol{\mu}_g||^2 + \text{tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}). \quad (2)$$

Here, $\boldsymbol{\mu}_r$ and Σ_r are the first and second moments of the latent-feature distribution Z_r of the human motion-capture data X, whereas $\boldsymbol{\mu}_g$ and Σ_g are the first and second moments of the latent-feature distribution Z_g of the generated gestures $\hat{X}$. Z_r and Z_g were extracted by the same feature extractor, which was obtained as the encoder part of a motion-reconstructing autoencoder. We used a CNN-based autoencoder trained on the challenge dataset

[3] See github.com/genea-workshop/genea_numerical_evaluations

Table 3. Summary Statistics of Responses from All User Studies, with 95% Confidence Intervals

(a) Full-body

	Median	Appropriateness			
	human-	Num. responses			Percent matched
ID	likeness	M.	Tie	Mism.	(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

	Median	Appropriateness			
	human-	Num. responses			Percent matched
ID	likeness	M.	Tie	Mism.	(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]

"M." stands for "matched" and "Mism." for "mismatched". "Percent matched" identifies how often subjects preferred matched over mismatched motion. Human-likeness values are between 0 and 100; higher is better.

following the implementation in Yoon et al. [2020]. Lower values are better.

6.6.5 System Ranking Comparison. A good objective metric might help in evaluating the performance of a system, especially when such a metric correlates with a subjective measure. To get more insight into whether the objective metrics in our study may be used as a proxy for subjective evaluation results, we calculated the correlation between the ranking of the conditions on median human-likeness, and the result on the objective metrics listed above. For this, we used Kendall's τ rank correlation coefficient, and associated statistical tests [Kendall 1970].

Of the objective metrics we studied, only CCA compares output poses directly to the corresponding reference motion-capture poses. All other metrics are invariant to permutation, in the sense that changing the order of the different sequences (mismatching them with other speech/reference motion) will not change the value. They thus cannot measure appropriateness, which is why we only consider how those metrics correlate with human-likeness scores.

7 Results

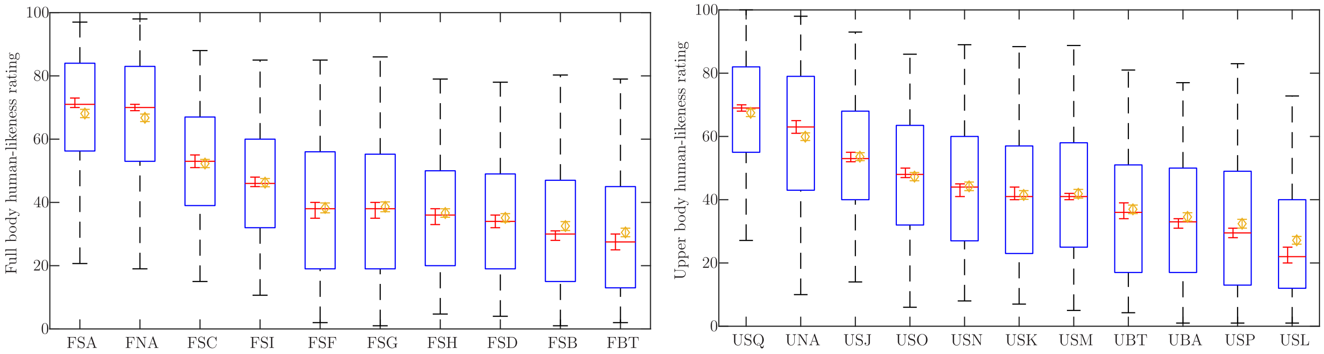
The results of the challenge are significant and thought-provoking. It is the first time that we find a system generating 3D gesture motion that exceeds the source data in terms of human-likeness, whilst simultaneously laying bare the extent of the gap between natural and synthetic gesture motion in terms of their appropriateness for speech. We furthermore find that all objective metrics except for the FGD correlate so poorly with subjective human-likeness scores as to be statistically indistinguishable from chance correlation. More detail is provided in the sections below, first reporting the results of the subjective evaluation and thereafter the objective metrics. Discussion of the various findings is reserved for Section 8.

7.1 Analysis and 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 4, with summary statistics (sample median and sample mean) for the ratings of all conditions in each of the two human-likeness studies given in the first half of Table 3, together with 95% confidence intervals for the true median. These confidence intervals were computed using order statistics, leveraging the binomial distribution cdf, while those for the mean used a Gaussian assumption (i.e., using Student's t-distribution cdf, rounded outward to ensure sufficient coverage); see Hahn and Meeker [1991]. We note that statistics regarding the mean should be interpreted with caution, since responses should be seen as ordinal rather than numerical, and it is therefore improper from a perceptual perspective to perform averaging on the ratings.

The distributions in Figure 4 are seen to be quite broad. This is common in evaluations like HEMVIP [Jonell et al. 2021], 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 judgements. 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.

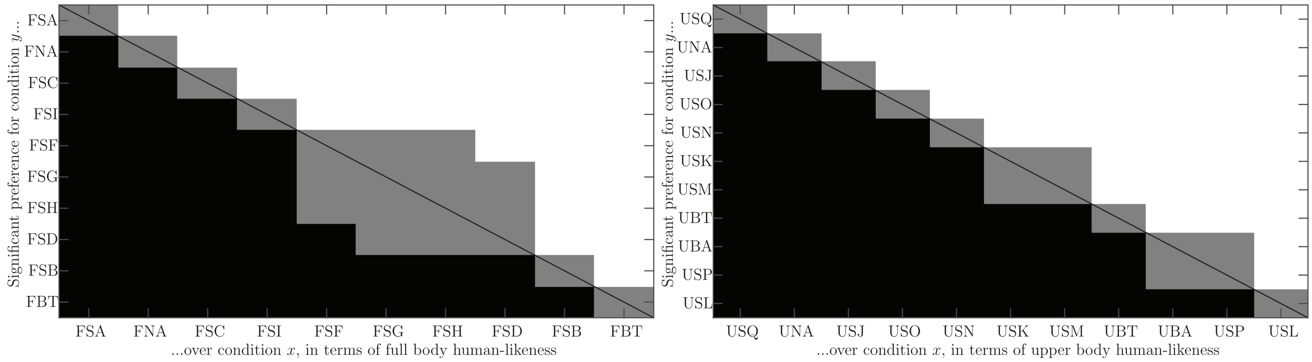
Despite the wide range of the distributions, the fact that the conditions were rated in parallel on each page enables using pairwise statistical tests to factor out many of the above sources of variation. To analyse the significance of differences in median rating between different conditions, we applied 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 [Kucherenko et al. 2021b].) This closely follows the analysis methodology used throughout recent Blizzard Challenges and, 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 and were assigned valid ratings were included in the analysis of significant differences. (Recall that not all conditions were rated on all pages due to the limited number of sliders and the



(a) Full-body

(b) Upper-body

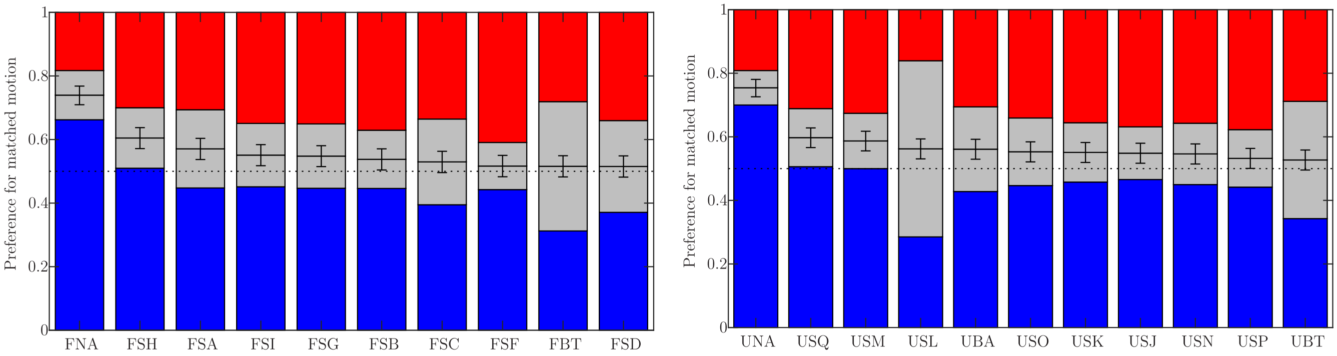
Fig. 4. Box plots visualising the ratings distribution in the human-likeness studies. Red bars are medians and yellow diamonds are means, each with a 0.05 confidence interval and a Gaussian assumption for the means. Box edges are at 25 and 75 percentiles, while whiskers cover 95% of all ratings for each condition. Conditions are ordered descending by sample median for each tier.



(a) Full-body

(b) Upper-body

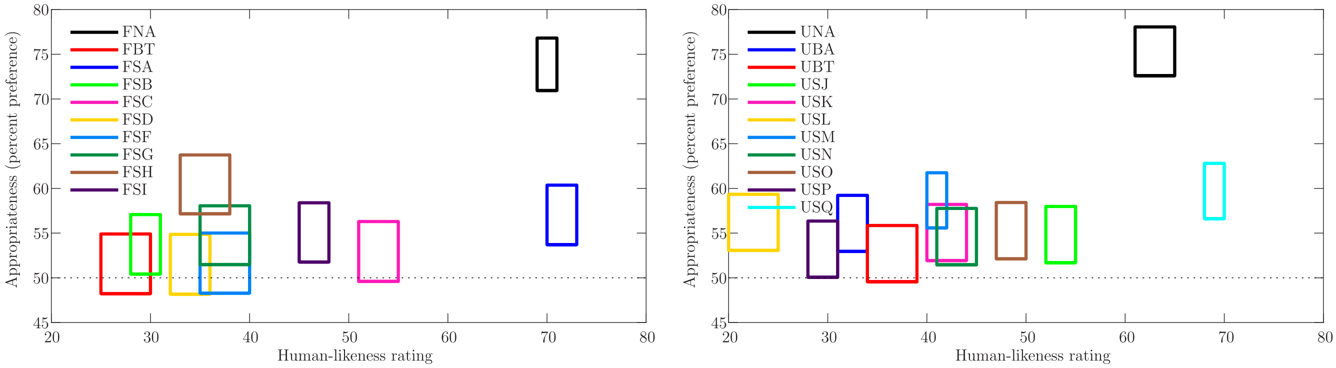
Fig. 5. Significant differences in human-likeness. White means the condition listed on the y-axis rated significantly above the condition on the x-axis, black means the opposite (y rated below x), and grey means no statistically significant difference at level $\alpha = 0.05$ after Holm–Bonferroni correction. Conditions use the same order as the corresponding subfigure in Figure 4.



(a) Full-body

(b) Upper-body

Fig. 6. 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

Fig. 7. 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.

presence of attention checks.) This meant that every statistical significance test was based on at least 615 pairs of valid ratings in the full-body study, and 603 pairs of valid ratings in the upper-body study. 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 4, 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 [Holm 1979], which is uniformly more powerful than conventional Bonferroni correction at keeping the **family-wise error rate** (**FWER**), often referred to as alpha-level, at or below $\alpha = 0.05$

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 in the two studies are visualised in Figure 5 which uses the same condition order as the box plot and shows which conditions were found to be rated significantly above or below which other conditions.

7.2 Analysis and 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. Every condition received at least 873 responses in the full-body study and 983 in the upper-body study. Raw response statistics for all conditions in each of the two studies are shown in the second half of Table 3, 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 confidence intervals were rounded outward to ensure sufficient coverage. The response distributions in the two studies are further visualised through bar plots in Figure 6, whilst Figure 7 visualises the results of the entire challenge in a single coordinate system per tier.

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 [Barnard 1945] to identify statistically significant differences at the level $\alpha = 0.05$ between all pairs of distinct conditions, applying the Holm-Bonferroni method [Holm 1979] to correct for multiple comparisons as before. (Here and forthwith, we only consider the relative preference in the sample after dividing ties equally.) Barnard's test is considered more appropriate than Fisher's exact test for a product of two independent binomial distributions [Lydersen et al. 2009], as here.

Our statistical 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. As before, the significant differences we identified in the two studies are visualised in Figure 5 which uses the same condition order as the box plot and shows which conditions were found to be rated significantly above or below which other conditions. No other pairwise differences were statistically significant in either study.

Instead of comparing the appropriateness of different synthesis approaches against one another, one may instead compare to 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, FNA, FSA, FSB, FSG, FSH, and FSI are significantly more appropriate than chance in the full-body study, and all conditions except UBT are more appropriate than chance in the upper-body study. Unlike other significance tests in the subjective evaluation, these assessments do not include a correction for multiple comparisons.

7.3 User Comments

As part of the post-evaluation questionnaire, we asked study participants to comment on the user studies, including positive and

Table 4. Objective Evaluation Results

(a) Full-body

Condition	Average jerk	Average accel.	Global CCA	Hellinger distance	FGD
FNA	**31,300 ± 6,590**	**798 ± 208**	**1**	**0**	**0**
FSA	14,600 ± 2,970	**668 ± 161**	0.849	**0.041**	**3.18**
FSC	5,130 ± 2,120	332 ± 129	0.818	0.125	16.4
FSI	7,370 ± 1,710	345 ± 98	0.789	0.111	**4.87**
FSF	**22,600 ± 6,240**	**666 ± 223**	**0.916**	0.195	7.49
FSG	5,560 ± 2,380	282 ± 127	**0.992**	**0.060**	10.1
FSH	8,630 ± 2,440	313 ± 92	**0.968**	0.104	**4.02**
FSD	8,690 ± 8,320	405 ± 257	0.886	0.132	43.4
FSB	**27,200 ± 4,680**	**628 ± 116**	0.782	**0.050**	16.3
FBT	3,510 ± 1,090	177 ± 56	0.738	0.267	28.6

(b) Upper-body

Condition	Average jerk	Average accel.	Global CCA	Hellinger distance	FGD
UNA	**33,000 ± 7,030**	**842 ± 222**	**1**	**0**	**0**
USQ	**15,400 ± 3,190**	**710 ± 173**	0.685	**0.043**	**2.84**
USJ	8,280 ± 1,460	375 ± 81	0.640	0.197	**4.83**
USO	5,450 ± 2,260	353 ± 138	0.812	0.129	16.4
USN	7,510 ± 3,400	384 ± 127	0.789	0.092	194
USK	8,180 ± 2,450	311 ± 99	**0.962**	0.137	15.5
USM	6,840 ± 3,200	385 ± 172	**0.991**	**0.039**	**2.17**
UBT	3,760 ± 1,170	190 ± 60	0.707	0.248	18.2
UBA	**18,000 ± 14,900**	513 ± 326	**0.964**	0.244	17.0
USP	**28,500 ± 4,960**	**661 ± 123**	0.769	**0.051**	18.0
USL	7,730 ± 5,420	258 ± 157	0.849	0.306	28.4

The word "acceleration" has been abbreviated to "accel."; ± shows the standard deviation per sequence. The best two or three numbers in each column, i.e., those closest to the numbers from the held-out motion-capture data (FNA/UNA, first row of values), are bold. Except for FNA/UNA, conditions (rows) are ordered by decreasing median human-likeness rating. Numbers have generally been rounded to three significant digits.

Table 5. Correlations Between Objective and Subjective Results)

(a) Full-body

Metric	Average	Average	Global		Hellinger	FGD
	jerk	accel.	CCA		distance	
Versus	Hum.	Hum.	Hum.	App.	Hum.	Hum.
τ	−0.09	−0.36	−0.36	−0.38	−0.36	−0.49
p-value	0.72	0.15	0.16	0.15	0.15	0.048

(b) Upper-body

Metric	Average	Average	Global		Hellinger	FGD
	jerk	accel.	CCA		distance	
Versus	Hum.	Hum.	Hum.	App.	Hum.	Hum.
τ	−0.11	−0.26	0.11	−0.49	−0.40	−0.51
p-value	0.64	0.27	0.64	0.041	0.085	0.029

Rank correlations (Kendall's τ) between the "error" in the objective metrics (how much each objective value differed from the reference FNA/UNA) and median human-likeness scores (here abbreviated "Hum.") or—only for CCA—the preference for matched motion after splitting ties (abbreviated "App."). A strong predictor of human scores will exhibit a τ-value close to negative unity combined with a low p-value.

negative aspects they perceived. 97% of the respondents in the user studies responded positively on whether the compensation was adequate. Additionally, they often commented positively on how interesting and engaging the study was.

We also asked participants regarding any negative aspects of the study. Here, 15% of the participants answered that they found repetitiveness a negative aspect of the study. Some users pointed out the lack of a proper human face on the humanoid and suggested incorporating that in future work. Others commented on the lack of real conversation and proposed to have the humanoid be part of an actual conversation. All responses to these questions can be found in our data release.

7.4 Objective Evaluation Results

The values of the objective metrics we computed are listed in Table 4. For each number in the table, we also calculated how much it differed from the corresponding value for the reference system (FNA/UNA), and then computed the rank correlation between the absolute value of these differences and the median human-likeness scores from the subjective evaluation. The idea is that systems exhibiting values closer to FNA/UNA should appear more human-like. The resulting rank correlations and p-values can be found in Table 5. For median human-likeness, we only found a statistically significant ($p < 0.05$) rank correlation with FGD, for both the full and upper-body tier (Kendall's $\tau = -0.49$ and -0.51, respectively). The negative sign is expected since a smaller difference from FNA/UNA should be associated with better-looking motion and higher human-likeness scores. Figure 8 visually compares the subjective human-likeness ratings and objective metric results.

CCA is the only metric we computed that can indicate appropriateness since it directly compares each generated sequence to the corresponding reference motion-capture poses. We, therefore, computed its rank correlations with the appropriateness data as well. Here we found a statistically significant effect ($\tau = -0.49$) for the upper-body tier, but not for the full body.

8 Discussion

We now discuss our results and how they may be interpreted, first for human-likeness (in Section 8.1), then for appropriateness (in Section 8.2), and then for the objective metrics (in Section 8.3). We connect our discussion of each part to the other evaluations we performed and to previous literature. Based on our findings, we then formulate a number of take-home messages regarding what matters most in gesture generation (in Section 8.4) and give examples of how materials from the challenge can be used by the field (in Section 8.5).

8.1 Discussion of Human-likeness Results

Generating convincing 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 forward compared to

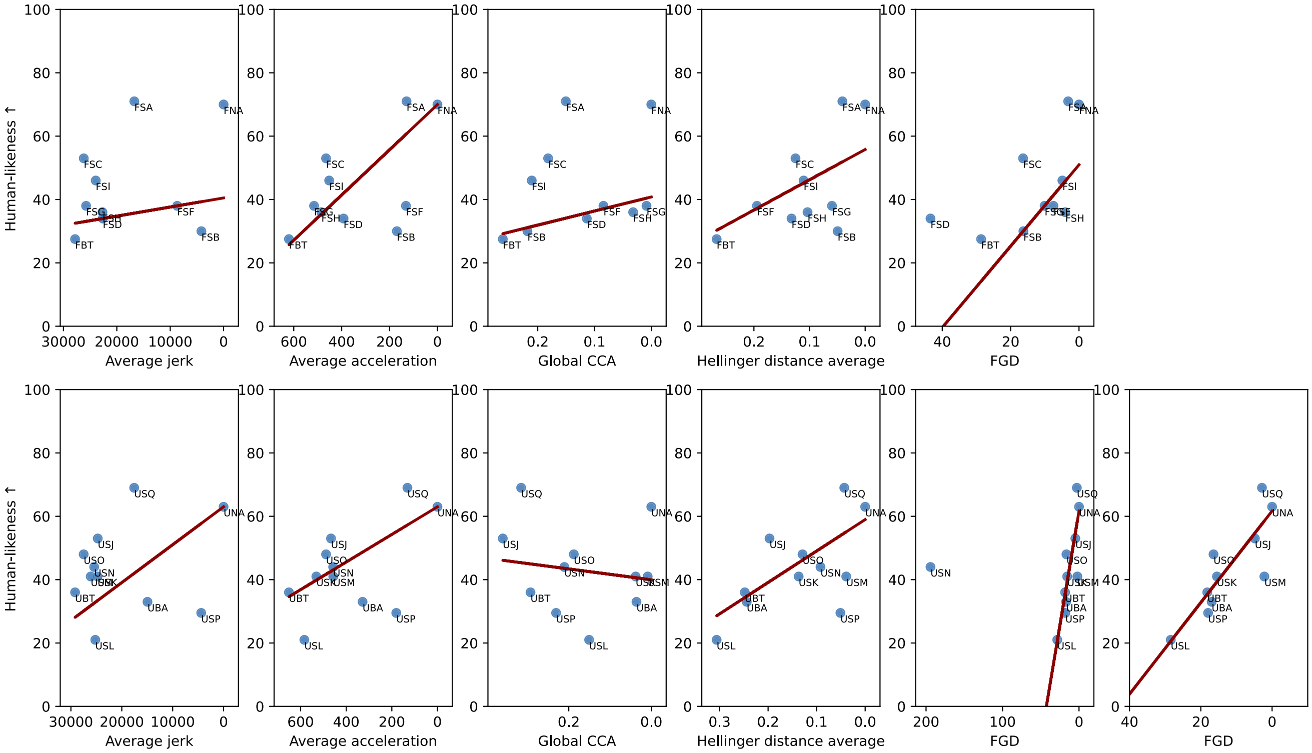


Fig. 8. Scatterplots comparing objective metrics and human-likeness ratings. The first row is for the full-body tier and the second row is for the upper-body tier. The x-axis shows the absolute magnitude of the difference between the objective value for each system and the corresponding value for the reference motion FNA/UNA, with the scale reversed such that the systems most similar to the reference are on the right. Regression lines (from the Theil–Sen regressor [Sen 1968; Theil 1992], which is robust to outliers) are also shown. The last plot in the second row is for FGD but with a narrower x-axis range for a better view.

GENEA 2020, and we believe it represents a motion quality not before seen in large-scale evaluations. Although there has been work, specifically Rebol et al. [2021], that reported a proposed motion-generation method as being statistically not significantly different from natural motion, they only evaluated a single method and their study was not based on motion-capture data but on 3D pose estimation from monocular video. We think that the choice of data source restricted the motion quality of their natural-motion condition to be less convincing (and thus easier to surpass) than our reference-motion conditions FNA/UNA. Furthermore, all differences between natural and synthetic conditions are significant in our study.

8.1.1 Interpreting the High Scores of FSA and USQ. Despite Zhou et al. [2022] (conditions FSA and USQ) being rated above the corresponding natural reference motion, we caution that this does not mean that the motion is “superhuman”, or even completely human-like—indeed, the median rating is much below 100, which would constitute “completely human-like” as per the instructions to test takers. What the result means is rather that the visualised motion in the majority of cases was perceived as more human-like 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. Most notably, the nominally natural motion is constrained by our ability to accurately capture the entire range of human motion, especially the fingers, using the technology we use. Finger motion capture is very difficult, and dataset limitations meant that the finger motion could not be chosen so as to look completely natural in all test segments evaluated, potentially degrading the ratings of FNA/UNA as a result. 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 FNA/UNA. This is in fact what was done for systems FSA and USQ, which only used selected training-data segments, manually chosen to have high motion quality, in generating their output gestures [Zhou et al. 2022].

Our ability to visualise human characters and their motion also plays a role in our findings. 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. If the challenge had involved generating additional modalities such as gaze and facial expression, the shortcomings of artificial systems may have become more clear, at the expense of increased complexity when running and taking part in the challenge.

8.1.2 On the Differences between the Two Tiers. There are fewer significant differences in the full-body evaluation than in the upper-body evaluation, perhaps meaning that full-body motion is more difficult to rate consistently. Although the difference is not substantial, we would naively expect the opposite, due to the correction for multiple comparisons being more conservative in the upper-body evaluation. There are many possible explanations for this finding, beyond the fact that the different teams did not all participate in both tiers. For example, our finding is consistent with an interpretation that full-body motion is a more difficult machine-learning problem, for instance due to the increased dimensionality of the output space and the increased number of behaviours that need to be learnt. This could explain why the best entry in the upper-body evaluation more clearly outperformed UNA, compared to the margin between the best entry in the full-body evaluation and FNA.

Another possible explanation for the same result is that the process of imposing full-body motion from a walking and talking human onto an avatar with a fixed lower body may not always yield completely natural results, and could sometimes give rise to incongruous motion. This could also explain the wider span (greater interquartile range) of ratings of UNA compared to FNA. Future GENEA challenges intend to only consider full-body motion.

8.2 Discussion of Appropriateness Results

We find the results of the appropriateness evaluation both thought-provoking and revealing about the state of the field. To begin with, the greatest relative preference, a 75% preference for matched motion, was observed for natural motion capture, i.e., FNA/UNA. This +25% effect size over the 50/50 bottom line validates that our methodology can well identify when gestures are appropriate for the speech and is about half the theoretical maximum value of +50% (a 100/0 split). A +25% effect size should be considered a good result, since previous studies that have incorporated mismatched stimuli, e.g., Jonell et al. [2020a] and Rebol et al. [2021], 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 (and do not, say, match audio of active speaking with a segment of motion corresponding to the character listening without speaking; cf. Wolfert et al. [2023]). Furthermore, both matched and mismatched motion stimuli here 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 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. It is therefore not surprising to find that the preference for matched motion over mismatched motion is not larger for FNA/UNA.

In line with expectations, 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. However, the synthetic conditions are all far behind natural human motion in terms of appropriateness. The measured effect sizes over the 50/50 bottom line range from +10% and down to 1.5% for all these conditions, compared to +25% for FNA/UNA, and all differences compared to FNA/UNA are highly statistically significant. This is a very substantial gap, and it is clear that generating meaningful and appropriate gestures is still far from a solved problem.

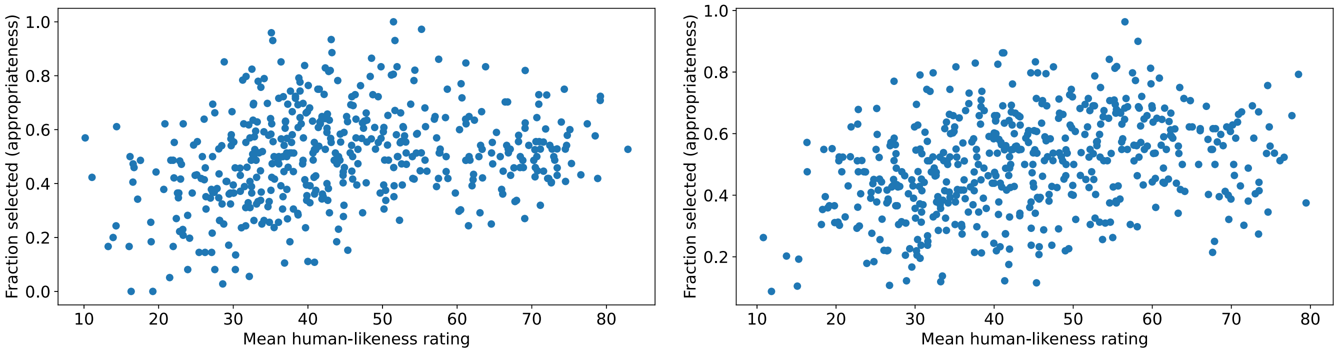
One other interesting trend is that a few conditions with relatively poor human-likeness, specifically FBT, UBT, and USL, show a noticeably larger fraction of tied responses, compared to other conditions. We hypothesise that this could be due to underarticulated motion, noting that a hypothetical, extremely underarticulated system that does not move at all should receive the response "They are equal" all the time. This hypothesis is consistent with the fact that these conditions all had the three lowest average acceleration values in Table 4, indicating little motion overall.

8.2.1 *Comparison to the Human-likeness Studies.*

Compared to the results for the human-likeness studies, we did not find as many differences between the submissions in terms of appropriateness. We can envision four factors that could contribute to this, which we list below, along with thoughts regarding potential mitigations:

- Responses are confined to much fewer categories, meaning that each response provides less information in an information-theoretic sense. This could potentially be addressed by having test-takers complement their response with an indication of the strength of their preference. We recommend that future developments in evaluation consider using a preference scale with more response options, e.g., five or seven possible responses. After the subjective evaluations described in this article concluded, such a scale was subsequently implemented by Mehta et al. [2023] and Kucherenko et al. [2023].
- Unlike the HEMVIP-based human-likeness studies, the responses to the appropriateness studies were not analysed using pairwise statistical tests to control for variation between subjects and stimuli. This might have led to reduced resolving power. It might be possible to improve on the statistical analysis using, e.g., statistical models that account for the effects of different test takers and different videos, or by changing the study setup to allow for pairwise statistical testing. One can furthermore gather more responses per condition, which we recommend in case the same statistical analysis methodology is used.
- Assessing appropriateness may be a more difficult task for humans than assessing human-likeness (where test takers assessed only motion in isolation, without any associated speech audio), meaning that there is more random variation in the responses relative to the human-likeness studies. In a signal-to-noise analogy, this means that the noise is higher. Mitigating this would probably require changing the evaluation and its task. For example, differences might become more obvious if segments were mismatched completely randomly, such that speech sometimes would be paired with motion from a segment where the character is not actively speaking, and vice versa (see Wolfert et al. [2023]), although doing so would essentially change the type of appropriateness that is being assessed.
- It may simply be that current artificial systems struggle to generate motion that is particularly appropriate to any specific input speech. In other words, in a signal-to-noise analogy, the signal is weaker. Consequently, there is less of a difference to be uncovered in the first place.

Although all of these factors may contribute to the results we observe, the big gap in effect size between natural motion capture and synthetic motion, and the fact that FNA/UNA were very significantly better than all other conditions, shows that our



(a) Full-body

(b) Upper-body

Fig. 9. Scatterplots illustrating the correlation between the rated mean human-likeness of a motion segment and how often that segment was selected by users in the appropriateness user study as being the more appropriate one—regardless of whether it was shown as the matched or mismatched motion. Each point represents a unique motion segment. Pearson correlation analyses reveal positive correlations of 0.26 (p-value < 0.01) and 0.32 (p-value < 0.01) for full-body and upper-body motion tiers, respectively.

methodology is sufficiently accurate to clearly resolve important differences between conditions. Coupled with the finding that FSA/USQ were significantly differently better than FNA/UNA when instead rating human-likeness, it is clear that our evaluations have managed to tap into and estimate different aspects of motion.

8.2.2 Analysis of Stimulus-level Correlation between Performance Measures. To quantify the degree of decoupling between the human-likeness ratings and appropriateness assessments, we ran a numerical analysis of the correlation at the stimulus (motion segment/excerpt) level. For each motion segment from the human-likeness study, we computed its (arithmetic) mean human-likeness rating, along with how often that motion segment was chosen as the more appropriate motion example when presented in the appropriateness study—regardless of whether it was the segment that matched the speech audio or not. Tied responses were split equally between both stimuli in the pair.

Scatterplots visualising the resulting data for the two tiers of the challenge can be found in Figure 9. Computing the Pearson correlation between the two quantities in the scatterplots (rated human likeness vs. empirical user preference in the appropriateness study) yielded a correlation of 0.26 (p-value <0.01) and 0.32 (p-value <0.01) for full-body and upper-body motion, respectively. The correlation analyses thus find that a higher mean human-likeness rating is on average associated with a greater probability of a segment being selected by users in the appropriateness study, regardless of whether or not it matched the speech.

However, although statistically significant, correlations were of moderate strength. To further control for the effect of human-likeness in the mismatching paradigm, we propose that future studies may (1) explicitly ask test takers to ignore visual motion quality in making their judgements (similar to the question formulation used in 2020 [Kucherenko et al. 2021b]), and/or (2) may choose to set up studies such that the matched and mismatched segments in each presented pairing have similar mean human-likeness ratings.

8.2.3 Comparison to Other Gesture-appropriateness Assessments. The distribution of the three different responses across the different conditions in Figure 6 is similar to that seen in the mismatching study reported in Jonell et al. [2020a], which used a similar methodology. On the other hand, we see fewer statistical differences compared to the appropriateness study in GENEA 2020 [Kucherenko et al. 2021b], which asked participants to rate the appropriateness of the stimuli on an absolute scale using HEMVIP. However, the ratings in that study were strongly biased towards conditions with high human-likeness, as discussed in Section 6.4. This is evidenced by the fact that mismatched natural motion (M) scored second best in terms of appropriateness there. In the new appropriateness evaluation paradigm, M would perform at chance rate by definition. Furthermore, in a segment-level re-analysis analogous to those in Section 8.2.2, the Pearson correlation between the 2020 mean human-likeness and mean appropriateness ratings is 0.51, which is both significantly different from zero (p-value $<$ 0.01) and numerically about twice as large as the correlations between human-likeness and appropriateness judgements in the 2022 evaluations. The reduced correlation in 2022 indicates that responses in the latest appropriateness studies are markedly less confounded by segment human-likeness, as was our goal. In effect, we traded the high-resolution, high-bias method from GENEA 2020 for a reduced-resolution, lower-bias method.

In addition to controlling for the effect of motion quality, our method for assessing appropriateness only requires comparing a system to itself. We believe this feature 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. Seeing that creating appropriate baseline systems is one of the sticking points both for carrying out research and for its subsequent assessment in peer review, this can be a major simplification compared to parallel methodologies like HEMVIP [Jonell et al. 2021] that involve simultaneously comparing and evaluating many different conditions against each other. Since responses in those studies are affected by what other videos are shown on the same page, their results thus cannot be directly compared unless stimuli or implementations of previous synthetic baseline conditions are included in the new study. Our recommendation for future research that uses the same methodology as this article is to report effect size and $\alpha = 0.05$

Clopper-Pearson confidence intervals similar to Table 3, to enable easy and accurate comparison between studies.

8.3 Discussion of Objective Metrics

The values of each of the six objective metrics in Table 4 span a wide range. From the acceleration and jerk values, we can observe that some systems, e.g., the text-based baselines from Yoon et al. [2019], exhibit much less movement than others. Unfortunately, most objective metrics are not well aligned with subjective human-likeness scores. In the full-body tier, one of the least human-like systems, FSB, received some of the best scores in terms of average absolute jerk, acceleration, and Hellinger distance. At the same time, one of the most human-like systems, FSC, is not in the top three according to any of the objective metrics used. In the upper-body tier, one of the least human-like systems, USP, was in the top three systems according to average jerk, acceleration, and Hellinger distance while one of the most human-like systems, USO, is not in the top three according to any of the objective metrics. The rank correlations in Table 5 make these observations more precise, by showing that most correlations are not statistically significantly different from zero. The one exception is the FGD. Although the correlations we found there are moderate (around -0.5) and system USN shows an outlying value, this metric might have some potential as an objective evaluation metric useful for faster evaluation in the development phase, although it is not clear how well it will resolve smaller differences between systems. Further development and validation of objective metrics would benefit the research community, as exemplified by a recent study that improved the FGD and generalised it to non-human motion [Maiorca et al. 2023].

As for speech appropriateness, only the CCA metric takes reference motion into account and thus has any possibility to measure this aspect. (None of the studied metrics explicitly considers information from the speech itself.) The CCA results are not clear-cut, but nonetheless somewhat encouraging, seeing that the systems with the best appropriateness (namely, FSH, USQ, and USM) also exhibit some of the highest CCA values, of 0.96 and above, and we found a statistically significant correlation for one of the tiers.

All in all, we want to emphasise that objective evaluation of generated gestures is still an open problem. Subjective evaluation, as used by this challenge, remains the gold standard for comparing gesture-generation models [Wolfert et al. 2022], and none of the objective evaluation metrics can replace subjective user studies.

8.4 Take-home Messages

In this section we combine salient points from our findings with information that the teams provided about their challenge entries, in order to see what we can learn about what aspects matters most in gesture-generation methods, data processing, and evaluation.

8.4.1 What Have We Learnt about the Gesture-generation Problem? The challenge results—with some entries performing very well in human-likeness, but none coming close to human-level appropriateness—indicate that generating random generic gesturing movements is much easier than tailoring gestures to fit the speech well. This could be due to the fact that there is a strong correlation between consecutive frames of movement, whilst the correlation between speech and gestures is relatively weak. One simple argument that the latter correlation is weak is that the same speech may be accompanied by different gestures, and the same gesture conversely may accompany different speech audio clips. Furthermore, fastText [Bojanowski et al. 2017], the most commonly used text representation among challenge entries, embeds each word individually regardless of context, and is far from the state-of-the-art in language modelling and text representation; cf. Wang et al. [2019]. (Indeed, it is arguably weaker than some of the text representations used in 2020 and referenced in Section 5.2.) This likely impeded the ability of the models to learn semantically appropriate gestures.

8.4.2 What Have We Learnt about Successful Gesture-generation Methods? Section 5 provides an overview of the submitted systems, and Table 1 lists all submissions with their corresponding system properties, sorting them according to their human-likeness scores. We can note that all systems except the text-based baseline used audio as an input modality, fewer systems used text, and even fewer used speaker IDs. There seems to be no clear indication that using any given combination of modalities necessarily gives better results than others, as some systems using only audio are on the top and others on the bottom of the list. However, the fact that so many of them did use audio input suggests a perception among teams that taking audio into account is important. This is reinforced by the finding that BT, the only system exclusively based on text, did not perform well in the subjective evaluations.

When it comes to the techniques used, RNNs were the most popular choice and used almost by all the systems, followed closely by auto-regression. Again, for most of these, there seems to be no strong indication that certain choices are necessarily better than others. This is not unexpected, given the many aspects and ways in which challenge submissions differ, all of which are likely to have affected the results in different ways. Even by aggregating results across multiple challenges, associating outcomes with individual design decisions remains difficult; cf. King [2014].

Our main and clear observation is that the state-of-the-art in human-likeness is not to use deep learning for everything (or at least not to generate the gesture poses), seeing that the most human-like system, GestureMaster, is based on motion graphs [Arikan and Forsyth 2002; Kovar et al. 2002; Lee et al. 2002] and a library of carefully selected high-quality motion segments. For methods that used deep learning to generate output poses (i.e., all other submissions), the two approaches demonstrating the greatest human-likeness in both tiers relied on probabilistic approaches (VAEs or VQ-VAEs) with stochastic output generation. This resembles the state of the previous challenge, where the entry with the greatest human-likeness [Alexanderson 2020] was based on pose sequences sampled from an autoregressive normalising flow, although the difference to the most human-like non-probabilistic submission [Korzun et al. 2021] was not statistically significant at the more stringent $\alpha = 0.01$ level [Kucherenko et al. 2021b].

Although perhaps initially surprising, the strong showing from a playback-based method echoes the long dominance of concatenative (i.e., exemplar-based) systems in terms of speech-synthesis naturalness [King 2014]. Furthermore, the use of motion graphs

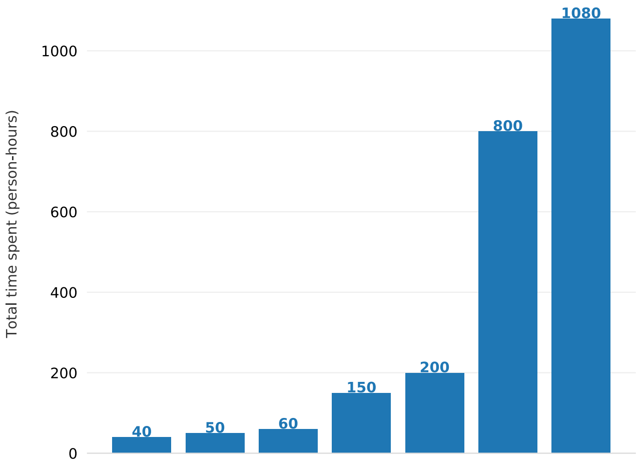


Fig. 10. The number of person hours different responding teams reported spending on the GENEA Challenge 2022, sorted in ascending order.

also resembles the prevailing approach to achieving high visual quality in 3D rendering, which is to combine small constituent images (bitmaps) as textures on a mesh; pure machine-learning approaches to rendering have taken a very long time to become competitive in terms of graphical quality [Mildenhall et al. 2021]. In general, it appears that – unless one has the methods and data needed to build an exceptionally strong deep generative model – an approach based on concatenating shards of real-world observations is instead the better path to achieving convincing results. In such an approach, it is feasible to ensure that all individual units hold high quality (they may for example be taken wholesale from the real world, making them completely natural by definition), leaving only the task of joining them together with minimal artefacts.

However, it is not only the best submissions that have gotten better in the recent challenge. The baselines from the 2020 challenge both performed relatively poorly in terms of human-likeness in 2022. Together, these findings offer evidence that the quality of gestures in the challenge as a whole is increasing, which may indicate that gesture-generation methodologies in general are getting better.

8.4.3 What Have We Learnt about Gesture-data Processing? Figure 10 shows how much time different teams spent on their submissions. We can see a very high variation, with some teams spending between 40 and 60 person hours whilst some others spent 800 hours or more. The two teams who spent 800 or more hours on their submissions reported devoting a large amount of time on data pre-processing, which other teams did not. One of the former teams is the top-performing team in terms of gesture human-likeness scores. This suggests that spending time on data preparation is likely to pay off in better model performance, which is consistent with trends from the Blizzard Challenges in TTS, where top teams often spend significant resources on manual data acquisition and processing. Data processing tasks included cropping the recordings into shorter segments and annotating those short segments for, e.g., motion quality, and similar. Some teams found it important to remove segments where the character was listening rather than talking since the character exhibits little gesture motion in these segments, which can make deterministic gesture-generation approaches regress towards the mean pose and thus produce less vivid movement.

There were variations in how challenge teams represented audio and motion in their entries, but we did not find strong evidence that certain representations were better than others. Systems that used modern, learnt audio representations such as WavLM [Chen et al. 2022] and PASE+ [Ravanelli et al. 2020] (not seen in the 2020 challenge) did not show superior performance compared to systems that used conventional MFCC audio features. For motion, submissions likewise exhibited a more diverse set of representation approaches than what was seen in 2020. In this challenge, there was a weak finding that systems using motion representations based on rotation matrices, including 6D representations [Zhou et al. 2019] or 2-axis rotation matrices [Zhang et al. 2018], obtained better human-likeness scores than systems that used exponential maps [Grassia 1998] of axis-angle representations. However, this finding is not conclusive due to the small number of examples and the multitude of factors affecting system performance, and might simply reflect the fact that systems with less time put into their development were more likely to use the data pipeline and motion representation of the provided baseline code, which used exponential map representations.

Another important aspect when it comes to the data is post-processing, such as hip-centering or smoothing (cf. Kucherenko et al. [2021a]), of the output motion. As seen in Table 1, most of the systems (good and bad performance alike) applied motion smoothing in some form. This suggests that they found smoothing to be beneficial for gesture generation, although the user studies do not allow us to make a statistical conclusion about the importance of smoothing the output motion.

Finally, modelling the motion of the fingers or having them fixed emerged as another important decision. Roughly half of the systems in the evaluation used fixed fingers. Some of these systems achieved good performance whilst others did not. This does not allow us to make strong statistical conclusions about the importance of modelling fingers. However, we may surmise that finger motion may be especially difficult to make natural, otherwise all teams would presumably have included finger motion in their submissions.

8.4.4 What Have We Learnt about Evaluating Gesture Generation? Previous work shows it is not easy to disentangle perceived human-likeness from appropriateness as more human-like systems are often ranked as more appropriate [Kucherenko et al. 2021b]. In this challenge, we made a concerted effort to disentangle these two aspects. Specifically, we (1) muted the audio during the human-likeness evaluation, to remove any influence speech may exert on perceived appropriateness (cf. Jonell et al. [2020a]), and (2) compared each model with a mismatched version of itself (having similar human-likeness), to control for the effect of human-likeness when evaluating appropriateness. This effort paid off, seeing that different conditions performed best on the two performance measures, with differences being statistically significant, whilst simultaneously ensuring that a speech-independent system (like condition M in 2020) can no longer score better than chance. However, improving the statistical resolution of the evaluation procedure would be beneficial.

8.5 How Materials from the Challenge Can be Used

We believe the materials released together with the challenge have many benefits for gesture-generation research. To illustrate this, we provide a list of possible use cases, often with references to prior work similarly that re-used resources from the previous GENEA Challenge (from 2020) in a similar manner. One may, for instance…

- Benchmark/compare new models to the state-of-the-art using our public data and existing motion or video stimuli, like Ferstl et al. [2021] and Yazdian et al. [2022] did with previous open stimuli.
- Evaluate models using our open-sourced code for the evaluation interface and analyses, similar to the re-use of HEMVIP code from Jonell et al. [2021] by Wolfert et al. [2021].
- Use our questions and evaluation structure for evaluating new proposed methods, similar to how Teshima et al. [2022] re-used previous evaluation designs.
- Use our public visualisation avatar and/or code to simplify development and obtain more standardised and comparable visuals, as done by Alexanderson et al. [2023], and similar to prior re-use of the GENEA Challenge 2020 open upper-body visualisation in Mehta et al. [2023], Saund and Marsella [2021], Teshima et al. [2022], Wang et al. [2021], and Zhang et al. [2023].
- Evaluate new models objectively using the same metrics that showed the most promise here, similar to how Ahuja et al. [2022], Liang et al. [2022], and Ye et al. [2022] re-used metrics and sometimes code from Ahuja et al. [2020] and Yoon et al. [2020].
- Use our large dataset of subjective evaluation responses to build and/or validate new automatic quality-assessment methods, similar to He [2022], or perform in-depth analyses of human preferences using the individual response data, perhaps linking these to the time taken by study participants, their questionnaire responses, and so on.
- Use our materials and those released by participating teams to probe reproducibility in the field.

9 Limitations

Despite being a large evaluation with many conditions and raters, there are inevitable limitations to the challenge and its results, imposed by scope, systems, data, visualisation, and evaluation choices. We discuss some of these limitations below.

9.1.1 Scope and Scale. The ten teams participating in the 2022 challenge do not represent the full spectrum of all gesture-generation approaches available today. Although ten teams (plus the top line and baselines) are more systems than considered in any other joint comparison of gesture-generation systems we are aware of, it is still not large enough to, e.g., make strong conclusions regarding which system architectures to prefer. We hope to attract more teams to participate in the challenge in future years.

9.1.2 Data. Motion capture is a remarkable technology, but it does not yet perfectly capture every aspect of human pose and figure. There are hardware issues such as calibration, and software challenges in estimating poses of diverse humanoid skeletons whilst dealing with problems like reflective marker displacements, occlusion, and markers in close proximity. Together, these issues may lead to problems with the data, commonly seen as artefacts in the produced motions (e.g., twitching or unnatural bone rotations), especially in the fingers. Although we have worked to exclude low-quality parts of the data and process it to make it more amenable to deep learning, some artefacts still remain. We suspect that this is an important reason why generated motion could surpass the notionally natural motion capture in terms of human-likeness. More, and more high-quality, motion data might allow for generating better gesture motion and constitute a stronger top line.

Some useful information is also missing from the current data. On the verbal side, this includes speech information removed for anonymisation. On the non-verbal side, one prominent missing aspect is facial data, which is an important communication channel but was not recorded in the current dataset. Neither was body form, such as muscle mass, body fat, and skin nor how these deform when muscles flex and extend, since the data has abstracted the humanoid form down to only a skeletal hierarchy. Future challenges should maintain awareness of new datasets being published, and their data quality and modalities captured. One modality worth investigating further is face motion, as it may help systems learn more appropriate gestures that relate to facial expressions and emotions.

9.1.3 Visualisation. The gesture visualisation used in the challenge has several limitations. Some are dictated by the data, and some are deliberate choices to, e.g., reduce complexity. The result is a virtual character that, whilst representative of typical gesture-generation visualisations, lacks both skin deformations and many human communication channels, such as gaze, facial expression, and lip motion. Whilst the absence of such features can help focus attention on the body motion currently being studied, it does also lead to a less human-like character appearance overall. Our evaluation also deliberately obscured some aspects of motion, e.g., by cropping the view so as to not show potential foot sliding and (for the upper-body tier) fixing the legs of the virtual character, which is innately unnatural. Future challenges should consider incorporating additional communication channels, e.g., facial features on a 3D mesh, to improve the realism of the virtual characters and their gestures.

Aside from limitations on what agent behaviours are visualised and how, the interlocutor from the recorded conversations is missing entirely in both modelling and visualisation. This was a deliberate choice to not increase the complexity of the challenge too much, but the absence of such information prevents us from assessing interlocutor-dependent aspects of motion such as proxemics and behavioural alignment. (We deliberately excluded turn taking, back channels, and listening behaviour from the subjective evaluation, since these are likely to look odd without seeing both sides of the conversation.) Future challenges may opt to include information about both conversation parties in the evaluation so that study participants can be interlocutor-aware in their responses. However, any increases in complexity, whether due to adding additional inputs or output modalities, should be performed one step at a time, so that it is more clear which findings relate to which aspect of the complex problem that is gesture generation.

9.1.4 Evaluation. Our core evaluation only sought to quantify two performance measures, namely subjective human-likeness and perceived appropriateness for the given speech. Aspects such as gesture diversity, or generation speed and latency, were not measured. Furthermore, we only studied the overall appropriateness of the gestures for the speech, but there is value in evaluating appropriateness with respect to the speech rhythm and speech meaning separately, since these are distinct aspects. We hope to consider doing that in future challenges, for example by performing two user studies, each focused on a separate type of appropriateness: semantic appropriateness and rhythmic/temporal appropriateness. A further extension would be to break this down into individual gesture categories, e.g., beat, iconic, deictic, and metaphoric gestures.

There are also many other kinds of appropriateness that can be assessed, e.g., appropriateness for the given speaker, and for the interlocutor behaviour as discussed above. (See the discussion of *grounding* in Nyatsanga et al. [2023] for a more extensive list.) None of these were considered in the present challenge, either due to dataset limitations or to keep the complexity to a manageable level. A difficult but important long-term goal is to pursue a more "ecologically valid" evaluation, to eventually compare different gesture-generation methods in human interaction, similar to He et al. [2022].

10 Conclusions and Implications

We have hosted the GENEA Challenge 2022 to compare many different gesture-generation methods and assess the state-of-the-art in data-driven co-speech gesture generation for full-body and upper-body avatars. The central design goals of the challenge were (1) to enable direct comparison between many different gesture-generation methods whilst controlling for factors of variation external to the model, namely, data, embodiment, and evaluation methodology, and (2) to disentangle the effects of motion human-likeness and motion appropriateness in the evaluations.

Our evaluation results show that, with the right approach, synthetic motion can attain human-likeness ratings equal to or better than the underlying motion-capture data. This is a big step forward, although most systems did not come close to this level of performance. The results also suggest that the field is advancing measurably since most submissions performed significantly better than the previously published baseline methods. However, using a careful 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 are thus only at the beginning of the road when it comes to generating co-speech motion that is appropriate for the specific speech. Finally, most objective metrics we computed did not exhibit any statistically significant correlations with our subjective human-likeness ratings, with the FGD being the lone exception to the rule. Objective metrics should thus only be used with great caution.

10.1 Implications

The challenge findings have implications for both research and practice. We summarise our perspectives below.

10.1.1 Implications for Practical Systems. If you are building a gesture-generation system and want to reach top-of-the-line human-likeness, you should currently consider using "playback-based" methods like motion graphs [Arikan and Forsyth 2002; Kovar et al. 2002; Lee et al. 2002] as demonstrated by GestureMaster [Zhou et al. 2022] to generate the pose sequences, instead of relying solely on deep learning to go all the way from input features to motion. Playback-based systems need less data, and the quality of the motion material is then a higher priority than database size, in contrast to current deep-learning trends. Machine learning is still useful for deciding which gestures to generate (e.g., which motion clips to concatenate). In all cases, it appears important to spend time on data processing.

10.1.2 Implications for Research and Evaluation. We believe the challenge adds value to the research community in several ways. A lot can doubtlessly be learnt from the system-description articles 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) have broad utility for future research, system building, and benchmarking in gesture generation, similar to the community uptake of the resources from the GENEA Challenge 2020. Furthermore, the methodology we demonstrate for assessing motion appropriateness for speech is much more accurate at controlling for the effect of subjective motion quality and does not involve subjects making any direct comparisons between videos generated by different conditions, which is beneficial for efficient benchmarking against previous publications (see Sections 8.2.1 and 8.2.2 for further details and recommendations).

10.1.3 Implications for Future Developments in the Field. 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 (at least ones based on motion playback) should be able to advance to more consistently match, or possibly even exceed, motion capture in terms of human-likeness. Systems that generate poses directly from deep learning are likely to improve in human-likeness as well, as larger datasets with more accurate motion become available (e.g., Liu et al. [2022b]). This would be similar to recent developments in verbal behaviour generation, where neural language models [Brown et al. 2020] and speech synthesisers [Li et al. 2019; Shen et al. 2018] trained on large datasets are approaching the text and speech produced by humans in terms of surface quality (but not necessarily appropriateness). Gesture generation may be lagging behind due to the relative scarcity of high-quality 3D motion data, compared to text and audio, since accurate motion estimation from monocular in-the-wild video remains a challenging problem.

As the above 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 what specific properties are appropriate for each individual gesture instance to be generated, even for the hard case of semantically motivated, communicative gestures from speech [Ao et al. 2022; Kucherenko et al. 2021c, 2022]. We also believe that there is great potential for devising better objective metrics, using challenge materials to validate these, and that the adoption of meaningful and validated objective metrics may further accelerate progress in the field.

10.1.4 Implications for Future Challenges. 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 Jonell et al. [2020a] and Woo [2021]. Generating interlocutor-aware full-body gestures was, therefore, a focus of the GENEA Challenge 2023 Kucherenko et al. [2023]. This should be coupled with further method development to obtain methodologies for conducting and analysing appropriateness tests with increased resolving power whilst still controlling for motion human-likeness. 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 towards 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, Tony Belpaeme, Jieyeon Woo, and Rajmund Nagy for feedback on the manuscript; and Jim Royal for video production.

References

Chaitanya Ahuja, Dong Won Lee, Ryo Ishii, and Louis-Philippe Morency. 2020. No gestures left behind: Learning relationships between spoken language and freeform gestures. In *Findings of the Association for Computational Linguistics (EMNLP'20 Findings)*. 1884–1895. DOI: https://doi.org/10.18653/v1/2020.findings-emnlp.170

Chaitanya Ahuja, Dong Won Lee, and Louis-Philippe Morency. 2022. Low-resource adaptation for personalized co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*. 20566–20576. DOI: https://doi.org/10.1109/CVPR52688.2022.01991

Simon Alexanderson. 2020. The StyleGestures entry to the GENEA challenge 2020. In *Proceedings of the GENEA Workshop (GENEA'20)*. DOI: https://doi.org/10.5281/zenodo.4088599

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

Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. 2023. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics* 42, 4 (2023), 1–20. DOI: https://doi.org/10.1145/3592458

Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. 2022. Rhythmic Gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM Transactions on Graphics* 41, 6, Article 209 (2022), 19 pages. DOI: https://doi.org/10.1145/3550454.3555435

Okan Arikan and David A. Forsyth. 2002. Interactive motion generation from examples. *ACM Transactions on Graphics* 21, 3 (2002), 483–490. DOI: https://doi.org/10.1145/566570.566606

Molly Babel and Jamie Russell. 2015. Expectations and speech intelligibility. *Journal of the Acoustical Society of America* 137, 5 (2015), 2823–2833. DOI: https://doi.org/10.1121/1.4919317

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

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

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)*. Retrieved from https://pub.uni-bielefeld.de/record/2392953

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)*. 76–89. DOI: https://doi.org/10.1007/978-3-642-04380-2_12

Kirsten Bergmann, Stefan Kopp, and Friederike Eyssel. 2010. Individualized gesturing outperforms average gesturing – Evaluating gesture production in virtual humans. In *Proceedings of the International Conference on Intelligent Virtual Agents (ICA'10)*. 104–117. DOI: https://doi.org/10.1007/978-3-642-15892-6_11

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)*. 2027–2036. DOI: https://doi.org/10.1145/3474085.3475223

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)*. 77–80. DOI: https://doi.org/10.21437/Interspeech.2005-72

Yochai Blau and Tomer Michaeli. 2018. The perception-distortion tradeoff. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'18)*. 6228–6237. DOI: https://doi.org/10.1109/CVPR.2018.00652

Piotr Bojanowski, Édouard Grave, Armand Joulin, and Tomáš Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146. DOI: https://doi.org/10.1162/tacl_a_00051

Hans Rutger Bosker and David Peeters. 2021. Beat gestures influence which speech sounds you hear. *Proceedings of the Royal Society B* 288, 1943 (2021), 20202419. DOI: https://doi.org/10.1098/rspb.2020.2419

Elif Bozkurt, Engin Erzin, and Yücel Yemez. 2015. Affect-expressive hand gestures synthesis and animation. In *Proceedings of the International Conference on Multimedia and Expo (ICME'15)*. 1–6. DOI: https://doi.org/10.1109/ICME.2015.7177478

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 *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS'20)*. 1877–1901. Retrieved from https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Michael Büttner and Simon Clavet. 2015. Motion matching – the road to next gen animation. In *Proceedings of the Nucl.ai*. Retrieved from https://youtu.be/z_wpgHFSWss

Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2001. BEAT: The behavior expression animation toolkit. In *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'01)*. 477–486. DOI: https://doi.org/10.1145/383259.383315

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)*. 784–789. DOI: https://doi.org/10.1145/3536221.3558060

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)*. 1564–1568. DOI: https://doi.org/10.21437/Interspeech.2013-395

Kang Chen, Zhipeng Tan, Jin Lei, Song-Hai Zhang, Yuan-Chen Guo, Weidong Zhang, and Shi-Min Hu. 2021. ChoreoMaster: Choreography-oriented music-driven dance synthesis. *ACM Transactions on Graphics* 40, 4, Article 145 (2021), 13 pages. DOI: https://doi.org/10.1145/3450626.3459932

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing* 16, 6 (2022), 1505–1518. DOI: https://doi.org/10.1109/JSTSP.2022.3188113

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)*. 152–166. DOI: https://doi.org/10.1007/978-3-319-21996-7_17

Steven B. Davis and Paul Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28, 4 (1980), 357–366. DOI: https://doi.org/10.1109/TASSP.1980.1163420

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)*. 4171–4186. DOI : https://doi.org/10.18653/v1/N19-1423

European Broadcasting Union. 2020. Loudness Normalisation and Permitted Maximum Level of Audio Signals. EBU Recommendation EBU R 128v4. Retrieved 11 June 2024 from https://tech.ebu.ch/docs/r/r128.pdf

Ylva Ferstl and Rachel McDonnell. 2018. Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the ACM International Conference on Intelligent Virtual Agents (IVA'18)*. 93–98. DOI : https://doi.org/10.1145/3267851.3267898

Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2021. ExpressGesture: Expressive gesture generation from speech through database matching. *Computer Animation and Virtual Worlds* 32, 3–4 (2021), e2016. DOI : https://doi.org/10.1002/cav.2016

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)*. 778–783. DOI : https://doi.org/10.1145/3536221.3558068

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)*. 1551–1555. DOI : https://doi.org/10.21437/Interspeech.2019-1783

F. Sebastian Grassia. 1998. Practical parameterization of rotations using the exponential map. *Journal of Graphics Tools* 3, 3 (1998), 29–48. DOI : https://doi.org/10.1080/10867651.1998.10487493

Gerald J. Hahn and William Q. Meeker. 1991. *Statistical Intervals: A Guide for Practitioners*. John Wiley and Sons. DOI : https://doi.org/10.1002/9780470316771

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 ACM International Conference on Intelligent Virtual Agents (IVA'22)*. Article 8, 8 pages. DOI : https://doi.org/10.1145/3514197.3549697

Zhiyuan He. 2022. Automatic quality assessment of speech-driven synthesized gestures. *International Journal of Computer Games Technology* 2022, Article 1828293 (2022), 11 pages. DOI : https://doi.org/10.1155/2022/1828293

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS'17)*. Retrieved from https://proceedings.neurips.cc/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf

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

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6, 2 (1979), 65–70. Retrieved from https://www.jstor.org/stable/4615733

Wen Chin Huang, Erica Cooper, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi. 2022. The VoiceMOS Challenge 2022. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech'22)*. 4536–4540. DOI : https://doi.org/10.21437/Interspeech.2022-970

International Telecommunication Union, Telecommunication Standardisation Sector. 1996. *Methods for Subjective Determination of Transmission Quality*. Recommendation ITU-T P.800. Retrieved from https://www.itu.int/rec/T-REC-P.800-199608-I

Carlos T. Ishi, Daichi Machiyashiki, Ryusuke Mikata, and Hiroshi Ishiguro. 2018. A speech-driven hand gesture generation method and evaluation in android robots. *IEEE Robotics and Automation Letters* 3, 4 (2018), 3757–3764. DOI : https://doi.org/10.1109/LRA.2018.2856281

Ryo Ishii, Taichi Katayama, Ryuichiro Higashinaka, and Junji Tomita. 2018. Generating body motions using spoken language in dialogue. In *Proceedings of the International Conference on Intelligent Virtual Agents (IVA'18)*. 87–92. DOI : https://doi.org/10.1145/3267851.3267866

Patrik Jonell, Taras Kucherenko, Gustav Eje Henter, and Jonas Beskow. 2020a. 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)*. Article 31, 8 pages. DOI : https://doi.org/10.1145/3383652.3423911

Patrik Jonell, Taras Kucherenko, Ilaria Torre, and Jonas Beskow. 2020b. 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)*. Article 30, 8 pages. DOI : https://doi.org/10.1145/3383652.3423860

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)*. 707–711. DOI : https://doi.org/10.1145/3462244.3479957

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)*. 753–757. DOI : https://doi.org/10.1145/3536221.3558061

Maurice G. Kendall. 1970. *Rank Correlation Methods* (4 ed.). Charles Griffin and Co.

Simon King. 2014. Measuring a decade of progress in text-to-speech. *Loquens* 1, 1, Article e006 (2014), 12 pages. DOI : https://doi.org/10.3989/loquens.2014.006

Vladislav Korzun, Anna Beloborodva, and Arkady Ilin. 2022. ReCell: Replicating recurrent cell for auto-regressive pose generation. In *Proceedings of the Companion publication of the ACM International Conference on Multimodal Interaction (ICMI'22 Companion)*. 94–97. DOI : https://doi.org/10.1145/3536220.3558801

Vladislav Korzun, Ilya Dimov, and Andrey Zharkov. 2021. Audio and text-driven approach for conversational gestures generation. In *Proceedings of the Computational Linguistics and Intellectual Technologies (DIALOGUE'21)*. DOI : https://doi.org/10.28995/2075-7182-2021-20-425-432

Lucas Kovar, Michael Gleicher, and Frédéric Pighin. 2002. Motion graphs. *ACM Transactions on Graphics* 21, 3 (2002), 473–482. DOI : https://doi.org/10.1145/566654.566605

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)*. 97–104. DOI : https://doi.org/10.1145/3308532.3329472

Taras Kucherenko, Dai Hasegawa, Naoshi Kaneko, Gustav Eje Henter, and Hedvig Kjellström. 2021a. Moving fast and slow: Analysis of representations and post-processing in speech-driven automatic gesture generation. *International Journal of HumanComputer Interaction* 37, 14 (2021), 1300–1316. https://doi.org/10.1080/10447318.2021.1883883

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)*. 242–250. DOI : https://doi.org/10.1145/3382507.3418815

Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. 2021b. 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)*. 11–21. DOI : https://doi.org/10.1145/3397481.3450692

Taras Kucherenko, Rajmund Nagy, Patrik Jonell, Michael Neff, Hedvig Kjellström, and Gustav Eje Henter. 2021c. 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)*. 145–147. DOI : https://doi.org/10.1145/3472306.3478333

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

Taras Kucherenko, Rajmund Nagy, Youngwoo Yoon, Jieyeon Woo, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2023. The GENEA Challenge 2023: A large-scale evaluation of gesture generation models in monadic and dyadic settings. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI'23)*. 792–801. DOI : https://doi.org/10.1145/3577190.3616120

Quoc Anh Le and Catherine Pelachaud. 2012. Evaluating an Expressive Gesture Model for a Humanoid Robot: Experimental Results. Retrieved 11 June 2024 from https://www.researchgate.net/publication/268257868_Evaluating_an_Expressive_Gesture_Model_for_a_Humanoid_Robot_Experimental_Results

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)*. 763–772. DOI : https://doi.org/10.1109/ICCV.2019.00085

Jehee Lee, Jinxiang Chai, Paul S. A. Reitsma, Jessica K. Hodgins, and Nancy S. Pollard. 2002. Interactive control of avatars animated with human motion data. *ACM Transactions on Graphics* 21, 3 (2002), 491–500. DOI : https://doi.org/10.1145/566654.566607

Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun. 2010. Gesture controllers. *ACM Transactions on Graphics* 29, 4, Article 124 (2010), 11 pages. DOI : https://doi.org/10.1145/1778765.1778861

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. DOI : https://doi.org/10.1609/aaai.v33i01.33016706

Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. 2021. AI Choreographer: Music conditioned 3D dance generation with AIST++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'21)*. 13401–13412. DOI : https://doi.org/10.1109/ICCV48922.2021.01315

Yuanzhi Liang, Qianyu Feng, Linchao Zhu, Li Hu, Pan Pan, and Yi Yang. 2022. SEEG: Semantic energized co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*. 10473–10482. DOI: https://doi.org/10.1109/CVPR52688.2022.01022

Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022b. BEAT: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV'22)*. 612–630. DOI: https://doi.org/10.1007/978-3-031-20071-7_36

Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. 2022a. Learning hierarchical cross-modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*. 10462–10472. DOI: https://doi.org/10.1109/CVPR52688.2022.01021

Yu Liu, Gelareh Mohammadi, Yang Song, and Wafa Johal. 2021. Speech-based gesture generation for robots and embodied agents: A scoping review. In *Proceedings of the International Conference on Human-Agent Interaction (HAI'21)*. 31–38. DOI: https://doi.org/10.1145/3472307.3484167

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)*. 900–904. DOI: https://doi.org/10.1109/ICASSP39728.2021.9414660

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, 790–796. DOI: https://doi.org/10.1145/3536221.3558059

Stian Lydersen, Morten W. Fagerland, and Petter Laake. 2009. Recommended tests for association in 2×2 tables. *Statistics in Medicine* 28, 7 (2009), 1159–1175. DOI: https://doi.org/10.1002/sim.3531

Antoine Maiorca, Hugo Bohy, Youngwoo Yoon, and Thierry Dutoit. 2023. Objective evaluation metric for motion generative models: Validating Fréchet motion distance on foot skating and over-smoothing artifacts. In *Proceedings of the ACM SIGGRAPH Conference on Motion, Interaction and Games (MIG'23)*. Article 2, 11 pages. DOI: https://doi.org/10.1145/3623264.3624443

Zofia Malisz, Gustav Eje Henter, Cassia Valentini-Botinhao, Oliver Watts, Jonas Beskow, and Joakim Gustafson. 2019. Modern speech synthesis for phonetic sciences: A discussion and an evaluation. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS'19)*. 487–491. DOI: https://doi.org/10.31234/osf.io/dxvhc

Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264, 5588 (1976), 746–748. DOI: https://doi.org/10.1038/264746a0

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

Shivam Mehta, Siyang Wang, Simon Alexanderson, Jonas Beskow, Éva Székely, and Gustav Eje Henter. 2023. Diff-TTSG: Denoising probabilistic integrated speech and gesture synthesis. In *Proceedings of the ISCA Speech Synthesis Workshop (SSW'23)*. Retrieved from https://openreview.net/forum?id=PCZ16_vl_ee

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2021. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM* 65, 1 (2021), 99–106. DOI: https://doi.org/10.1145/3503250

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)*. 1748–1752. DOI: https://doi.org/10.21437/Interspeech.2020-2382

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)*. 1325–1328. DOI: https://doi.org/10.21437/Interspeech.2010-413

Gretchen Montgomery and Yan Bing Zhang. 2018. Intergroup anxiety and willingness to accommodate: Exploring the effects of accent stereotyping and social attraction. *Journal of Language and Social Psychology* 37, 3 (2018), 330–349. DOI: https://doi.org/10.1177/0261927X17728361

Pietro Morasso. 1981. Spatial control of arm movements. *Experimental Brain Research* 42, 2 (1981), 223–227. DOI: https://doi.org/10.1007/BF00236911

Mikhail S. Nikulin. 2001. Hellinger distance. *Encyclopedia of Mathematics*. Springer. http://encyclopediaofmath.org/index.php?title=Hellinger_distance Accessed: 2021-01-31.

Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. 2023. A comprehensive review of data-driven co-speech gesture generation. *Computer Graphics Forum* 42, 2 (2023), 569–596. DOI: https://doi.org/10.1111/cgf.14776

Kunkun Pang, Taku Komura, Hanbyul Joo, and Takaaki Shiratori. 2020. CGVU: Semantics-guided 3D body gesture synthesis. In *Proceedings of the GENEA Workshop (GENEA'20)*. DOI: https://doi.org/10.5281/zenodo.4090878

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. 1532–1543. DOI: https://doi.org/10.3115/v1/D14-1162

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML'21)*. 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, Joao Monteiro, Jan Trmal, and Yoshua Bengio. 2020. Multi-task self-supervised learning for robust speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'20)*. 6989–6993. DOI: https://doi.org/10.1109/ICASSP40776.2020.9053569

Manuel Rebol, Christian Güti, and Krzysztof Pietroszek. 2021. Passing a non-verbal Turing test: Evaluating gesture animations generated from speech. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces (VR'21)*. 573–581. DOI: https://doi.org/10.1109/VR50410.2021.00082

Manuel Sam Ribeiro, Junichi Yamagishi, and Robert A. J. Clark. 2015. A perceptual investigation of wavelet-based decomposition of $f0$ for text-to-speech synthesis. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech'15)*. 1586–1590. DOI: https://doi.org/10.21437/Interspeech.2015-368

Najmeh Sadoughi and Carlos Busso. 2019. Speech-driven animation with meaningful behaviors. *Speech Communication* 110 (2019), 90–100. DOI: https://doi.org/10.1016/j.specom.2019.04.005

Khaled Saleh. 2022. Hybrid seq2seq architecture for 3D co-speech gesture generation. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI'22)*. 748–752. DOI: https://doi.org/10.1145/3536221.3558064

Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2013. To err is human(-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics* 5, 3 (2013), 313–323. DOI: https://doi.org/10.1007/s12369-013-0196-9

Maha Salem, Stefan Kopp, Ipke Wachsmuth, Katharina Rohlfing, and Frank Joublin. 2012. Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics* 4, 2 (2012), 201–217. DOI: https://doi.org/10.1007/s12369-011-0124-9

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 *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'11)*. 247–252. DOI: https://doi.org/10.1109/ROMAN.2011.6005285

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

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

Carolyn Saund and Stacy Marsella. 2021. The importance of qualitative elements in subjective evaluation of semantic gestures. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG'21)*. 1–8. DOI: https://doi.org/10.1109/FG52635.2021.9667023

Pranab Kumar Sen. 1968. Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association* 63, 324 (1968), 1379–1389. DOI: https://doi.org/10.1080/01621459.1968.10480934

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)*. 4799–4783. DOI: https://doi.org/10.1109/ICASSP.2018.8461368

Akihito Shimazu, Chie Hieida, Takayuki Nagai, Tomoaki Nakamura, Yuki Takeda, Takenori Hara, Osamu Nakagawa, and Tsuyoshi Maeda. 2018. Generation of gestures during presentation for humanoid robots. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'18)*. 961–968. DOI: https://doi.org/10.1109/ROMAN.2018.8525621

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)*. 3335–3339. Retrieved from https://aclanthology.org/L12-1513/

Hitoshi Teshima, Naoki Wake, Diego Thomas, Yuta Nakashima, Hiroshi Kawasaki, and Katsushi Ikeuchi. 2022. Deep gesture generation for social robots using type-specific libraries. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'22)*. 8286–8291. DOI: https://doi.org/10.1109/IROS47612.2022.9981734

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).* 718–723. DOI: https://doi.org/10.1109/ECTI-CON51831.2021.9454931

Henri Theil. 1992. A rank-invariant method of linear and polynomial regression analysis. In *Proceedings of the Henri Theil's Contributions to Economics and Econometrics: Econometric Theory and Methodology.* Baldev Raj and Johan Koerts (Eds.), Springer, 345–381. DOI: https://doi.org/10.1007/978-94-011-2546-8_20

Bruce Thompson. 1984. *Canonical Correlation Analysis: Uses and Interpretation.* Sage. Retrieved from https://uk.sagepub.com/en-gb/eur/book/canonical-correlation-analysis

George Toderici, Lucas Theis, Nick Johnston, Eirikur Agustsson, Johannes Ballé, Fabian Mentzer, Wenzhe Shi, and Radu Timofte. 2020. CLIC 2020: Overview, and Analysis of the Competition Results. Retrieved March 27, 2024 from https://youtu.be/iXzgFrRWNEg

Yoji Uno, Mitsuo Kawato, and Rika Suzuki. 1989. Formation and control of optimal trajectory in human multijoint arm movement. *Biological Cybernetics* 61, 2 (1989), 89–101. DOI: https://doi.org/10.1007/BF00204593

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. arXiv:1609.03499. Retrieved from https://arxiv.org/abs/1609.03499

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

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS'19).* Retrieved from https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html

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).* 177–185. DOI: https://doi.org/10.1145/3462244.3479914

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).* 802–810. DOI: https://doi.org/10.1145/3577190.3616116

Stephen J. Winters and David B. Pisoni. 2004. Perception and comprehension of synthetic speech. In *Proceedings of the Research on Spoken Language Processing Progress Report No. 26.* Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, IN, 95–138. Retrieved from https://citeseerx.ist.psu.edu/pdf/8e10a4c4d279e9540cd5af5aae692fe9907409ff

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).* 494–502. DOI: https://doi.org/10.1145/3462244.3479889

Pieter Wolfert, Gustav Eje Henter, and Tony Belpaeme. 2023. "Am I listening?", Evaluating the quality of generated data-driven listening motion. In *Proceedings of the Companion Publication of the ACM International Conference on Multimodal Interaction (ICMI'23 Companion).* 6–10. DOI: https://doi.org/10.1145/3610661.3617160

Pieter Wolfert, Taras Kucherenko, Hedvig Kjelström, and Tony Belpaeme. 2019. Should beat gestures be learned or designed? A benchmarking user study. In *Proceedings of the ICDL-EpiRob Workshop on Naturalistic Non-Verbal and Affective Human-Robot Interactions (ICDL-EpiRob'19 Workshop).* 4 pages. Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-255998

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. DOI: https://doi.org/10.1109/THMS.2022.3149173

Jieyeon Woo. 2021. Development of an interactive human/agent loop using multimodal recurrent neural networks. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI'21).* 822–826. DOI: https://doi.org/10.1145/3462244.3481275

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).* 758–763. DOI: https://doi.org/10.1145/3536221.3558066

Payam Jome Yazdian, Mo Chen, and Angelica Lim. 2022. Gesture2Vec: Clustering gestures using representation learning methods for co-speech gesture generation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'22).* 3100–3107. DOI: https://doi.org/10.1109/IROS47612.2022.9981117

Sheng Ye, Yu-Hui Wen, Yanan Sun, Ying He, Ziyang Zhang, Yaoyuan Wang, Weihua He, and Yong-Jin Liu. 2022. Audio-driven stylized gesture generation with flow-based model. In *Proceedings of the European Conference on Computer Vision (ECCV'22).* 712–728. DOI: https://doi.org/10.1007/978-3-031-20065-6_41

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* 39, 6, Article 222 (2020), 16 pages. DOI: https://doi.org/10.1145/3414685.3417838

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 *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'19).* 4303–4309. DOI: https://doi.org/10.1109/ICRA.2019.8793720

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

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 ACM International Conference on Multimodal Interaction (ICMI'22).* 736–747. DOI: https://doi.org/10.1145/3536221.3558058

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).* 342–346. DOI: https://doi.org/10.21437/Interspeech.2016-847

Fan Zhang, Naye Ji, Fuxing Gao, and Yongping Li. 2023. DiffMotion: Speech-driven gesture synthesis using denoising diffusion model. In *Proceedings of the International Conference on Multimedia Modeling (MMM'23).* 231–242. DOI: https://doi.org/10.1007/978-3-031-27077-2_18

He Zhang, Sebastian Starke, Taku Komura, and Jun Saito. 2018. Mode-adaptive neural networks for quadruped motion control. *ACM Transactions on Graphics* 37, 4, Article 145 (2018), 11 pages. DOI: https://doi.org/10.1145/3197517.3201366

Kai Zhang, Shuhang Gu, and Radu Timofte. 2020. NTIRE 2020 challenge on perceptual extreme super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR'20 Workshop).* 492–493. DOI: https://doi.org/10.1109/CVPRW50498.2020.00254

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).* 764–770. DOI: https://doi.org/10.1145/3536221.3558063

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2019. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19).* 5745–5753. DOI: https://doi.org/10.1109/CVPR.2019.00589

Received 15 February 2023; revised 24 November 2023; accepted 27 February 2024