The GENEA Challenge 2023: A large-scale evaluation of gesture generation models in monadic and dyadic settings

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

Rajmund Nagy*
rajmund@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

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

Jieyeon Woo
woo@isir.upmc.fr
ISIR – Sorbonne University
Paris, France

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

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

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

ABSTRACT

This paper reports on the GENEA Challenge 2023, in which participating teams built speech-driven gesture-generation systems using the same speech and motion dataset, followed by a joint evaluation. This year's challenge provided data on both sides of a dyadic interaction, allowing teams to generate full-body motion for an agent given its speech (text and audio) and the speech and motion of the interlocutor. We evaluated 12 submissions and 2 baselines together with held-out motion-capture data in several large-scale user studies. The studies focused on three aspects: 1) the human-likeness of the motion, 2) the appropriateness of the motion for the agent's own speech whilst controlling for the human-likeness of the motion, and 3) the appropriateness of the motion for the behaviour of the interlocutor in the interaction, using a setup that controls for both the human-likeness of the motion *and* the agent's own speech. We found a large span in human-likeness between challenge submissions, with a few systems rated close to human mocap. Appropriateness seems far from being solved, with most submissions performing in a narrow range slightly above chance, far behind natural motion. The effect of the interlocutor is even more subtle, with submitted systems at best performing barely above chance. Interestingly, a dyadic system being highly appropriate for agent speech does not necessarily imply high appropriateness for the interlocutor. Additional material is available via the project website at svito-zar.github.io/GENEAchallenge2023/.

*Equal contribution and joint first authors.

CCS CONCEPTS

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

KEYWORDS

gesture generation, embodied conversational agents, evaluation paradigms, dyadic interaction, interlocutor awareness

1 INTRODUCTION

Verbal communication is not the only way to convey messages; non-verbal behaviour is also a significant factor in human communication [32]. Much of non-verbal behaviour consists of hand gestures and body gestures, which closely relate to speech content and have been shown to improve speech understanding [16].

Automatic gesture generation for embodied conversational agents (ECAs) has been studied for more than two decades. Synthetic gestures were initially created using rule-based systems, e.g., [7, 37], but data-driven approaches that learn from human motion data have recently emerged [2, 5, 10, 25, 30, 47, 48]. For more in-depth reviews of speech gesture generation, refer to Nyatsanga et al. [34].

The present study focuses on a fair, systematic comparison of systems that automatically generate non-verbal behaviour. By comparing different methods and evaluating their effectiveness, it is possible to accurately assess and improve the current state of the art. Additionally, this comparison helps to identify important aspects of gesture generation and where the main issues lie. We accomplished this by running an open challenge, the GENEA[1] Challenge 2023.

---

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

In this paper, we report on the challenge, in which we provided the same dataset, evaluation criteria, and visualisation process to ensure that all major sources of variation are accounted for, except for a gesture-generation system itself. The GENEA Challenge 2023 builds on previous GENEA Challenges [26, 50], especially the 2022 challenge described in Yoon et al. [50] and Kucherenko et al. [28]. In addition to being an exercise in benchmarking both new and previously-published gesture-generation methods, the results of the two previous challenges have since helped improve gesture-generation benchmarking in other ways as well. Researchers have, for example, used the data splits [3, 22], the visualisation [40], and the objective [6] and subjective [43, 44, 49] evaluation methodologies, as a basis for future research. The data has also been used to benchmark subsequent gesture-generation models [12, 46], and even for automatic quality assessment [15]. The main difference from previous GENEA Challenges is that we considered gesture generation for dyadic scenarios (meaning considering both conversation partners) and evaluated the interlocutor appropriateness of submitted systems.

Our contributions include: 1) providing common elements of dataset, visualisation, and evaluation setup for a fair comparison between gesture generation systems; 2) organising an open challenge to promote the development of the field and encourage researchers to share their ideas; and 3) three large-scale user studies for human-likeness of motion, speech appropriateness, and interlocutor appropriateness that evaluate 14 gesture-generation systems (including two baselines).

2 EVALUATION SETUP

The focus of the GENEA Challenge is on large-scale, crowdsourced subjective evaluation of the generated gestures through multiple user studies. This year, we evaluated three aspects of the generated gestures:

Human-likeness: Whether the motion of the virtual character looks like the motion of a real human, without explicitly considering the speech or any interlocutor behaviour.

Appropriateness for agent speech: Whether the motion of the virtual character is appropriate for the given speech, controlling for the overall human-likeness of the motion.

Appropriateness for interlocutor behaviour: Whether the motion of the virtual character is appropriate for the given interlocutor behaviour (both speech and motion), controlling for the overall human-likeness of the motion and (as far as possible) for the effect of the agent's own speech.

We sometimes call the first two studies "monadic" and the third "dyadic". Note that the models always had the interlocutor information as input, even in monadic evaluations. All evaluations are built on the same avatar and visualisation software as used in last year's challenge (see [27]). More details about the three evaluations are given in Secs. 3 to 5, respectively.

2.1 Data

In order to evaluate gesture generation in a dyadic context, we supplement the recordings in last year's dataset [50] with information from the conversation partner in the original Talking With Hands (TWH) data [29]. Specifically, we provide speech (audio and transcription), speaker ID, and motion for two parties; the *main agent*

(for which the task is to generate motion) and the *interlocutor* (which is the other party in the conversation). As a minor improvement to last year's data, we set the height of the hips to the same for all speakers, since they all use the same skeleton. This makes the character's feet touch the ground in a consistent manner for all speakers. The dataset is publicly available at doi.org/10.5281/zenodo.8199132.

We provided three splits to the challenge participants: 1) a training split, composed of last year's training and validation data [50], 2) a validation split, equivalent to last year's test data, and 3) a test split, containing data that was completely left out from the 2022 challenge but with main-agent motion omitted. We augmented the training data by making each recorded conversation appear twice, with the conversational roles flipped. There was also an option for participating teams to leverage a version of the BEAT dataset [31], retargeted and adapted to the challenge skeleton, but none of the submitting teams used it.

We created a *core test set* consisting of 41 *chunks* of approximately one minute each, restricted to recordings with finger motion tracking for the speaker chosen to be the main agent. We further extended the test set with 29 additional chunks in which the two sides of the conversation were mismatched, for use in the interlocutor awareness study (see Sec. 5.1 for details). Hence, the extended test set (core and additional test sets) consists of 70 chunks in total.

2.2 Conditions evaluated

In total, 12 teams participated in the GENEA Challenge 2023 evaluation. The evaluation also included two baseline systems and natural motion taken from the motion-capture recordings from the speakers in the database.

We call each source of motion a *condition* rather than a "system" or "model", since the evaluation also includes avatar motion based on human mocap. This natural motion was labelled **NA** (for "natural") in all evaluations. We prefer not to use the term "ground truth", since there is no single, true way to move to a given speech and interlocutor. The NA condition is intended as a top line, but in practice there can be some mocap artefacts and the motion is thus not always completely human-like, as seen from the stimuli and results of the 2022 challenge [50].

Sources of artificial, generated motion may be referred to as *systems* or *models*. The 12 systems entered into the challenge by participating teams [9, 11, 14, 21, 23, 24, 38, 39, 41, 45, 51, 52] are often called *entries* or *submissions*.

In addition to natural mocap and the different entries, the evaluations also included two baseline systems for automatic gesture-generation, both based on an entry from the GENEA Challenge 2022, specifically Chang et al. [8], which won the reproducibility award. This means that a total of 15 conditions were included in the evaluation. The two baselines were included with the intention to provide continuity and easier comparison between different years of the challenge, and also to track the progress of the field with respect to a fixed baseline. (Note that none of the systems used as baselines in the 2020 evaluation were included this year, since those systems now are too far behind the state of the art to constitute reasonable points of comparison.) The two baseline conditions were: 1) Monadic baseline (**BM**): The 2022 challenge entry from Chang et al. [8] with slightly adjusted hyper-parameters, since the dataset is not the exact same. 2) Dyadic baseline (**BD**): An adaptation of
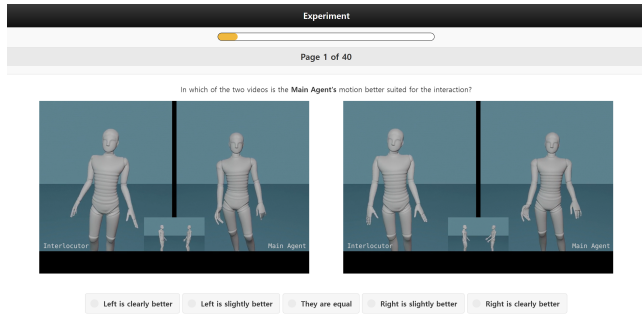
Figure 1: Screenshots of the response interfaces from the appropriateness for the interlocutor evaluation, showing questions, response options, and snapshots of stimulus videos.

BM that also takes information from the interlocutor into account as an extra input to the model.

In the results, each condition was given a two-character label, a.k.a. condition ID. The first character identifies the condition type, with 'N' for motion-capture of natural human motion, 'B' for a baseline, and 'S' for a submission from a participating team. The second character distinguishes between conditions within the same tier and type. In particular, the challenge entries were labelled **SA**–**SL**. These labels are anonymous and have no relationship to the team names or identities.

2.3 Setup of the user studies

Study participants (a.k.a. *test takers*) were recruited through the crowdsourcing platform Prolific. Participants were required to reside in a set of six English-speaking countries, specifically UK, IE, USA, CAN, AUS, and NZ. A Prolific user could take any number of our studies, but could only participate in each study at most once.

While the specifics of the three user studies differed (see Secs. 3 to 5 for details), they also shared some common elements. In particular, we applied the same counterbalancing approach as in last year's challenge [50]. Studies began with a screen of instructions, followed by a designated training page showing a fixed set of videos with different motions, to familiarise participants with the task. Furthermore, after completing the study, participants filled in a short questionnaire to gather broad, anonymous demographic information. We set the payment to 6 GBP for each study.

The studies also incorporated four attention checks per test taker similar to last year [50], to make sure that participants were paying attention to the task and to remove insincere test-takers. Each attention check was a modified stimulus that (partway through the stimulus) instructed the participant, using either a text overlay or a synthesised voice (only in studies where the audio was not muted), to select a specific response option. Test takers who failed two or more attention checks were removed from the study without pay. You can find statistics about the test takers in Table 1.

2.4 Visualisation

Like in previous years, we developed and maintained a standardised visualisation pipeline to which challenge participants had access. This allowed participants to preview their generated motions and to roughly assess how the final stimuli videos during evaluation would

look like. We used the same virtual avatar (Figure 1) in all rendered videos during the challenge and the evaluation. The avatar was the same as last year [50]. The visualiser is capable of visualising both monadic and dyadic interactions; either one character or two characters are visible at a time. All teams had access to the official visualisation and rendering pipeline, in the form of code intended for local use with Blender. The code is open source and is available at github.com/TeoNikolov/genea_visualizer/ and the rendered stimuli videos at doi.org/10.5281/zenodo.8211448.

3 HUMAN-LIKENESS EVALUATION

The human-likeness evaluation for the GENEA Challenge 2023 used the HEMVIP methodology [20], same as in 2022, with some minor changes to GUI design, question formulation and response interface, and the statistical analysis.

3.1 Data selection for stimuli

From the 41 chunks of core test set data, we selected 41 short segments of test speech and corresponding test motion to be used in the two agent-focussed subjective evaluations. Our rules for selecting these segments were as follows: (i) Segments should be around 8 to 10 seconds long. (ii) Segments should only be taken from parts of the interaction where the person chosen as the main agent is the active speaker (so no turn-taking within a segment, but backchannels from the interlocutor were OK). (iii) Segments should not contain any parts where Lee et al. [29] had replaced the speech with silence for anonymisation. (iv) Segments should be more or less complete phrases, starting at the start of a word and ending at the end of a word. (v) The recorded motion capture in the segments (i.e., the NA motion) should not contain any significant artefacts such as whole-body vibration.

The 41 segments selected in this way were between 7 and 13 seconds in duration, with the average duration being 9 seconds.

3.2 Evaluation procedure

On each page (a.k.a. screen) in the HEMVIP evaluation, eight different motion examples were presented in parallel [20], all corresponding to the same speech segment, but different conditions.

Each page asked participants "Please indicate on a sliding scale how human-like the gesture motion appears". Study participants gave their ratings in response to this question on a scale from 0 (worst) to 100 (best) by adjusting an individual GUI slider for each video. In contrast to previous years, however, the numbers corresponding to the exact rating were not shown, and test takers were not explicitly told, e.g., that the sliders had 101 steps. By using sliders without numerical labels, it is more clear to test takers that the scale is intended to be ordinal.

Like in Jonell et al. [20], Kucherenko et al. [26], the rating scale was anchored by partitioning the sliders into five equal-length intervals labelled (from best to worst) "Excellent", "Good", "Fair", "Poor", and "Bad". These labels were based on those associated with the 5-point scale used in the Mean Opinion Score (MOS) [18] standard for audio quality evaluation. An example of the evaluation interface can be seen in the supplement.

Since it has been found that speech content can influence gesture perception and confound motion evaluations [19], the videos seen

Table 1: Statistics on test takers and user-study responses.

Study	No. test takers	Videos per page	test taker	No. responses per condition min	max	Median time taken
Human-likeness	200	8	80	938	2000	25.3 min
Speech approp.	600	2	80	1766	1815	23.8 min
Interloc. approp.	423	2	80	993	1019	30.3 min
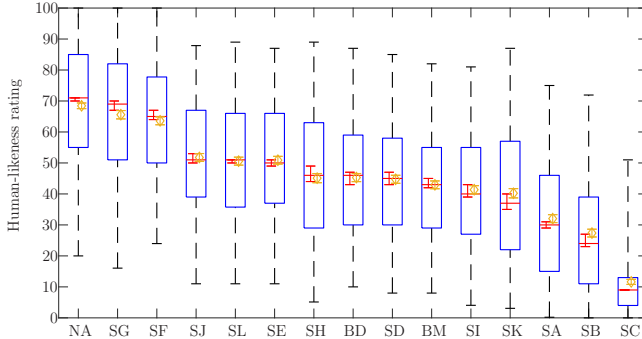
Figure 2: Box plot visualising the rating distribution in the human-likeness study. Red bars are the median ratings (each with a 0.05 confidence interval); yellow diamonds are mean ratings (also with a 0.05 confidence interval). Box edges are at 25 and 75 percentiles, whilst whiskers cover 95% of all ratings for each condition. Conditions are ordered by descending sample median rating.

by participants in these human-likeness evaluations (although they all corresponded to the same speech input and had the same length) were completely silent and did not include any audio. This way, ratings can only depend on the motion seen in the videos.

Each participant completed 10 pages of ratings for the evaluation. The evaluation design was balanced in the same way as last year [50] such that each segment appeared on pages 1 through 10 with approximately equal frequency across all participants (segment order).

3.3 Response data

Confidence intervals for the median were computed using order statistics, leveraging the binomial distribution cumulative distribution function, cf. [13], while those for the mean used a Gaussian assumption (i.e., using Student's t-distribution cdf, rounded outward to ensure sufficient coverage).

The rating distribution in the study is further visualised through the box plot in Figure 2. The rating distributions in the figure are seen to be quite broad. This is common in evaluations like HEMVIP, 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, due to the large number of ratings collected for each condition.

3.4 Significant differences

To analyse the significance of differences in median rating between different conditions, we applied two-sided Mann-Whitney U tests
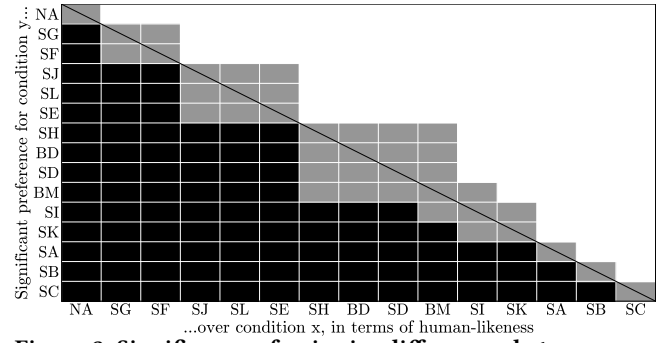
Figure 3: Significance of pairwise differences between conditions in human-likeness. White means that the condition listed on the y-axis rated significantly above the condition on the x-axis, black means the opposite, and grey means no statistically significant difference at the level $\alpha = 0.05$ after Holm-Bonferroni correction. Conditions are listed in the same order as in Figure 2.

to all unordered pairs of distinct conditions in each study. This is an unpaired nonparametric test, which differs from the pairwise Wilcoxon signed-rank tests used in previous GENEA Challenges. The reason for this change is that, whilst pairwise tests can more easily control for variation between segments and raters, such tests only can be used for data from screens where both conditions under comparisons were shown to participants. Since this year's challenge is comparing more conditions than before but still uses the same number of sliders (8 in parallel), such pairings are relatively rare. An unpaired test can use all the available ratings of any given condition, making it more powerful in this instance.

Unlike Student's t-test, which assumes that the test statistic follows a Gaussian distribution, this analysis is valid also for ordinal response scales, like those we have here. For each condition pair, only pages for which both conditions were assigned valid ratings were included in the analysis of significant differences.

The p-values computed in the significance tests were adjusted for multiple comparisons using the Holm-Bonferroni method [17], which is uniformly more powerful than conventional Bonferroni correction, to keep the family-wise error rate (FWER) at or below $\alpha = 0.05$. After FWER correction, our statistical analysis found all but 12 out of 105 condition pairs to be significantly different at the level $\alpha = 0.05$. Which conditions that were found to be rated significantly above or below which other conditions in the study is shown in Figure 3.

3.5 Discussion

Looking at the results, we see that the challenge submissions span a wide range of different human-likeness scores, with a greater concentration near the middle. Unlike the GENEA Challenge 2022 [50], no submission achieved median human-likeness ratings that exceeded those of the human motion capture, although two submissions were not far behind. This helps establish the trend that, whilst human-like gesture motion is not altogether a solved problem, strong performance in that regard is possible (but not necessarily easy) with current data-driven methods. The baselines BD and BM performed similarly to each other and achieved ratings near the middle of the pack. This resembles the performance of

the underlying system last year [8, 28], when it was a challenge submission.

Similar to past years' challenges, natural mocap (NA) is rated significantly lower than perfect human-likeness. Among other things, this may be attributed to shortcomings with the mocap, the 3D model used (which deliberately appears neutral and lacks mouth and gaze), and human biases against assigning high scores.

One other noteworthy observation to keep in mind for later is the fact that ordering conditions by their median human-likeness (as in Figure 2) produces a quite different ordering compared to ordering systems by appropriateness in Secs. 4 and 5. This indicates that we managed to disentangle human-likeness and appropriateness in our evaluation.

4 APPROPRIATENESS FOR AGENT SPEECH

Similar to the GENEA Challenge 2022 [50], we also evaluated the appropriateness of the generated motion for the main agent's speech, using a similar but improved methodology compared to 2022.

The GENEA Challenge 2022 adapted a new mismatching methodology [19, 36] with the design goal to evaluate motion appropriateness whilst controlling for the human-likeness of the motion in an effective way. Test takers were presented with a pair of videos, each video showing motion corresponding to a different speech recording. We replaced the audio track of one of the videos with the speech audio of the other clip, resulting in one video where the motion matched the audio, and another video where audio and motion were from unrelated sources. Participants were then asked to pick one video from the pair where motion, according to them, best matched the speech. Their responses reveal how specific the gesture motion is to the given speech.

Since last year's appropriateness evaluation, with only three possible response options, had quite wide confidence intervals, we asked test takers to indicate their degree of preference this year, so as to get information value out of each response. This change was based on a suggestion for future evaluations in Kucherenko et al. [28] and piloted in Mehta et al. [33]. More specifically, test takers were asked "Which character's motion matches the speech better, both in terms of rhythm and intonation and in terms of meaning?" The five response options were adapted from [33], merely changing "much" to "clearly". Which of the two videos (left or right) was the matched one was random. A screenshot of the evaluation interface used for the appropriateness studies is presented in the supplement.

In this type of evaluation, a system whose output does not depend on the input will perform at chance rate. This avoids the issues seen in the appropriateness evaluation of the GENEA Challenge 2020 [26], and we no longer recommend that earlier methodology.

4.1 Stimuli

Table 2 provides an overview of the different stimuli used in the different user studies. The speech appropriateness evaluation leveraged the same evaluation segments chosen for the human-likeness evaluation in Sec. 3.1. Concretely, we created the mismatched stimuli by taking the speech and motion for the 41 segments selected there, and then permuting the motion in between them such that no motion segment ever remained in its original place. Since the different segments generally do not have the same length, sometimes a longer or shorter segment of motion had to be excerpted from the test-set chunks (original or generated), so as to match the

new speech duration, but the starting point of the motion video was always the same as in the respective matched stimulus video (i.e., corresponding to the start of a phrase). In the terminology of Table 2, the mismatched motion segments were excerpted using the start time of segment 2 but the duration of segment 1. This is exactly the same as was done in 2022.

4.2 Evaluation interface and question asked

Because we are deliberately avoiding comparisons across conditions (as different conditions may have different motion qualities), test takers were only asked to view and compare a single pair of videos at a time. For pairwise comparisons like these, asking for a preference between two items can be more efficient than asking participants to rate each video [42]. We therefore asked for preference rather than absolute ratings in these studies. The question asked was "Which character's motion matches the speech better, both in terms of rhythm and intonation and in terms of meaning?".

4.3 Response data

For analysis, the five possible responses were converted to integer values $\{-2, -1, 0, 1, 2\}$ in order, with -2 meaning the mismatched stimulus was rated as clearly better and 2 meaning the matched stimulus was rated as clearly better. Raw response statistics for all conditions in each of the two studies are available in the appendix.

The table also includes 95% confidence intervals for the average numerical value of the per-condition user responses, computed using a Student's t-distribution. We refer to this average, used for our appropriateness statistical analyses, as *mean appropriateness score* (or *MAS*).

The response distributions in the study are further visualised through the bar plot in Figure 4a. The bar plot also visualises the confidence interval for the mean appropriateness score, but scaled linearly from the interval $[-2, 2]$ to $[0\%, 100\%]$ in order to make the intervals comparable with the rest of the plot.

4.4 Statistical analysis

Two different statistical analyses were applied to the results, testing:
  (1) Whether the mean appropriateness score of a given condition is statistically different from chance performance (without correcting for multiple comparisons).
  (2) Whether the mean appropriateness score of any given condition was statistically different from the mean appropriateness score (correcting for multiple comparisons).
All statistical tests were carried out at the $\alpha = 0.05$ level.

Remember that in our evaluation, a system whose output does not depend on the input will theoretically perform at a chance rate.

4.4.1 MAS differences from chance performance. The dotted line at 50% (equal to an MAS of zero) in Figure 4a marks chance performance. Any condition whose 95% MAS confidence interval does not overlap with zero demonstrates performance that is statistically different from chance (without any correction for multiple comparisons). On this measure, all conditions except SL and SC were statistically better than chance in terms of MAS.

4.4.2 MAS differences between conditions. To assess whether any two conditions were statistically different from one another, we used Welch's t-test. This is an unpaired statistical test; the study

Table 2: Overview of the stimuli used in the different studies. Numbers 1 and 2 indicate which of two test-set segments – the matched or the mismatched – that data was sourced from. Each matched segment was paired with a corresponding mismatched segment elsewhere in the test data. Abbreviations: "Hum.-like." stands for "Human-likeness", "approp." stands for "appropriateness", "segs." stands for "segments", "ID" stands for "Speaker ID", and "Interloc." for "Interlocutor". The motion for condition NA was taken from segment 1 in matched stimuli and 2 in mismatched. All other conditions used motion synthesised from the inputs listed under "Source of test-time system inputs".

| | Studies featured | | | Active | Source of test-time system inputs | | | | | Source of what is presented in stimuli | | |
| | Hum.- | Speech | Dyadic | speaker | Main agent | | Interlocutor | | | Main agent | Interlocutor | |
Stimulus type	like.	approp.	approp.	in segs.	Speech	ID	Speech	Motion	ID	speech	Speech	Motion
Matched (speaking)	✓	✓		Agent	1	1	1	1	1	1	N/A (monadic)	
Mismatched (speaking)		✓		Agent	2	2	2	2	2	1	N/A (monadic)	
Matched (listening)			✓	Interloc.	1	1	1	1	1	1	1	1
Mismatched (listening)			✓	Interloc.	1	1	2	2	2	1	1	1
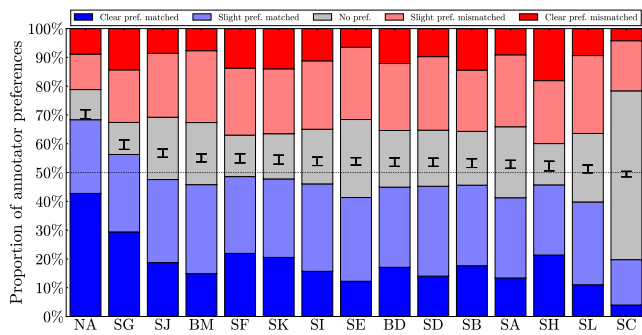
(a) Appropriateness for agent speech
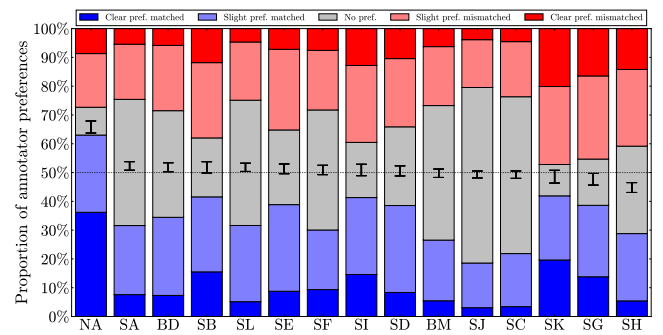
(b) Appropriateness for the interlocutor

Figure 4: Bar plots visualising the response distribution in the appropriateness studies. On top of each bar is also a confidence interval for the mean appropriateness score, scaled to fit the current axes. The dotted black line indicates chance-level performance. Conditions are ordered by mean appropriateness score.

Table 3: Mean appropriateness scores (MAS) from both appropriateness studies with confidence intervals at the level $\alpha = 0.05$. Submissions are ordered alphabetically.

Con-dition	MAS for… Speech	MAS for… Interloc.	Con-dition	MAS for… Speech	MAS for… Interloc.
NA	0.81±0.06	0.63±0.08	SF	0.20±0.06	0.04±0.06
BD	0.14±0.06	0.07±0.06	SG	0.39±0.07	−0.09±0.08
BM	0.20±0.05	−0.01±0.06	SH	0.09±0.07	−0.21±0.07
SA	0.11±0.06	0.09±0.06	SI	0.16±0.06	0.04±0.08
SB	0.13±0.06	0.07±0.08	SJ	0.27±0.06	−0.03±0.05
SC	−0.02±0.04	−0.03±0.05	SK	0.18±0.06	−0.06±0.09
SD	0.14±0.06	0.02±0.07	SL	0.05±0.05	0.07±0.06
SE	0.16±0.05	0.05±0.07			

design was not optimised for pairwise testing since a re-analysis of the data from Mehta et al. [33] to compare the effect of paired versus unpaired t-tests for analysis produced no noteworthy differences.

A correction for multiple comparisons was performed to control the false discovery rate (FDR), i.e., the fraction of erroneous rejections of the null hypothesis out of all rejected null hypotheses. Specifically, we used the BH non-adaptive one-stage linear step-up procedure [4] as implemented in the MultiPy package. Note that controlling the FDR is not the same as controlling for the family-wise error rate (which controls the probability of making any false rejections at all), as done in Sec. 3.

After FDR correction, our statistical analysis found 56 of 105 condition pairs to be significantly different. Figure 5a visualises the statistically significant differences between conditions.

4.5 Discussion

Just as in the GENEA Challenge 2022 results, natural motion capture (condition NA) exhibited the greatest difference between matched and mismatched stimuli. NA attained an MAS of 0.81 ± 0.06 on the scale from −2 to 2. A more in-depth discussion of why this magnitude of difference is reasonable (and why one should not expect 100%) can be found in [28].

Unlike the human-likeness evaluation, however, no submission came close to NA in terms of specific appropriateness to the speech in this evaluation, with the best synthesis system achieving a MAS around 0.39 and a 62% preference for matched motion after splitting ties. Also unlike the case for human-likeness, most submissions are confined to a narrow range of MAS scores between 0.27 and 0.10, or so. Clearly, current synthesis methods are far from as successful as humans are in producing gestures that match the specific speech, although the current evaluation does not reveal to what extent the shortcomings seen are related to mistimed gesturing, or to failures to generate semantically relevant gesticulation. Either way, it is not clear if relatively non-specific gesture motion, as generally produced by the systems in the challenge, enhances human-agent communication. Since better communication is a major motivator of research into (and applications of) embodied conversational

(a) Appropriateness for agent speech

(b) Appropriateness for the interlocutor
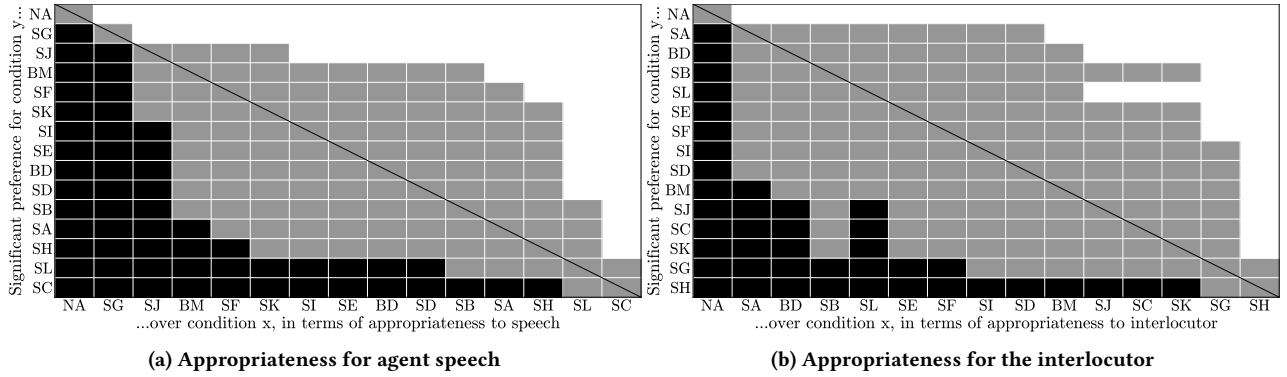
Figure 5: Significant differences between conditions in the two appropriateness studies. White means the condition listed on the y-axis achieved an MAS significantly above the condition on the x-axis, black means the opposite , and grey means no statistically significant difference at level $\alpha = 0.05$ after correction for the false discovery rate. Conditions use the same order as the corresponding subfigure in Figure 4.

agents in the first place, our results suggest that improving the appropriateness of synthesised gestures for agent speech is an important research target.

5 APPROPRIATENESS FOR THE INTERLOCUTOR

Although there were many similarities to the appropriateness evaluation in Sec. 4, this study differed in several important aspects, namely in what was being studied, how the motion contained in the stimuli was obtained, and how the stimuli were presented for evaluation.

For this final evaluation, we wanted to use mismatching to quantify motion appropriateness for the interaction whilst controlling for both motion human-likeness *and* for the effect of the main agent's own speech as far as possible. This required carefully crafted stimuli.

5.1 Data creation for stimuli

We reasoned that interlocutor-aware behaviour might be most salient when the interlocutor is the active speaker, and the main agent is reacting to them. For this reason, this appropriateness evaluation exclusively considered segments where the interlocutor was the active speaker.

Prior to the test set being shared with challenge teams, we selected 29 segments from the core test set where the main agent exhibited dyadic adaptations to the interlocutor, for example head nods. The segments selected in this way were between 5.5 and 15 seconds in duration, with the average duration being 10.4 seconds.

For each segment that we selected from the test set, we proceeded to create a second version of the same interaction, where the main agent was the same (same WAV, TSV, and speaker ID) but the interlocutor was replaced by mismatched interlocutor behaviour (WAV, TSV, and speaker ID) taken from another interaction. The mismatched interlocutor behaviour was chosen such that the interlocutor was the active speaker for the duration of the segment used. These mismatched interactions, one interaction for each segment used for the present study, formed the extended test set mentioned in Sec. 2.1. When participating teams generated motion for the extended test set, they would thus create motion examples based on the same main-agent speech as the matched segments, but using different interlocutor behaviour as input. This can be used to create

stimulus pairs where any systematic differences in main agent motion is attributable only to whether or not the motion was generated for the matched interlocutor behaviour or not. See Table 2 for a complementary specification of this setup.

5.2 Stimuli

Figure 1 shows static images of example video stimuli that contains two frontal views, one of each party in the conversation, and a third view showing the two interacting characters from the side.

Unlike the stimuli in previous appropriateness studies, the audio in the stimuli video clips was presented in stereo, with the interlocutor audio in the left channel and the main agent audio in the right channel. This made it possible to tell who was who and who was saying what despite the two avatars not having mouths.

Stimuli (whether matched or mismatched) would only use audio from the matched segment/interaction. Similarly, the interlocutor motion would also be taken from the matched interaction for both matched and mismatched stimuli. Thus the only difference between the two video stimuli in a pair would be in how the main agent moves. In the matched case, it would be based on the speech and the interlocutor motion seen in the video, but in the mismatched case, it would be based on the same agent speech but different interlocutor speech and motion, leveraging the artificially-constructed extended test set. The only exception would be for condition NA, where the mismatched main-agent motion would, by necessity, be based on mismatched rather than matched agent speech.

5.3 Evaluation interface and question asked

The evaluation interface for the user study of motion appropriateness for the interlocutor was exactly the same as in Sec. 4.2. This time, test takers were asked "In which of the two videos is the Main Agent's motion better suited for the interaction?" A screenshot of the complete evaluation interface can be seen in Figure 1.

After an instructions page and a training page, each study participant evaluated 40 pages (i.e., 40 pairs of videos) in a counterbalanced design. This means that each person watched 80 videos in total. A design goal of this study was that each condition would receive approximately 1,000 or more responses, similar to before.

Despite the same number of videos watched, the time taken per participant was greater than in Sec. 4. One explanation might be

that the task is more demanding, with multiple characters to pay attention to.

5.4 Response data

The statistical analysis used mean appropriateness scores based on the integer values $\{-2, -1, 0, 1, 2\}$ as before. Raw response statistics for all conditions in each of the two studies are shown in Table 3, along with 95% confidence intervals for the MAS, all computed as before.

Like before, the response distributions from the study are further visualised through the bar plot in Figure 4b, together with confidence intervals for the mean appropriateness score, scaled to fit the rest of the plot.

5.5 Statistical analysis

The same two statistical analyses as in Sec. 4.4 were applied.

5.5.1 MAS differences from chance performance. The dotted line at 50% in Figure 4b marks chance performance. Any condition whose 95% MAS confidence interval does not overlap with zero (which coincides with the dotted line in the graph) demonstrates performance that is statistically different from chance (without any correction for multiple comparisons). On this measure, NA, BD, SA, and SL are significantly better than chance, whilst SG and SH instead are significantly worse. The remaining conditions were not statistically different from chance.

5.5.2 MAS differences between conditions. To assess whether any two conditions were statistically different from one another, we once again used Welch's *t*-test together with the BH non-adaptive one-stage linear step-up procedure [4] to control the false discovery rate. After FDR correction, our statistical analysis found 45 of 105 condition pairs to be significantly different. Figure 5b visualises these statistically significant differences using the same condition order as the box plot.

5.6 Discussion

To begin with, it must be pointed out that condition NA in this study is not comparable to the others, since its motion is mismatched not only with respect to the interlocutor, but also with respect to the main agent's own speech. This cannot be avoided, since the interactions in the extended test set do not have any counterparts in real interactions. The performance of NAs is thus in a sense a (likely unattainable) upper bound on MAS in this experiment. However, it does establish that mismatched behaviour can reliably be detected by the test takers, even for cases where the main agent is listening rather than speaking, albeit with a slightly lower MAS than for the speaking segments in Table 3.

In general, we may expect the difference between matched and mismatched stimuli to be smaller in the dyadic study, since for the monadic study, all inputs to the models differed when generating the two motions being compared, whereas only the model inputs corresponding to the interlocutor differed for this study. The greater fraction of "They are equal" responses in Figure 4b compared to Figure 4a indicates that this is indeed the case.

Having 5 out of 14 artificial systems differ from chance performance at level $\alpha = 0.05$ is very unlikely to occur by random

happenstance despite the absence of a correction for multiple comparisons. This strongly suggests that a notable fraction of the dyadic systems indeed create meaningfully different motion depending on the behaviour of the interlocutor, although the effect in absolute terms is small.

Looking at individual conditions, we can make some additional observations. Unlike previous studies, the two baselines are now far apart in the rankings, with BD (as hoped) being significantly different from chance performance, whilst BM (as expected) was not. The MAS of conditions SG and SH are also notable. Condition SG has gone from one of the most appropriate conditions (for the agent speech) to one of the least appropriate (for the interlocutor/interaction) in Table 3. Condition SH, meanwhile, exhibited the largest effect size of any artificial system in this dyadic evaluation, but in the direction of being *less* appropriate than random chance. It is not clear to what extent this may be influenced by to the use of listening rather than speaking, or by the "chimeric" and artificial inputs in the extended test set, or even by the use of a dyadic visualisation.

6 CONCLUSION

We have hosted the GENEA Challenge 2023 to investigate and evaluate different techniques of speech-driven gesture-generation. Via our evaluation, we remark that it is challenging to generate human-like and speech-appropriate gestures that also address the problem of assuring appropriateness with respect to the behaviour of the interlocutor. Nevertheless, it is promising to see that a substantial amount of submissions managed to respond to the interlocutor. Notably, from the dyadic systems that consider the information from the interlocutor, we remark that the interlocutor's information surely plays a role in rendering reactive motions that match the interlocutor's behaviours. Communication is not only mono-directional or bi-directional, but also multi-party. We envision observing more research conducted on dyadic interaction in the short-term and on multi-party interaction in the mid-term future to address all types of interaction. ECAs use synthetic speech to render their utterances. Recent studies show that human voice is perceived as more natural and intelligible than synthesised speech [1, 35]. We are interested in exploring the overall impression of the agent of whether the synthesised speech influences the global perceptive effect of the agent's human-likeliness. Further research awaits for study around gesture-generation to create human-like and appropriate ECAs capable of interacting in all types of conversations. The proposed challenge and similar ones can serve as the foundation for the development of such ECAs by identifying and sharing key aspects of gesture generation with the research community.

REFERENCES

[1] Amal Abdulrahman and Deborah Richards. 2022. Is natural necessary? Human voice versus synthetic voice for intelligent virtual agents. *Multimodal Technologies and Interaction* 6, 7 (2022), 51.

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

[3] 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 (TOG)* 42, 4 (2023), 1–20.

[4] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B Met.* 57, 1 (1995), 289–300.

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

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

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

[8] Che-Jui Chang, Sen Zhang, and Mubbasir Kapadia. 2022. The IVI Lab Entry to the GENEA Challenge 2022 – A Tacotron2 Based Method for Co-Speech Gesture Generation With Locality-Constraint Attention Mechanism. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI '22)*. ACM, 784–789. https://doi.org/10.1145/3536221.3558060

[9] Ankur Chemburkar, Shuhong Lu, and Andrew Andrew. 2023. Discrete Diffusion for Co-Speech Gesture Synthesis. In *Companion Publication of the 2023 International Conference on Multimodal Interaction (ICMI '23 Companion)*. Association for Computing Machinery.

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

[11] Anna Deichler, Shivam Mehta, Simon Alexanderson, and Jonas Beskow. 2023. Diffusion-based co-speech gesture generation using joint text and audio representation. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI '23)*. ACM.

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

[13] Gerald J. Hahn and William Q. Meeker. 1991. *Statistical Intervals: A Guide for Practitioners*. Vol. 92. John Wiley & Sons.

[14] Leon Harz, Hendric Voß, and Stefan Kopp. 2023. FEIN-Z: Autoregressive Behavior Cloning for Speech-Driven Gesture Generation. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI '23)*. ACM.

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

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

[17] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6, 2 (1979), 65–70.

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

[19] Patrik Jonell, Taras Kucherenko, Gustav Eje Henter, and Jonas Beskow. 2020. Let's face it: Probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings. In *Proceedings of the ACM International Conference on Intelligent Virtual Agents (IVA '20)*. ACM, Article 31, 8 pages. https://doi.org/10.1145/3383652.3423911

[20] Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, and Gustav Eje Henter. 2021. HEMVIP: Human evaluation of multiple videos in parallel. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI '21)*. ACM, 707–711. https://doi.org/10.1145/3462244.3479957

[21] Gwantae Kim, Yuanming Li, and Hanseok Ko. 2023. The KU-ISPL entry to the GENEA Challenge 2023-A Diffusion Model for Co-speech Gesture generation. In *Companion Publication of the 2023 International Conference on Multimodal Interaction (ICMI '23 Companion)*. Association for Computing Machinery.

[22] Gwantae Kim, Seonghyeok Noh, Insung Ham, and Hanseok Ko. 2023. MPE4G: Multimodal Pretrained Encoder for Co-Speech Gesture Generation. In *ICASSP*

[23] Geunmo Kim, Jaewoong Yoo, and Hyedong Jung. 2023. Co-Speech Gesture Generation via Audio and Text Feature Engineering. In *Companion Publication of the 2023 International Conference on Multimodal Interaction (ICMI '23 Companion)*. Association for Computing Machinery.

[24] Vladislav Korzun, Anna Beloborodova, and Arkady Ilin. 2023. The FineMotion entry to the GENEA Challenge 2023: DeepPhase for conversational gestures generation. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI '23)*. ACM.

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

[26] Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. 2021. A large, crowdsourced evaluation of gesture generation systems on common data: The GENEA Challenge 2020. In *Proceedings of the ACM International Conference on Intelligent User Interfaces (IUI '21)*. 11–21. https://doi.org/10.1145/3397481.3450692

[27] 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)*. ACM.

[28] Taras Kucherenko, Pieter Wolfert, Youngwoo Yoon, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2023. Evaluating gesture-generation in a large-scale open challenge: The GENEA Challenge 2022. *arXiv preprint arXiv:2303.08737* (2023).

[29] Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha S. Srinivasa, and Yaser Sheikh. 2019. Talking With Hands 16.2M: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV '19)*. IEEE, 763–772. https://doi.org/10.1109/ICCV.2019.00085

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

[31] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022. 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)*. Springer, 612–630.

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

[33] 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)*. ISCA.

[34] Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. 2023. A Comprehensive Review of Data-Driven Co-Speech Gesture Generation. In *Computer Graphics Forum*, Vol. 42. Wiley Online Library, 569–596.

[35] David Obremski, Helena Babette Hering, Paula Friedrich, and Birgit Lugrin. 2022. Exploratory Study on the Perception of Intelligent Virtual Agents With Non-Native Accents Using Synthetic and Natural Speech in German. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI '22)*. 15–24.

[36] 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)*. IEEE, 573–581. https://doi.org/10.1109/VR50410.2021.00082

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

[38] Viktor Schmuck, Nguyen Tan Viet Tuyen, and Oya Celiktutan. 2023. The KCL-SAIR team's entry to the GENEA Challenge 2023 Exploring Role-based Gesture Generation in Dyadic Interactions: Listener vs. Speaker. In *Companion Publication of the 2023 International Conference on Multimodal Interaction (ICMI '23 Companion)*. Association for Computing Machinery.

[39] Rodolfo Luis Tonoli, Leonardo Boulitreau de Menezes Martins Marques, Lucas Hideki Ueda, and Paula Paro Dornhofer Costa. 2023. Gesture Generation with Diffusion Models Aided by Speech Activity Information. In *Companion Publication of the 2023 International Conference on Multimodal Interaction (ICMI '23 Companion)*. Association for Computing Machinery.

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

[41] Jonathan Windle, Iain Matthews, Ben Milner, and Sarah Taylor. 2023. The UEA Digital Humans entry to the GENEA Challenge 2023. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI '23)*. ACM.

[42] Pieter Wolfert, Jeffrey M. Girard, Taras Kucherenko, and Tony Belpaeme. 2021. To rate or not to rate: Investigating evaluation methods for generated co-speech gestures. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI '21)*. ACM, 494–502. https://doi.org/10.1145/3462244.3479889

[43] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, Ming Cheng, and Long Xiao. 2023. DiffuseStyleGesture: Stylized Audio-Driven Co-Speech Gesture Generation with Diffusion Models. *arXiv preprint arXiv:2305.04919* (2023).

[44] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, and Haolin Zhuang. 2023. QPGesture: Quantization-Based and Phase-Guided Motion Matching for Natural Speech-Driven Gesture Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2321–2330.

[45] Sicheng Yang, Haiwei Xue, Zhensong Zhang, Minglei Li, Zhiyong Wu, Xiaofei Wu, Songcen Xu, and Zonghong Dai. 2023. The DiffuseStyleGesture+ entry to the GENEA Challenge 2023. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI '23)*. ACM.

[46] 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.

[47] 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. https://doi.org/10.1145/3414685.3417838

[48] 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)*. IEEE, 4303–4309. https://doi.org/10.1109/ICRA.2019.8793720

[49] 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. https://doi.org/10.1145/3472749.3474789

[50] 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)*. ACM.

[51] Weiyu Zhao, Liangxiao Hu, and Shengping Zhang. 2023. DiffuGesture: Generating Human Gesture From Two-person Dialogue With Diffusion Models. In *Companion Publication of the 2023 International Conference on Multimodal Interaction (ICMI '23 Companion)*. ACM.

[52] Zeyu Zhao, Nan Gao, Zhi Zeng, Guixuan Zhang, Jie Liu, and Shuwu Zhang. 2023. Gesture Motion Graphs for Few-Shot Speech-Driven Gesture Reenactment. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI '23)*. ACM.