The GENEA Challenge 2020: Benchmarking gesture-generation systems on common data

Taras Kucherenko*
tarask@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

Patrik Jonell*
pjjonell@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

Youngwoo Yoon*
youngwoo@etri.re.kr
ETRI / KAIST
Daejeon, South Korea

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

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

ABSTRACT

Automatic gesture generation is a field of growing interest, and a key technology for enabling embodied conversational agents. Research into gesture generation is rapidly gravitating towards data-driven methods. Unfortunately, individual research efforts in the field are difficult to compare: there are no established benchmarks, and each study tends to use its own dataset, motion visualisation, and evaluation methodology. To address this situation, we launched the GENEA gesture-generation challenge, wherein participating teams built automatic gesture-generation systems on a common dataset, and the resulting systems were evaluated in parallel in a large, crowdsourced user study. Since differences in evaluation outcomes between systems now are solely attributable to differences between the motion-generation methods, this enables benchmarking recent approaches against one another and investigating the state of the art in the field. This paper provides a first report on the purpose, design, and results of our challenge, with each individual team's entry described in a separate paper also presented at the GENEA Workshop. Additional information about the workshop can be found at genea-workshop.github.io/2020/.

KEYWORDS

gesture generation, conversational agents, evaluation paradigms

1 INTRODUCTION

In human communication, nonverbal behaviour plays a key role in conveying a message. A large part of nonverbal behaviour consists of so called co-speech gestures, spontaneous hand gestures that have a close relation with the content of speech, and have shown to improve understanding [8]. Embodied conversational agents (ECAs) fare well with gesticulation, as gesticulation improves interaction with for example social robots [21]. For ECAs to have gesticulation, knowledge of how and when to gesture is needed.

Producing gestures used to be based on rule-based systems, but more recently, data-driven approaches are emerging. Recent work [1, 12, 16, 27] shows the improvement in gesticulation production for ECAs. However, the results of these studies are not one-by-one comparable since a variety of objective and subjective evaluation metrics are used. In this paper, we present the GENEA Challenge 2020. The aim of the challenge is not to select the best team, but to be able to compare different approaches and outcomes, which could lead to new standardized evaluation and generation methods. Unique for this field is the cross-comparison of different systems by different researchers on one and the same dataset.

The contributions of the Gesture Generation Challenge are:

(1) Evaluating several state-of-the-art gesture generation models on a common dataset with a common rendering pipeline.
(2) Two large-scale user studies evaluating human-likeness and appropriateness of submitted motions.
(3) Bringing together researchers in order to advance the state of the art in gesture generation.
(4) Making code and other challenge material publicly available in the spirit of reproducible research.[1]

2 RELATED WORK

Most of the previous work proposing new gesture generation methods, presented evaluation results to show the merits of their methods. There is no widely accepted objective measure due to the highly subjective aspect of human gestures, so most of them conducted human assessment. However, previous subjective evaluations had several drawbacks, the main ones being: the number of systems being compared and the study scale. Proposed models were compared only with a few previous or ablated models [16, 20]. Yoon et al. [26, 27] failed to show statistical significance for the majority of pairs of compared systems, due to the low number of evaluation participants.

Wolfert et al. [25] conducted a benchmarking user study for beat gestures. They compared the data-driven gesture generation method [15] and manually crafted beat gestures. However, only one generation model using speech audio context was used and the video stimuli were realized as stick figures which made it difficult to assess gestures.

Although there is no directly related work on challenges that benchmark co-speech gestures in ECAs, other fields have shown to do well with challenges to standardize evaluation techniques and benchmarks. For example, the Blizzard Challenge [14] has since its inception in 2005 lead to improvements in research on text-to-speech generation. Participants in the Blizzard Challenges are provided a dataset of speech audio and associated text transcriptions, and use these to build a synthetic voice (text-to-speech

*Equal contribution and joint first authors.

[1] See zenodo.org/communities/genea2020/.

system). This challenge is defined by the use of common data and its open participation. This has lead to the development of new and novel methods, driven by past results, and since participants had access to the same data, great steps have been made.

Challenges are active also in the computer vision community. In the CLIC [23] and AIM [18] challenges, participating systems for image compression and super-resolution were compared, by incorporating subjective human assessment similar to the challenge described in this paper.

3 TASK

We pose the problem of speech-driven gesture generation as follows: given a sequence of speech features $\boldsymbol{s}$ (which could involve either audio or text or the combination of the two) the task is to generate a corresponding pose sequence $\hat{\boldsymbol{g}}$ of gestures that an agent might perform while uttering this speech.

The task we focus on with the GENEA Challenge was to compare recent data-driven gesture generation in a fair way. To do so we made sure that every system participating in the evaluation was trained on the same gesture-speech dataset and was visualized on the same virtual avatar.

3.1 Dataset

We used the Trinity Speech-Gesture Dataset [4], comprising 244 minutes of audio and motion capture recordings of one male actor speaking freely on a variety of topics. We removed lower-body data, retaining 15 upper-body joints out of the original 69. Finger motion was also removed due to poor capture quality.

To obtain verbal information from the speech, we first transcribed the audio recordings using Google Cloud automatic speech recognition (ASR) [6], followed by a thorough manual review to correct recognition errors and add punctuation for both the training and test parts of the dataset. All names of non-fictive persons were removed and replaced by unique tokens. The data used in the challenge has been made publicly available in the original dataset repository[2].

For a better understanding of the vocabulary used in the database a table of word frequencies is provided at tinyurl.com/y22h6rtt.

3.2 Challenge Rules

3.2.1 Limits of participation. Each participating team could only submit one system per team for evaluation.

3.2.2 Challenge timeline. We followed the following timeline:

(1) 1st July 2020 – Challenge dataset released to participants
(2) 7th Aug 2020 – Test inputs released to participants
(3) 15th Aug 2020 – Participants submit generated gestures

After generated gestures were submitted we conducted the evaluations.

3.2.3 Synthesising test motion. Synthetic gesture motion was submitted at 20 frames per second (fps) in a format otherwise identical to that used by the challenge gesture database (BVH format, same skeleton, etc.). To prevent optimising for the specific evaluation used in the challenge and to encourage motion generation approaches with long-term stability, participants were asked to synthesise motions for 20 min of test speech in long contiguous segments, from which a subset of clips were extracted for the user studies, similar to many Blizzard Challenges. Manually tweaking of the output motion was not allowed, since the idea was to evaluate how systems would perform in an unattended setting. All the submitted motion in BVH format is publicly available at zenodo.org/record/4088324.

4 SYSTEMS AND TEAMS

We recruited challenge participants from a public call for participation. Sixteen teams were signed up for the challenge. Five teams completed the challenge and the remaining 11 teams dropped out. Two of the withdrawing teams explained it was due to their lack of capacity to complete the challenge and to their unsatisfactory results; there were no reported withdrawals due to the challenge data or task.

The challenge evaluation contained 9 different *conditions*: 2 toplines (which are the best possible motion in terms of naturalness or appropriateness), 2 previously published *baselines*, and 5 challenge *entries/submissions*. Table 1 lists all conditions, together with participating team names and (abbreviated) affiliations. We anonymized the teams in the present paper by not revealing team ID assignments.

The two toplines were:

N Natural motion capture from the actor for the input speech segment in question. Surpassing this system would essentially entail superhuman performance.

M *Mismatched* natural motion capture from the actor, corresponding to another speech segment than that played together with the video. This was accomplished by permuting the motion segments from condition N in such a way that no segments remained in its original position. This represents the performance attainable by a system that produces very high-quality motion (same as N, so a topline), but whose behaviour is completely unrelated to the speech (making it a kind of bottom line).

Since there has been no previous general study that compares systems to each other and what the state of the art is, it is hard to choose a "best" baseline system. Therefore the choice was more subjective and based on code availability. The two baseline systems we compared against were chosen from recent data-driven gesture generation papers which had their code available and were easy to reproduce. The baselines used were the following:

BA The system from [15], which only takes speech audio into account when generating system output. This model uses a chain of two neural networks: one maps from speech to pose representation and another decodes representation to pose.

BT The system from [27], which only takes text transcript information (which includes word timing information) into account when generating system output. This model consists of a encoder for text understanding and a decoder for frame-by-frame pose generation.

We also intended to have another baseline from Ginosar et al. [5], but we could not get any good results with this system using our dataset, and it was therefore omitted from the evaluation.

[2] trinityspeechgesture.scss.tcd.ie

These baselines systems were taken directly from previous gesture-generation publications [15, 27] and adapted to the challenge dataset by their original authors. Source code and hyperparameters for both baseline systems are available on Github[3]. These implementations and hyperparameters were also made available to participating teams during the challenge.

In BA, the representation of upper-body poses in the challenge dataset was different from the dataset used in the original publication and hence a new hyperparameter search was conducted to find optimal hyperparameters. Another change was that the resulting motion was represented using exponential map and was smoothed using Savitzky–Golay filter [22] with window length 9 and polynomial order 3.

In BT, the representation of upper-body poses in the challenge dataset was different to the TED dataset used in the original publication. Accordingly, the pose representation was changed from 2D Cartesian coordinates of 8 upper-body joints to 3x3 rotational matrices for each of 15 joints. The data dimension for a pose was 135 (3x3x15). The number of layers and loss function were the same to the original paper. The hyperparameters of learning rate and loss term weights were adjusted manually. Also, pretrained FastText word vectors [2] were used instead of GloVe [19].

The individual challenge entries are described in detail in separate workshop papers submitted by the participating teams and published in the GENEA Workshop 2020 proceedings.[4]

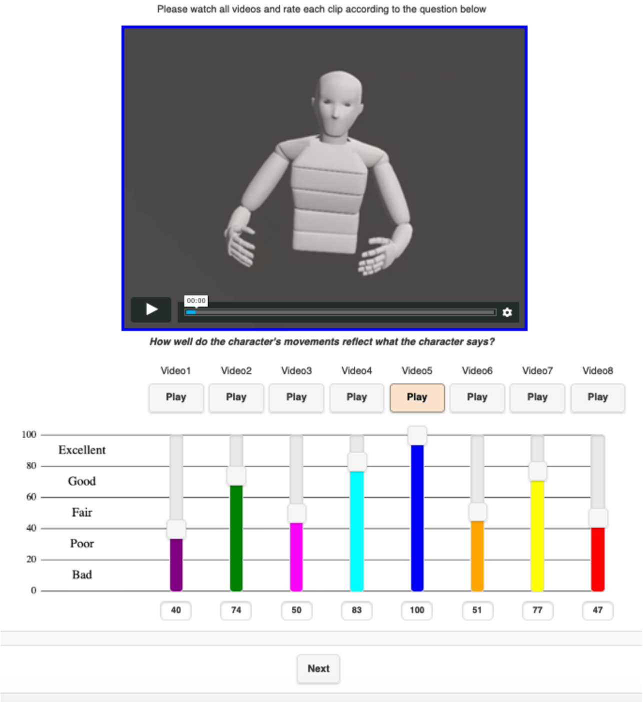
5 EVALUATION

We conducted a large-scale, crowdsourced, joint, and parallel evaluation of the motion submitted by the participating teams, along with some other conditions. The evaluation focused on gesture quality of the various submitted systems. The systems were evaluated in terms of human-likeness of the motion as well as appropriateness of the gestures for a given input speech. The central difference from other gesture-generation evaluations is that all systems in the GENEA evaluation used the same motion data, the same visualisation/embodiment, and were scored 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.

Jonell & Kucherenko et al. [13] recently showed that the results from crowdsourcing evaluations were not significantly different from in-lab evaluations in terms of results and consistency. We therefore adopted an entirely crowdsourced approach, as opposed to for example the Blizzard Challenge, which has used a mixed approach. We employ attention checks as a means of finding participants which were not paying attention (explained below).

5.1 Stimuli

The organisers selected 40 non-overlapping speech segments from the test inputs (average segment duration 10 s) to use in the user-study evaluation. These speech segments were selected across the test inputs to be full and/or coherent phrases. The motion from the corresponding intervals in the BVH files submitted by participating teams was extracted and converted to a motion video clip using the visualisation server (described in Section 5.1.2) provided to participants (albeit at a higher resolution of 960×540). All the stimuli used are publicly available at zenodo.org/record/4080919.



Please watch all videos and rate each clip according to the question below

How well do the character's movements reflect what the character says?

Figure 1: A screenshot of a page with stimuli from the evaluation interface. The question asked in the image ("How well do the character's movements reflect what the character says?") is from a pre-study used to validate the evaluation paradigm on stimuli with known differences from a previous work, and was changed for each of the two evaluations detailed in this text.

5.1.1 Virtual avatar. We used the same virtual avatar for all renderings during the challenge (during the challenge, and for the evaluation). The avatar can be seen in Fig 1. The avatar had originally 71 joints (full body including fingers) but only 15 joints, corresponding to the upper body and excluding fingers, were used for the challenge. The hands and fingers had a static pose, in which the hands were lightly cupped (see Fig 1).

5.1.2 Visualization server. For the challenge we developed a visualization server for the participants. The purpose was for all the participating teams to be able to produce visualizations which were identical (except of the resolution) with the stimuli that were going to be evaluated. The visualization was implemented using a python-based web-server which interfaced Blender 2.83[5]. The participants were able to send a BVH file to the visualization server and it would be put in a queue and processed in order as soon as a

[3] BA: github.com/GestureGeneration/Speech_driven_gesture_generation_with_autoencoder/tree/GENEA_2020
BT: github.com/youngwoo-yoon/Co-Speech_Gesture_Generation

[4] See zenodo.org/communities/genea2020/.

[5] www.blender.org

			Inputs used?		Representation or features		Stochastic
Name or description	Origin	ID	Aud.	Text	Input speech	Output motion	output?
Natural motion	-	N	✓	✓	–	–	✓
Mismatched motion	-	M	✗	✗	–	–	✓
Audio-based baseline	Kucherenko et al. [15]	BA	✓	✗	MFCC	Exp. map	✗
Text-based baseline	Yoon et al. [27]	BT	✗	✓	FastText†	Rot. matrix	✗
AlltheSmooth	CSTR lab, UEDIN, Scotland	S...	✓	✗	MFCC	Joint pos.	✗
Edinburgh CVGU	CVGU lab, UEDIN, Scotland	S...	✓	✓	BERT† and mel-spectrogram	Rot. matrix	✓
FineMotion	ABBYY lab, MIPT, Russia	S...	✓	✓	GloVe† and mel-spectrogram	Exp. map	✗
Nectec	HCCR unit, NECTEC, Thailand	S...	✓	✓	Phoneme, Spacy word vectors†, and audio features	Exp. map	✗
StyleGestures	TMH division, KTH, Sweden	S...	✓	✗	Mel-spectrogram	Exp. map	✓

Table 1: Conditions participating in the evaluation. Teams are sorted alphabetically by name. The anonymised IDs of submitted entries begin with the letter 'S' followed by a second, randomly-assigned letter in the range A through E, but which letter is associated which each team is not revealed in order to preserve anonymity. † indicates a use of word vectors pretrained on external data.

rendering-worker became available. The same visualization server was used for the final stimuli, however, the resolution was increased to 960×540 instead of 480×270. The lower resolution was used in order to increase performance and throughput of the visualization server, since there were sixteen teams competing initially. The input to the visualization server was expected to be 20 fps. Upon paper publication, a link to the code for the visualization server will be provided.

5.2 Evaluation interface

In order to efficiently evaluate a large number of relatively similarly-performing systems in parallel, we used a methodology inspired by the MUSHRA test (MUltiple Stimuli with Hidden Reference and Anchor) [10], which is a standard published by the International Telecommunication Union (ITU). However, there are numerous differences between a MUSHRA test for audio and our evaluation, for instance the reference and anchor are both missing, corresponding to the letters R and A.

Fig. 1 shows an example of the user interface used for rating stimuli in the subjective evaluation. The participants were first met with a screen with instructions and how to use the evaluation interface. They were then presented with 10 pages where on each page they would evaluate all of the participating systems, toplines, and baselines, as seen in the figure. The order of the conditions was generated randomly and it was not identified which was which. Randomly coloured sliders and video borders were used to make it easier to know which slider the currently-playing video was associated with. After the 10 pages of stimuli, raters were presented with a page asking for demographics and their experience of the test.

As can be seen in Fig. 1, the 100-point rating scale was anchored by dividing it into successive 20-point 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 for Mean Opinion Score (MOS) [11] tests, another evaluation standard developed by the ITU.

5.3 Study design

Each study was balanced using a greedy procedure such that each segment appeared on each page with approximately equal frequency (segment order), and each condition was associated with each slider with approximately equal frequency (condition order). For any given participant and study, each page would use different speech segment. Every page would contain condition N and (where relevant) condition M, but one other condition was randomly omitted from each page to limit the maximum number of sliders on a page to 8 or less, depending on the study.

Three attention checks were incorporated into the pages for each study participant. These either displayed a brief text message over the gesticulating avatar reading "Attention! Please rate this video XX.", or they temporarily replaced the audio with a synthetic voice speaking the same message. XX would be a number from 5 to 95, and the participant had to set the corresponding slider to the requested value, plus or minus 3, to pass the attention check. The numbers 13 through 19, as well as even multiples of 10 from 30 to 90, were not used for attention checks due to their acoustic ambiguity. Which sliders on which pages that were used for attention check was uniformly random, except that no page had more than one attention check, and condition N and M were never replaced by attention checks.

We evaluated two aspects of the gesture motion, each in a separate study:

Human-likeness This study asked participants to rate "How human-like does the gesture motion appear?", with the intention of measuring the quality of the generated motion while ignoring its link to the input speech. This study did not have speech in the stimulus videos (they were silent).

Appropriateness This study asked participants to rate "How appropriate are the gestures for the speech?" This was intended to investigate the perceived link between motion and speech (both in terms of rhythm/timing and semantics), ignoring motion quality as much as possible. This study contained speech audio in the stimulus videos.

5.4 Test-participant recruitment

Study participants were recruited through the crowdsourcing platform Prolific (formerly Prolific Academic), restricted to a set of English-speaking countries (UK, IE, USA, CAN, AUS, NZ). There was no requirement to be a native speaker of English, since Prolific does not support screening participants based on that criterion. A participant could take either study or both studies, but not more than once each. Participants were remunerated 5.75 GBP for completing the human-likeness study (median time 33 min) and 6.50 for the appropriateness study (median time 34 min).

5.5 Objective evaluation metrics

Since subjective evaluation is costly and time-consuming it would be beneficial for the field to agree on objective evaluations to be used. As a step in this direction we use the numerical measures introduced in previous work. Namely we use two numerical objective evaluation measures:

Average jerk used extensively to evaluate motion smoothness [15, 17, 24]. We report average values of absolute jerk (defined using finite differences) for different motion segments.

Distance between velocity histograms used previously to evaluate gesture quality [15, 16] since well trained models should produce similar motion properties as the actor it was trained on and hence it should have a similar motion speed profile.

5.5.1 Average jerk. The third derivative of the coordinates $x(t)$ is called jerk and can be mathematically formulated as following: $jerk(x) = x'''(t)$. In our experiments we report absolute average values of jerk for different motion segments.

5.5.2 Comparing velocity histograms. This metric is based on the assumption that synthesised motion should follow a similar velocity distribution as the ground truth motion. To evaluate this we calculate velocity distribution histograms for all the systems and compare them to the velocity distribution of the ground truth by calculation Hellinger distance between the two: $H(h^1, h^2) = \sqrt{1 - \sum_i \sqrt{h^1_i \cdot h^2_i}}$.

For both of the numerical evaluation above the motion has been first converted from joint angles to 3D coordinates that are publicly available at zenodo.org/record/4088319. To enhance reproducibility the code for numerical evaluations is also available, at https://github.com/Svito-zar/genea_numerical_evaluations.

6 RESULTS OF CHALLENGE EVALUATION

This section describes the results of the subjective and objective evaluations, with discussion and interpretation of the results reserved for Sec. 7. First, Sec. 6.1 introduces demographic and other information gathered from the recruited participants. Sec. 6.2 then reports the results of the subjective evaluation of challenge conditions, which also are visualised in a number of different figures. Sec. 6.3 complements the subjective findings with numerical results that quantify different aspects of the motion evaluated in the challenge.

Table 2: Summary statistics of user-study ratings for all conditions in the two studies, with 0.01-level confidence intervals. The human-likeness of M was not evaluated explicitly, but is expected to be very close to N since it uses the same motion clips.

	Human-likeness		Appropriateness	
ID	Median	Mean	Median	Mean
N	72 ∈ [70, 75]	67.6 ± 1.8	81 ∈ [79, 83]	73.8 ± 1.8
M	"	"	56 ∈ [53, 59]	53.3 ± 2.0
BA	46 ∈ [44, 49]	46.2 ± 1.7	40 ∈ [38, 41]	40.4 ± 1.8
BT	55 ∈ [53, 58]	54.6 ± 1.8	38 ∈ [35, 40]	38.5 ± 1.9
SA	38 ∈ [35, 41]	40.1 ± 1.9	35 ∈ [31, 37]	36.4 ± 1.9
SB	52 ∈ [50, 55]	52.8 ± 1.9	43 ∈ [40, 45]	43.3 ± 2.0
SC	57 ∈ [55, 60]	55.8 ± 1.9	50 ∈ [48, 52]	50.6 ± 1.9
SD	60 ∈ [57, 61]	58.8 ± 1.7	49 ∈ [46, 50]	48.1 ± 1.9
SE	49 ∈ [47, 51]	49.6 ± 1.8	47 ∈ [44, 49]	45.9 ± 1.8

6.1 Data on test participants

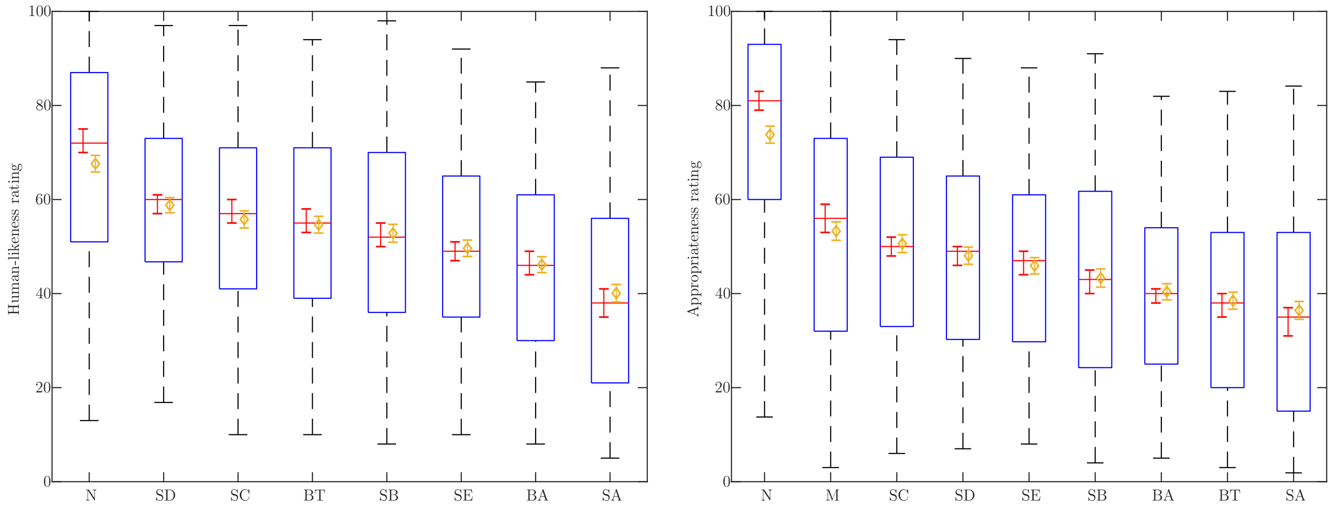
Each of the two user studies recruited 125 participants that passed all attention checks they encountered. In study 1 on Human-Likeness, the average age was 31.50 years (SD: 10.7), with 66 men, 57 women, and 2 others. There were in total 116 native speakers, and 9 non-native speakers of English. We asked participants on which continent they lived, and 69 participants were from Europe, 1 from Africa, 48 from North America, 2 from South America, and 5 from Asia. In study 2 on Appropriateness, the average age was 31.14 years old (SD: 11.7), with 60 men, 64 women, and 1 other. We asked participants on which continent they resided, and 78 answered Europe, 1 answered Africa, 39 answered North-America, 3 answered Asia, and 4 answered Oceania.

Participants that did not pass all attention checks (23 test-takers in the human-likeness study and 40 test-takers in the appropriateness study) were omitted from the analysis, as were scores from sliders used for attention check. The median successful completion time for the rating portion of the study (excluding reading instructions and answering the post-test questionnaire) was 24 minutes for the human-likeness study and 27 minutes for the appropriateness study, with the shortest successful completion times being 12 minutes in both studies.

6.2 Analysis and results of subjective evaluation

The study results are available at at zenodo.org/record/4088250. Summary statistics (sample median and sample mean) for all conditions in each of the two studies are shown in Table 2, together with a 99% confidence interval for the true median/mean. The confidence intervals were computed either using a Gaussian assumption for the means (i.e., with Student's t-distribution cdf, and rounded outward to ensure sufficient coverage), or using order statistics for the median (leverages the binomial distribution cdf, cf. [7]).

The ratings distributions in the two studies are further visualised through box plots in Fig. 2. The distributions are seen to be quite broad. This is common in MUSHRA-like evaluations, since the range of numbers not only reflects differences between systems, but also



(a) Human-likeness ratings

(b) Appropriateness ratings

Figure 2: Box plots visualising the ratings distribution in the two studies. Red bars are the median ratings (each with a 0.01 confidence interval); yellow diamonds are mean ratings (also with a 0.01 confidence interval). Box edges are at 25 and 75 percentiles, while whiskers cover 95% of all ratings for each system. Conditions are ordered descending by sample median, which leads to a different order in each of the two plots.

variation, e.g., between stimuli, in individual preferences, and in how harsh 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.

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 sample median between different conditions, we applied two-sided pairwise Wilcoxson signed-rank tests to all pairs of distinct conditions in each study. This closely follows the analysis methodology used throughout recent Blizzard Challenges. Unlike Student's t-test, this test does not assume that rating differences follow a Gaussian distribution, which would be inappropriate as we can see from the box plots in Fig. 2 that ratings distributions are skewed and thus likely non-Gaussian. For each condition pair, only pages for which both conditions were assigned valid scores were included in the analysis. Recall that not all systems were scored on all pages due to the limited number of sliders and the presence of attention checks. This meant that every statistical significance test was based on at least 796 pairs of valid ratings in each of the studies. The p-values computed in the significance tests were adjusted for multiple comparisons using the Holm-Bonferroni method [9] (which is uniformly more powerful than regular Bonferroni correction) to keep the family-wise error rate (FWER) at or below 0.01 in each of the two studies. This statistical analysis found all but 4 out of 28 condition pairs to be significantly different in the human-likeness study, which the corresponding numbers being 7 out of 36 condition pairs in the appropriateness study. Which conditions that were found to be rated significantly above or below which other conditions in the two studies is visualised in Fig. 3.

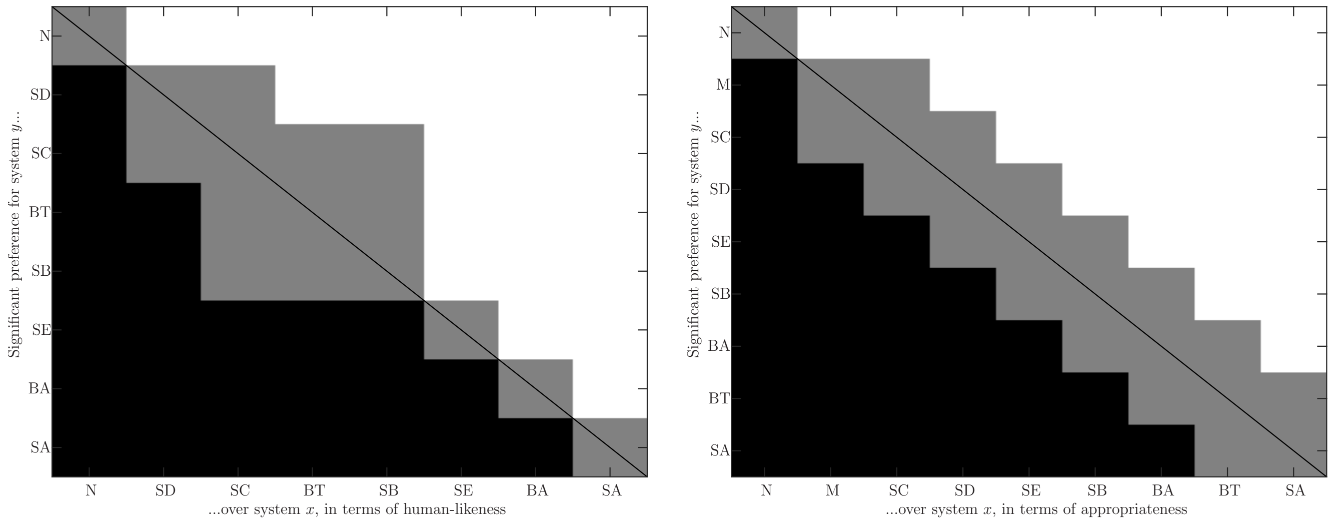
System	Jerk	Hell. dist. (left wrist)	Hell. dist. (right)
N	151.52 ± 35.57	0	0
BA	65.59 ± 4.42	0.08436	0.09029
BT	45.84 ± 2.14	0.13048	0.09662
SA	132.37 ± 27.64	0.06475	0.05931
SB	189.39 ± 4.66	0.12557	0.11389
SC	84.44 ± 8.48	0.08261	0.08825
SD	72.06 ± 7.91	0.07277	0.06221
SE	97.85 ± 9.34	0.04892	0.04925

Table 3: Results from the objective evaluations.

Finally, we present two diagrams that put the results of the two studies together. Fig. 4, in particular, visualises the relative (partial) ordering between different conditions implied by the results of the two studies in Fig. 3. Although there are similarities, the two orderings are meaningfully different. This suggests that the two studies managed to disentangle aspects of perceived motion quality (human-likeness) from the perceived link between gesture and speech (appropriateness). Second, Fig. 5, visualises confidence regions for the median rating as boxes whose horizontal and vertical extents are given by the corresponding confidence intervals in Table 2. Once again, different systems are found to be good at different things. The numerical gap between natural and synthetic gesture motion is seen to be more pronounced in the case of appropriateness than in the case of the human-likeness.

6.3 Results of objective evaluation

Results of objective evaluations are given in Table 3. The first column contains the average jerk across all the joints. We report



(a) Human-likeness study

(b) Appropriateness study

Figure 3: Significance of pairwise differences between conditions. White means that 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 the 0.01 level after Holm-Bonferroni correction. Conditions are listed in the same order as in Fig. 2, which is different for each of the two studies.

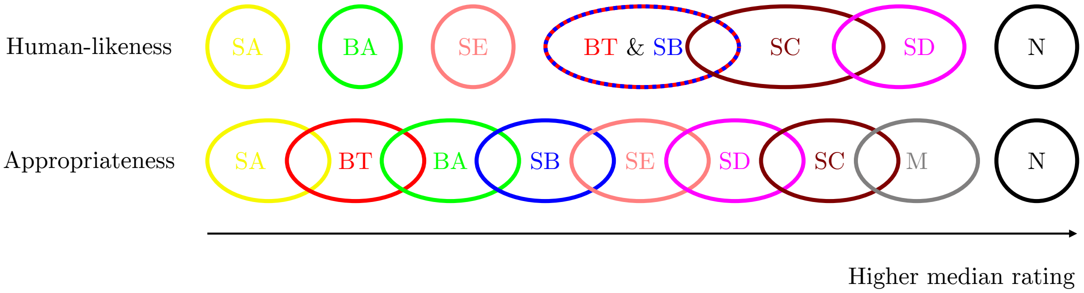


Figure 4: Partial ordering between conditions in the two studies. Each condition is an ellipse; overlapping or (in one case) coinciding ellipses signify that the corresponding conditions were not statistically significantly different in the evaluation. Colours were adapted from [3]. There is no scale on the axis since the figure visualises ordinal information only.

mean and standard deviation for the 20 minutes of test motions. The second and third columns contain Hellinger distance between velocity histogram for left and right wrists. For more details on the evaluation metrics we refer the reader to Section 5.5. We see that different systems performed best (came closest to the natural motion) on different objective measures.

We can see that that objective metrics are inconsistent with the subjective results. While SA showed the most similar jerk to natural motion (N), it was less preferred in the subjective evaluation. Similarly, SE showed the most similar Hellinger distances to N, but it was preferred moderately by the evaluation participants.

We stress that objective evaluation is a complementary measure and subjective evaluation is much more important.

7 DISCUSSION

In this section we analyse the results obtained in our evaluations, the limitations of the challenge, and what we the challenge brings to the scientific community.

7.1 Challenge results

First of all we want to note that gesture generation is a difficult problem which is hard from being solved, since no system could come even close to the natural motion.

It has been shown before that naturalness can strongly influence appropriateness of gestures during subjective evaluations [1, 16], and in our experiments we managed to separate the two only partially. From one side, we can observe at Figure 5 that different systems were good at different things: some scored better in naturalness and others in appropriateness. From another side, mismatched motion and system SC (which uses only audio) were preferred over

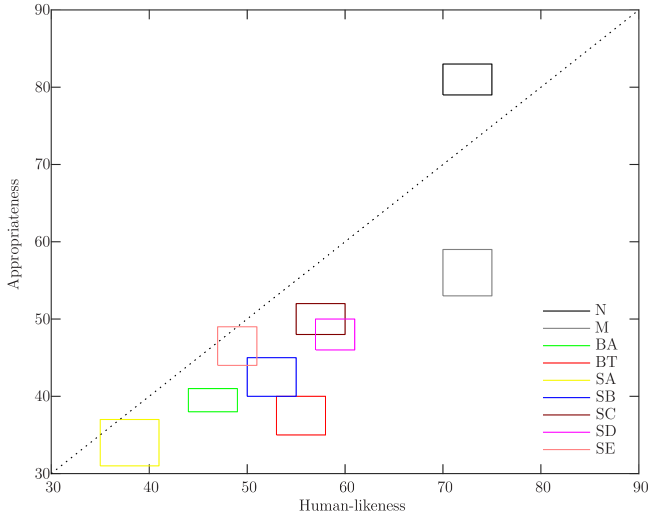


Figure 5: Confidence regions for the true median rating across both studies. The dotted black line is the identity, $x = y$. While the human-likeness (x-coordinate) of M was not evaluated directly, it is expected to be very close to N since it uses the same motion clips, and the horizontal extent of the confidence region for M was therefore copied from N.

all the systems in terms of appropriateness. This indicates that our evaluation probably did not capture semantic appropriateness well, since the question was slightly vague.

The gap between natural motion and that synthesis by machine learning models is greater in terms of appropriateness than in naturalness. This could indicate that appropriateness is a harder problem. It could be that available data may not allow to learn to generate appropriate gesture because the dataset is too small or the vocabulary used it too wide.

7.2 Limitations

We took the first step for benchmarking different gesture generation systems on a common dataset and stimuli, but our crowdsourced evaluation had a few limitations.

First, in measuring appropriateness of gestures (i.e., link between gestures and speech), semantic and rhythmic appropriateness was mixed together and there was no way to determine by which aspect of appropriateness the participants rated. In addition, rating on the appropriateness could be affected to some extent by motion quality even if we asked participants to rate regardless of motion quality.

Second, the dataset used in the challenge was limited to a single English speaker in a monologue scenario. The role of gesticulation may be expected to differ between different persons and languages as well as the speaking environment (e.g., conversation versus monologue). Benchmarking gestures in more complex scenarios of multiple speakers, multiple languages, and diverse environments should be considered in the future.

Another limitation is that we considered only upper-body gestures despite the fact that whole-body gestures including posture, stepping motion and stance, facial expression, and hand motion are important too in social interactions. Three teams stated that the most needed extension is to include whole-body or facial gestures. Some evaluation participants also found the absence of facial and finger motion to be a limitation of the challenge.

7.3 Lessons learned from the challenge

We have learned several things from the Challenge:

- A MUSHRA-like evaluation paradigm can be successfully used to benchmark multiple gesture generation models in parallel.
- Being human-like does not mean being appropriate for gestures of a virtual avatar, and synthetic systems can be strong and weak in different aspects.
- There is a need for future challenges, since there a big gap remains between natural and synthesised motion.
- Providing carefully pre-processed data and good code infrastructure helps challenge participants to focus on developing their system, instead of solving unrelated issues.

8 CONCLUSIONS

We have hosted a challenge to assess the state of the art in data-driven co-speech gesture generation. The central design goal of the challenge was to enable direct comparison between many different gesture-generation methods while controlling for factors of variation external to the model, namely data, embodiment, and evaluation methodology. Our results suggest that the field is advancing, since most submissions performed significantly better than the baselines published last year. Different systems were also found to be good at different things, on the two scales (quality and appropriateness) that we assessed. However, a substantial gap remains between synthetic and natural gesture motion, indicating that gesture generation is far from a solved problem.

We believe that the standardised challenge training and test sets of time-aligned audio, text and gestures, and the associated library of rated motion clips from the challenge provided at zenodo.org/communities/genea2020, will be useful for benchmarking future gesture-generation methods. Furthermore, we think challenges like the one described here are poised to play an important role in identifying key factors for convincing gesture generation in practice, and in driving and validating future progress toward the goal of endowing embodied agents with natural gesture motion.

We encourage the reader to seek out the system-description papers from the GENEA Workshop for additional lessons about system building and important factors in performance, as told by the team preparing the challenge entries.

ACKNOWLEDGMENTS

The GENEA Challenge 2020 used the Trinity Speech-Gesture Dataset collected by Ylva Ferstl and Rachel McDonnell. The challenge dataset was further processed by Taras Kucherenko, Simon Alexanderson, Jonatan Lindgren, and Jonas Beskow at KTH Royal Institute of Technology and by Pieter Wolfert at Ghent University.

The authors wish to thank Simon King for sharing his insights and experiences from running the Blizzard Challenge in speech synthesis. We are also grateful to Ulysses Bernardet for input and to Kevin El Haddad, Ulme Wennberg, Dmytro Kalpakchi, Bram

Willemsen, Jonas Beskow, and André Tiago Abelho Pereira for feedback on the paper preprint.

This research was partially supported by Swedish Foundation for Strategic Research contract no. RIT15-0107 (EACare), by IITP grant no. 2017-0-00162 (Development of Human-care Robot Technology for Aging Society) funded by the Korean government (MSIT), the Flemish Research Foundation grant no. 1S95020N, and by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

REFERENCES

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

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

[3] Robert M. Boynton. 1989. Eleven colors that are almost never confused. In *Proceedings of the SPIE Symposium: Human Vision, Visual Processing, and Digital Display*, Vol. 1077. SPIE International Society for OpticalEngineering, 322–332.

[4] Ylva Ferstl and Rachel McDonnell. 2018. Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the International Conference on Intelligent Virtual Agents*. ACM.

[5] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning Individual Styles of Conversational Gesture. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*. IEEE.

[6] Google. 2020. Google Cloud Speech-to-Text. https://cloud.google.com/speech-to-text Accessed: 2020-10-09.

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

[8] 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 & Review* 25, 5 (2018), 1900–1908.

[9] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6, 2 (1979), 65–70.

[10] International Telecommunication Union, Radiocommunication Sector 2015. *Method for the subjective assessment of intermediate quality levels of coding systems*. International Telecommunication Union, Radiocommunication Sector, Geneva, Switzerland. https://www.itu.int/rec/R-REC-BS.1534-3-201510-I/

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

[12] Ryo Ishii, Taichi Katayama, Ryuichiro Higashinaka, and Junji Tomita. 2018. Generating Body Motions Using Spoken Language in Dialogue. In *Proc. IVA* (Sydney, NSW, Australia) *(IVA '18)*. ACM, New York, NY, USA, 87–92.

[13] Patrik Jonell, Taras Kucherenko, Ilaria Torre, and Jonas Beskow. 2020. Can we trust online crowdworkers? Comparing online and offline participants in a preference test of virtual agents.. In *Proceedings of the ACM International Conference on Intelligent Virtual Agents (IVA'20)*.

[14] Simon King. 2014. Measuring a decade of progress in text-to-speech. *Loquens* 1, 1 (2014), 006.

[15] Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. 2019. Analyzing input and output representations for speech-driven gesture generation. In *Proceedings of the ACM International Conference on Intelligent Virtual Agents (IVA'19)*. ACM, New York, NY, USA, 97–104. https://doi.org/10.1145/3308532.3329472

[16] 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*.

[17] Pietro Morasso. 1981. Spatial control of arm movements. *Experimental brain research* 42, 2 (1981), 223–227.

[18] AIM Challenge organisers. 2020. AIM: Advances in Image Manipulation workshop and challenges on image and video manipulation. https://data.vision.ee.ethz.ch/cvl/aim20/ Accessed: 2020-10-05.

[19] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing*. 1532–1543.

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

[21] 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 *2011 Ro-Man*. IEEE, 247–252.

[22] Abraham Savitzky and Marcel J. E. Golay. 1964. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry* 36, 8 (1964), 1627–1639.

[23] George Toderici, Wenzhe Shi, Radu Timofte, Lucas Theis, Johannes Ballé, Eirikur Agustsson, Nick Johnston, and Fabian Mentzer. 2020. CLIC: Workshop and Challenge on Learned Image Compression. http://www.compression.cc/ Accessed: 2020-10-05.

[24] 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.

[25] Pieter Wolfert, Taras Kucherenko, Hedvig Kjelström, and Tony Belpaeme. 2019. Should beat gestures be learned or designed? A benchmarking user study. In *ICDL-EPIROB 2019 Workshop on Naturalistic Non-Verbal and Affective Human-Robot Interactions*. 1–4.

[26] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity. *ACM Trans. Graph.* 4 (2020).

[27] 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 Robotics and Automation Society, Piscataway, NJ, USA, 4303–4309. https://doi.org/10.1109/ICRA.2019.8793720