A Large, Crowdsourced Evaluation of Gesture Generation Systems on Common Data: The GENEA Challenge 2020

Taras Kucherenko[*]
tarask@kth.se
Division of Robotics, Perception and Learning, KTH Royal Institute of Technology
Stockholm, Sweden

Patrik Jonell[*]
pjjonell@kth.se
Division of Speech, Music and Hearing, KTH Royal Institute of Technology
Stockholm, Sweden

Youngwoo Yoon[*]
youngwoo@etri.re.kr
ETRI & KAIST
Daejeon, Republic of Korea

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

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

ABSTRACT

Co-speech gestures, gestures that accompany speech, play an important role in human communication. Automatic co-speech gesture generation is thus a key enabling technology for embodied conversational agents (ECAs), since humans expect ECAs to be capable of multi-modal communication. 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 Challenge, a 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 using the same motion-rendering pipeline. 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 in order to get a better impression of the state of the art in the field. This paper reports on the purpose, design, results, and implications of our challenge.

CCS CONCEPTS

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

KEYWORDS

gesture generation, conversational agents, evaluation paradigms

[*] Equal contribution and joint first authors.

IUI '21, April 14–17, 2021, College Station, TX, USA
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8017-1/21/04.
https://doi.org/10.1145/3397481.3450692

ACM Reference Format:
Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. 2021. A Large, Crowdsourced Evaluation of Gesture Generation Systems on Common Data: The GENEA Challenge 2020. In *26th International Conference on Intelligent User Interfaces (IUI '21), April 14–17, 2021, College Station, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3397481.3450692

1 INTRODUCTION

This paper is concerned with systems for automatic generation of nonverbal behaviour, and how these can be compared in a fair and systematic way in order to advance the state-of-the-art. This is of importance as nonverbal behaviour plays a key role in conveying a message in human communication [36]. A large part of nonverbal behaviour consists of so called co-speech gestures, spontaneous hand gestures that relate closely to the content of the speech, and that have been shown to improve understanding [17]. Embodied conversational agents (ECAs) benefit from gesticulation, as gesticulation, e.g., improves interaction with social robots [48] and willingness to cooperate with an ECA [46]. Knowledge of how and when to gesture is also needed. This can for example be learned from interaction data; see, e.g., [23] and references therein.

Synthetic gestures used to be based on rule-based systems, e.g., [8, 49]; see [56] for a review. These are gradually being supplanted by data-driven approaches, e.g., [3, 10, 28, 34], with recent work [2, 31, 61] showing improvements in gesticulation production for ECAs. However, the results in prior studies on gesture-generation are not directly comparable. First, prior studies make use of a variety of different evaluation metrics. Second, prior studies rely on different data sources, and train their models on these different sources. Lastly, visualisations of their generated gestures have different avatars and production values, which can obscure the quality of the underlying gesture-generation approach. All these differences are, however, external to the actual methods that drive the gesture generation.

In this paper, we present the GENEA Challenge 2020,[1] the first joint gesture-generation challenge that controls for all previous sources of between-paper variation, by providing a common dataset for building gesture-generation systems, along with common evaluation standards and a shared visualisation procedure. The aim of the challenge is not to select the best team – it is not a contest, nor a competition – but to be able to compare different approaches and outcomes. This makes it possible to assess and advance the state of the art in gesture generation, and measure the gap between it and natural co-speech gestures. Comparing the different methods and their performance also helps identify what matters most in gesture generation, and where the bottlenecks are. Challenge participants benefit by working on the same problem together with researchers interested in the same topic, strengthening the research community, and get an opportunity to compare their systems to other competitive systems in a large and carefully-executed joint evaluation.

Our concrete contributions are:

1. Jointly evaluating several state-of-the-art gesture-generation models on a common dataset using a common 3D model and rendering method.
2. Two large-scale user studies assessing human-likeness and appropriateness of submitted motion.
3. Providing open code and and high-quality data – comprising the pre-processed, multimodal training and test datasets, the standardised visualisation, a large number of subjective responses, and evaluation and analysis using open standards and code – in the spirit of reproducible research.
4. Bringing researchers together in order to advance the state-of-the-art in gesture generation, and enabling future research to compare and benchmark against systems from the challenge.

The remainder of this paper first presents prior work in terms of gesture-evaluation practices (and their shortcomings) and discusses how challenges have helped in other fields. We then describe the challenge setup and its results, and finally turn to consider the implications for future challenges and gesture generation as a whole.

2 RELATED WORK

Most previous work proposing new gesture-generation methods incorporates an evaluation to support the merits of their method. Human gesture perception is highly subjective, and there are currently no widely accepted objective measures of gesture perception, so most publications have conducted human assessments instead. However, previous subjective evaluations, as reviewed in [59], have several drawbacks, with major ones being the coverage of systems being compared and the scale of the studies. Like in [2, 30, 31, 45], proposed models are at most compared to one or two prior approaches (often a highly similar baseline) or possibly only to ablated versions of the same model. A large number of studies do not compare their outcomes with other methods at all. This creates an insular landscape where particular model families only are applied to particular datasets, and never contrasted against one another. As for scale, large evaluations are expensive, and studies may not be be able to recruit enough participants, thus leaving the differences between many pairs of studied systems unresolved and not statistically significant (cf. [60, 61]). Questionnaires, which are one popular evaluation methodology (cf. [4, 21, 47]) demand a lot of time and cognitive effort even before scaling up. In addition, the items used in questionnaires differs across studies and the set of questions used is often not standardised.

Sometimes, evaluations fail to anchor system performance against natural ("ground truth") motion from their database, e.g., [22, 33, 47]. Another significant difference between studies is how generated motion is visualised, where some prior work (e.g., [29, 58]) displays motion through stick figures, or applies it to a physical agent (e.g., [21, 47]). Neither of these may allow the same expressiveness or range of motion as 3D-rendered avatars in, e.g., [2, 31].

Although there is no directly related work on challenges that benchmark co-speech gestures in ECAs, other fields have done well using challenges to standardise evaluation techniques, establish benchmarks, and track and evolve the state of the art. For example, the Blizzard Challenges have since their inception in 2005 (see [5]) helped advance text-to-speech (TTS) technology and identified subtle but robust trends in the specific strengths and weaknesses in different speech-synthesis paradigms [26]. These challenges are open to both academia and industry. Participants are provided a common dataset of speech audio and associated text transcriptions, and use these to build a synthetic voice. The resulting voices are then evaluated in a large, joint evaluation. Challenge data, evaluation stimuli, and subjective ratings remain available after the challenge, and have been widely used both for benchmarking subsequent TTS systems, e.g., [9, 52], and for doing research on the perception of natural and artificial speech, e.g., [13, 37, 38, 50, 62].

Challenges are also actively used in the computer-vision community, for instance for benchmarking purposes. Recent CLIC [54] and NTIRE [41] challenges, for example, compared systems for image compression and super-resolution respectively, also incorporating subjective human assessments similar to the challenge described in this paper (although they used a MOS-like setup, which has been found to be less efficient than the side-by-side evaluation methodology we employ [44]). This addresses the over-reliance on objective metrics in computer-vision evaluation, which, just like in speech quality and gesture generation, do not always align with human perception. Inspired by the successes of challenges in other field of study, we conducted the first challenge in the field of gesture generation.

3 TASK

Our challenge focussed on data-driven gesture generation. We pose the problem of speech-driven gesture generation as follows: given input speech features $\boldsymbol{s}$ – which could involve either an audio waveform (a sequence of pressure samples) or text (a word sequence) or the combination of the two – the task is to generate a corresponding pose sequence $\hat{\boldsymbol{g}}$ describing gesture motion that an agent might perform while uttering this speech. To enable direct comparison of different data-driven gesture-generation methods, all methods evaluated in the challenge were trained of the same gesture-speech

[1] GENEA stands for "Generation and Evaluation of Non-verbal Behaviour for Embodied Agents". The paper extends a preliminary report, [32] (not peer reviewed), presented at the GENEA Workshop associated with the challenge.

dataset and their motion visualised using the same virtual avatar and rendering pipeline.

3.1 Dataset

We based the challenge on the Trinity Gesture Dataset [11], comprising 244 min of audio and motion-capture recordings of a male actor speaking freely on a variety of topics. This is one of the largest datasets of parallel speech and 3D motion (in joint-angle space) publicly available in the English language. We removed 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), 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 in the transcriptions.

Before releasing the data to challenge participants, it was split into training data (3 h and 40 min) and test data (20 min), with only the training data initially being shared with the participants. Both these data subsets have since been made publicly available in the original dataset repository at trinityspeechgesture.scss.tcd.ie.

3.2 Challenge rules

Each participating team could only submit one system for evaluation. As for timeline, the speech-motion training data was released to participants on July 1, 2020. Test input speech (but not motion output) was released to participants on August 7, with participants requested to submit their generated gesture motion for the test input speech on or before August 15. The joint evaluation took place after the generated gestures were submitted.

Synthetic gesture motion was required to be submitted at 20 frames per second (fps) in a format otherwise identical to that used by the challenge training data. 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. Manual tweaking of the output motion was not allowed, since the idea was to evaluate how systems would perform in an unattended setting.

4 SYSTEMS AND TEAMS

We recruited challenge participants from a public call for participation. Sixteen teams signed up for the challenge, and we distributed the dataset and baseline implementations to all of them. Five teams completed the challenge and the other teams were not able to submit results for evaluation. Two of the withdrawing teams explained it was (in one case) due to reduced manpower for completing the challenge and (in the other) due to unsatisfactory results. There were no reported withdrawals due to the challenge data or task.

The challenge evaluation contained 9 different *conditions* or *systems*: 2 toplines that represent human-quality gesture motions, 2 previously published *baselines*, and 5 challenge *entries/submissions*. Table 1 lists all conditions, together with participating team names and (abbreviated) affiliations. Following the practice established by the Blizzard Challenge, we anonymised the teams in the present paper, by not revealing which team was assigned which ID, but individual teams are free to disclose their ID if they wish. Papers from each team describing their submitted systems in detail are collected in the proceedings of the GENEA Workshop 2020.[2]

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 human-like motion (same as N, so a topline), but whose behaviour is completely unrelated to the speech (and thus can be considered as a bottom line in terms of motion appropriateness for the speech).

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 identify the “best” baseline systems to use. Therefore the choice was more subjective and based on code availability, with the two baseline systems chosen from recent data-driven gesture-generation papers that had their code available and were easy to reproduce. These were:

BA The system from [29], 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, generating motion frame by frame by sliding a window over the speech input.

BT The system from [61], which only takes text transcript information (which includes word timing information) into account when generating system output. This model consists of an encoder for text understanding and a decoder for frame-by-frame pose generation.

The original authors of the baseline systems updated their methods and code to perform well on the challenge material. In BA, the representation of upper-body poses in the challenge dataset was different from the data 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 the exponential map [14] and was smoothed using a Savitzky–Golay filter [51] with window length 9 and polynomial order 3.

In BT, the representation of upper-body poses in the challenge dataset was different to that of the TED dataset used in the original publication. Accordingly, the pose representation was changed from 2D Cartesian coordinates of 8 upper-body joints to 3×3 rotational matrices for each of 15 joints. The data dimension for a pose was 135 (3×3×15). The number of layers and loss function were the same as in the original paper. The hyperparameters of learning rate and loss term weights were adjusted manually. Also, pretrained FastText word vectors [6] were used instead of GloVe [43].

[2] Available at zenodo.org/communities/genea2020.

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.

			Inputs used		Representation or features		Stochastic
Name or description	Origin	ID	Aud.	Text	Input speech	Motion	output?
Natural motion	-	N	✓	✓	–	–	✓
Mismatched motion	-	M	✗	✗	–	–	✓
Audio-only baseline	Kucherenko et al. [29]	BA	✓	✗	MFCC	Exp. map	✗
Text-only baseline	Yoon et al. [61]	BT	✗	✓	FastText†	Rot. matrix	✗
AlltheSmooth [35]	CSTR lab, UEDIN, Scotland	S...	✓	✗	MFCCs	Joint pos.	✗
Edinburgh CVGU [42]	CVGU lab, UEDIN, Scotland	S...	✓	✓	BERT† & mel-spectr.	Rot. matrix	✓
FineMotion [27]	ABBYY lab, MIPT, Russia	S...	✓	✓	GloVe† & mel-spectr.	Exp. map	✗
Nectec [53]	HCCR unit, NECTEC, Thailand	S...	✓	✓	Phoneme, Spacy word vecs.†, MFCCs, & prosody	Exp. map	✗
StyleGestures [1]	TMH division, KTH, Sweden	S...	✓	✗	mel-spectr.	Exp. map	✓

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.

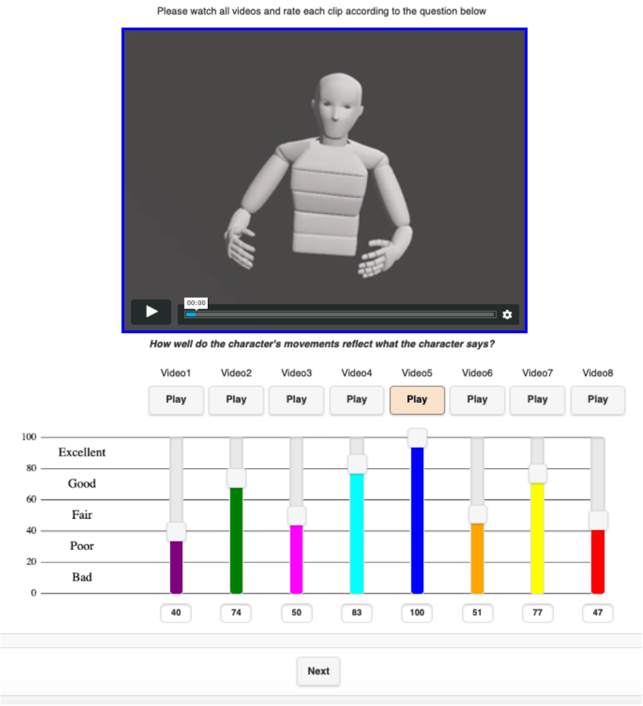
We also considered including a re-implementation of the system from Ginosar et al. [12] as a third baseline, but this was dropped due to unsatisfactory results. This might be due to the challenge dataset being smaller than needed for this method, or due to difficulties with tuning the particular implementation we used.

5 EVALUATION

We conducted a large-scale, crowdsourced, joint evaluation of gesture motion from the nine conditions in Table 1 in parallel using a within-subject design (i.e., every rater was exposed to and evaluated all conditions). The systems were evaluated in terms of the human-likeness of the gesture motion itself, as well as the appropriateness of the gestures for a given input speech. Jonell & Kucherenko et al. [24] recently found 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. Attention checks were used to exclude participants that were not paying attention, as detailed in Section 5.3.

5.1 Stimuli

Prior to motion being submitted, 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, which were not revealed to participants, 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 provided to participants (see Section 5.1), albeit at a higher resolution of 960×540 this time.


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: Screenshot of the rating interface from the evaluation. The question asked in the image ("How well do the character's movements reflect what the character says?") originates from [25], and was changed for each of the two evaluations in this paper.

We used the same virtual avatar for all renderings during the challenge and the evaluation. The avatar can be seen in Figure 1. The avatar originally had 69 joints (full body including fingers) but

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

only 15 joints, corresponding to the upper body and no fingers, were used for the challenge. Since hand and finger data had been omitted, these body parts were assigned a static pose, in which the hands were lightly cupped (again, see Figure 1).

We also developed a visualisation server that enabled all participating teams to produce gesture-motion visualisations identical (except in resolution) to the video stimuli evaluated in the challenge. This was implemented using a Python-based web server which interfaced Blender 2.83. Participants would send a send a 20 fps BVH file to the visualisation server, and these files were then processed as quickly as possible into videos visualising the motion on the avatar, in the order they came in. The same server was also used to render the final stimuli, but with the resolution increased to 960×540 instead of 480×270. (The lower resolution was used during the main part of the challenge to increase performance and throughput of the server, since 16 teams initially took part.) The visualisation server code is provided at github.com/jonepatr/genea_visualizer.

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 (MUltiple Stimuli with Hidden Reference and Anchor) test standard for audio-quality evaluation [19] from the International Telecommunication Union (ITU). However, there are a number of differences between the MUSHRA standard and our evaluation, e.g., our use of video rather than audio and the omission of a designated reference and a low-end anchor, which correspond to the letters R and A in the original acronym.

Figure 1 shows an example of the user interface used for the 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 compare and evaluate motion stimuli from all toplines, baselines, and most submitted systems, all for/with the same speech. It was possible for participants to return to previous conditions and change their rating after seeing other examples. Lastly they were presented with a page asking for demographics and their experience of the test. As can be seen in the figure, 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) [20] tests, another evaluation standard developed by the ITU.

For a detailed explanation of the evaluation interface we refer the reader to [25], which introduced and validated the evaluation paradigm for gesture-motion stimuli.

5.3 Study design

Each study was balanced such that each segment appeared on pages 1 through 10 with approximately equal frequency across all raters (segment order), and each condition was associated with each slider with approximately equal frequency across all pages (condition order). For any given participant and study, each page would use different speech segments. 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 7, 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 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 include speech in stimulus videos and only used text-based attention checks (all videos 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 stimuli, and each participant had to pass one text-based and two audio-based attention checks.

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 GBP 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 meaningful objective evaluations to use. As a step in this direction we consider two numerical measures previously used to evaluate co-speech gestures, namely average jerk and distance between gesture speed (i.e., absolute velocity) histograms.

5.5.1 Average jerk. The third time derivative of the joint positions is called *jerk*. Average jerk is commonly used to quantify motion smoothness [29, 39, 55]. We report average values of absolute jerk (defined using finite differences) across different motion segments. A perfectly natural system should have average jerk very similar to natural motion.

5.5.2 Comparing speed histograms. The distance between speed histograms has also been used to evaluate gesture quality [29, 31], since well-trained models should produce motion with similar properties to that of the actor it was trained on. In particular, it should have a similar motion-speed profile for any given joint. To evaluate this similarity we calculate speed-distribution histograms for all systems and compare them to the speed distribution of natural motion (condition N) by computing the Hellinger distance [40], $H(\boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)}) = \sqrt{1 - \sum_i \sqrt{h_i^{(1)} \cdot h_i^{(2)}}}$, between the histograms $\boldsymbol{h}^{(1)}$ and $\boldsymbol{h}^{(2)}$. Lower distance is better.

For both of the objective evaluations above the motion was first converted from joint angles to 3D coordinates. The code for the numerical evaluations has been made publicly available to enhance reproducibility.[4]

6 RESULTS AND FINDINGS OF THE CHALLENGE EVALUATION

This section describes and discusses the results of the subjective and objective evaluations. First, Section 6.1 introduces demographic and other information gathered from the recruited participants. Section 6.2 then reports the results of the subjective evaluation of challenge conditions, which also are visualised in a number of different figures. Section 6.3 complements the subjective findings with results on the objective measures introduced in Section 5.5. Section 6.4 provides a discussion of the results obtained in the challenge evaluation.

6.1 Data on test participants

Each user study recruited 125 participants that passed all attention checks they encountered. In the human-likeness study, average reported participant age was 31.5 years (standard deviation 10.7), with 66 men, 57 women, and 2 others. 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 the appropriateness study, average age was 31.1 years (standard deviation 11.7), with 60 men, 64 women, and 1 other. 78 participants reported residing in Europe, 1 in Africa, 39 in North America, 3 in Asia, and 4 in Oceania. Each study had 116 native and 9 non-native speakers of English.

23 test-takers in the human-likeness study and 40 test-takers in the appropriateness study did not pass all attention checks. These test-takers were not part of the 125 participants analysed. Scores from sliders used for attention checks were also omitted, leaving in total 8,375 and 9,625 ratings that were analysed in each of the two respective studies. The median successful completion time for the main part of the study was 24 min for the human-likeness study and 27 min for the appropriateness study, with the shortest successful completion time being 12 min in both studies. These figures exclude reading instructions and answering the post-test questionnaire, unlike the timings in Section 5.4.

6.2 Analysis and results of subjective evaluation

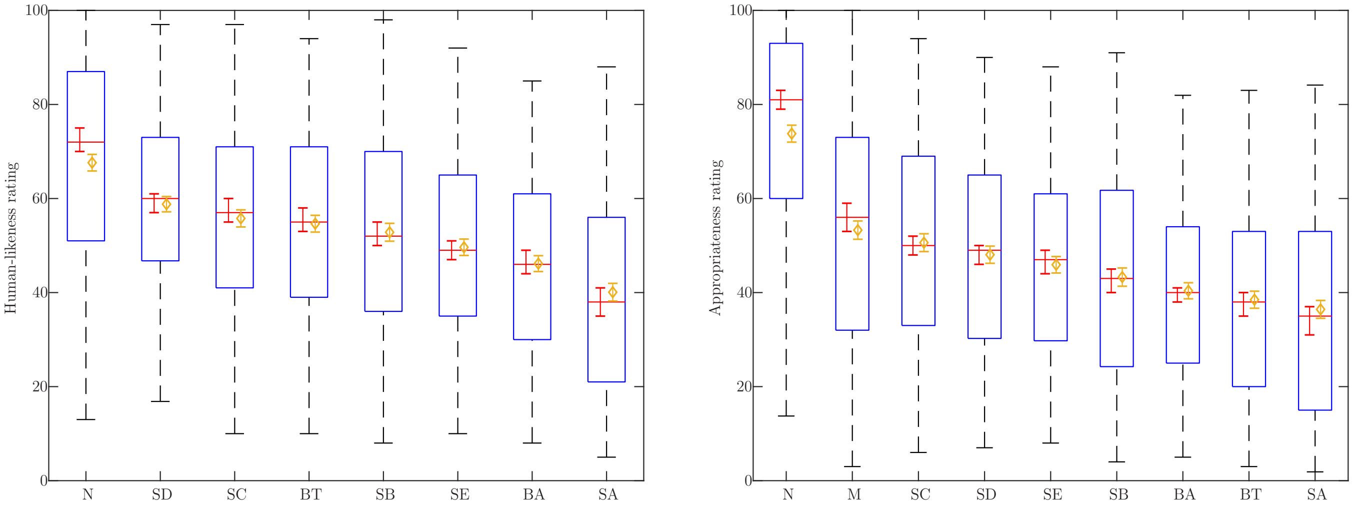
Summary statistics (sample median and sample mean) for all conditions in each of the two studies are shown in Table 2 (see page 8), together with a 99% confidence interval for the true median/mean. The confidence intervals were computed either using a Gaussian assumption for the means (i.e., with Student's t-distribution cdf, and rounded outward to ensure sufficient coverage), or using order statistics for the median (leverages the binomial distribution cdf, cf. [16]).

The ratings distributions in the two studies are further visualised through box plots in Figure 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 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.

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 Wilcoxon 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 likely be inappropriate, as we can see from the box plots in Figure 2 that ratings distributions are skewed and thus 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 [18] (which is uniformly more powerful than regular Bonferroni correction) 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, all at the level $\alpha = 0.01$. Which conditions that were found to be rated significantly above or below which other conditions in the two studies is visualised in Figure 3.

Finally, we present two diagrams that bring the results of the two studies together. Figure 4, in particular, visualises the relative (partial) ordering between different conditions implied by the results of the two studies in Figure 3. Although there are similarities, the two orderings are meaningfully different. This, together with the results in [25], reinforces a conclusion that the two studies managed to disentangle aspects of perceived motion quality (human-likeness) from the perceived link between gesture and speech (appropriateness). Figure 5, meanwhile, visualises confidence regions for the median rating as boxes whose horizontal and vertical extents are

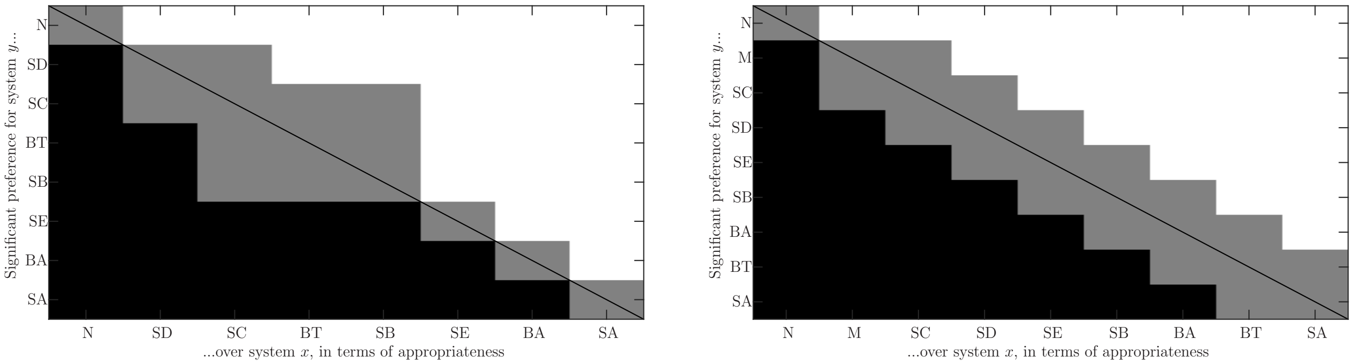
[4] See github.com/Svito-zar/genea_numerical_evaluations.



(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.



(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 Figure 2, which is different for each of the two studies.

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 for human-likeness.

6.3 Results of objective evaluation

Results of the objective evaluations from Section 5.5 are given in Table 3. The first column contains the average jerk across all the joints. We report mean and standard deviation for the full 20 min of test motion. The second and third columns contain the Hellinger distance between speed histograms for the left and right wrists.

Different systems performed best (coming closest to the natural motion N) in different objective measures. For example, systems SA and SB where the closest to the ground truth in terms of the jerk value, but SE and SD were among the closest to the ground truth as measured by Hellinger distance between speed histograms.

We also found that objective metrics deviate from the subjective results. While SA showed the most similar jerk to natural motion, it was less preferred in the subjective evaluation. Similarly, SE showed the Hellinger distances most similar to N, but was not close to being the most preferred synthetic system in the subjective evaluation. Considering this disparity, we stress that objective evaluation of

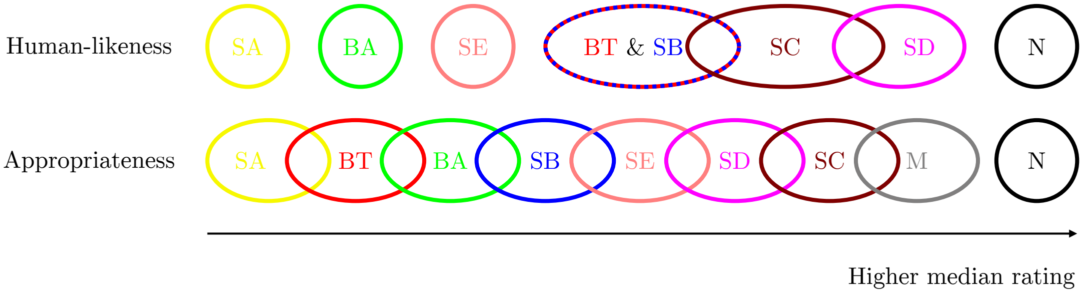


Figure 4: Partial ordering between conditions in the two studies. Each condition is an ellipse; overlapping or (in one case) coinciding ellipses signify that the corresponding conditions were not statistically significantly different in the evaluation. The diagram was inspired by [57] with colours adapted from [7]. There is no scale on the axis since the figure visualises ordinal information only.

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, since it uses the same motion clips as N.

	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

Table 3: Results from the objective evaluations. The Hellinger distance between natural and synthetic speed profiles was computed for the two wrist joints, since hand motion is of central importance for co-speech gestures.

		Hellinger distance	
ID	Jerk	Left	Right
N	151.52 ± 35.57	0	0
BA	65.59 ± 4.42	0.084	0.090
BT	45.84 ± 2.14	0.130	0.096
SA	132.37 ± 27.64	0.064	0.059
SB	189.39 ± 4.66	0.126	0.114
SC	84.44 ± 8.48	0.083	0.088
SD	72.06 ± 7.91	0.073	0.062
SE	97.85 ± 9.34	0.049	0.049

gesture motion is a complementary measure, and that subjective evaluation is much more important.

6.4 Discussion of the challenge results

It is obvious that gesture generation is a difficult problem which is far from being solved, seeing that no system came remotely close to the natural motion N. However, the fact that many submissions scored significantly better than the previously published baselines suggests that progress is being made. The numerical gap between natural motion and that synthesised by machine-learning models is greater in terms of appropriateness than human-likeness. This (along with the fact that no artificial system surpassed the speech-independent condition M) could indicate that appropriateness is a harder problem to solve. As one part of this, the available data may not be sufficiently rich to allow learning to generate appropriate gestures, especially semantically-meaningful gesticulation.

Previous studies suggest that motion quality (human-likeness) may influence gesture appropriateness ratings in subjective evaluations [31, 61]. Our experiments only partly managed to separate these two aspects of gesture perception. On the one hand, we can observe in Figure 4 that different systems were good at different things: some scored better than other on human-likeness, but worse on appropriateness. The human-likeness ratings, which did not include any speech information in the video stimuli, also have little potential to include any aspects of appropriateness. On the other hand, no machine-learning system was rated above mismatched motion M in terms of appropriateness, which contrasts against previous evaluations on other data such as [61]. This could be an effect of that data containing more pauses in speech and gesticulation, thus making a mismatch more apparent. Moreover, the high appropriateness rating reached by one of the audio-only systems may indicate that our evaluation did not capture semantic appropriateness well.

7 DISCUSSION AND IMPLICATIONS OF THE CHALLENGE

In this section we discuss challenge implications: what the challenge brings to the scientific community, the limitations of the challenge, and lessons learned from conducting it.

7.1 Implications of the challenge

We have taken the first step in jointly benchmarking different gesture generation systems on a common dataset and virtual avatar.

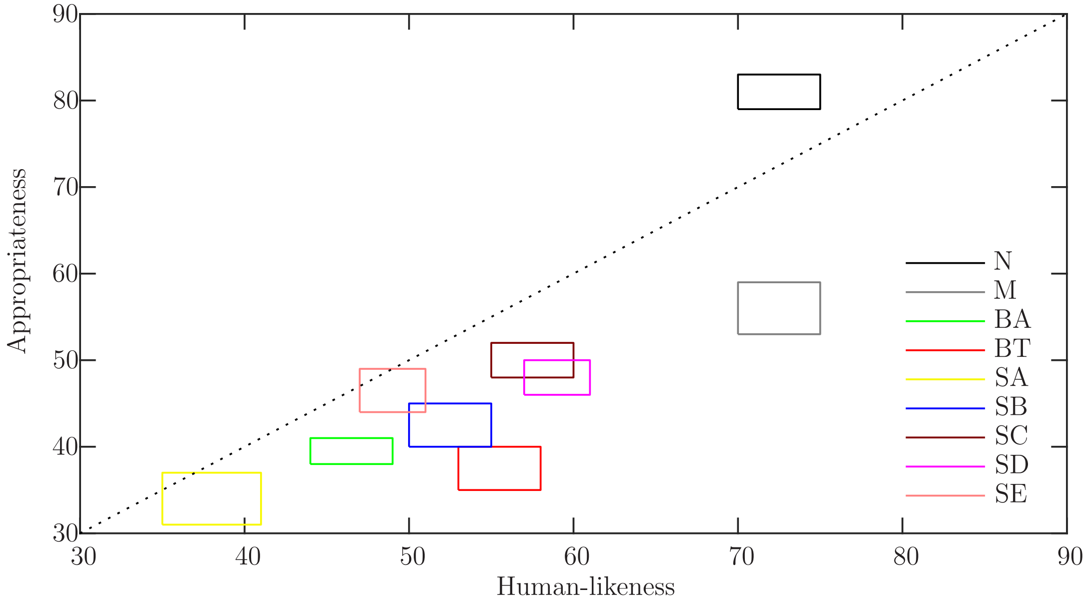


Figure 5: Confidence regions for the true median rating across both studies. The dotted black line is the identity, $x = y$. While the human-likeness (x-coordinate) of M was not evaluated directly, it is expected to be very close to N since it uses the same motion clips, and the horizontal extent of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

The below points summarise some of the added value we see for the gesture-generation field:

(1) We have defined the first benchmark for evaluating gesture-generation models, consisting of a dataset of speech audio, aligned text transcriptions, and 3D motion, as well as train-test splits and an evaluation procedure. Future research can make use of these components to compare new models with previous ones in a consistent way.

(2) All the motion clips generated by the systems evaluated in the challenge are publicly available, together with the rendering pipeline used.[5] This enables easy comparisons with these systems in the future, since their motions can be used directly, without the need to reproduce the systems.

(3) All the subjective and objective scores for the challenge submissions and analysis scripts we used are also available online.[6] This material could be used, e.g., to investigate human perception and to analyse the correlation between subjective perception and different objective measures (not only those in Section 5.5), to aid progress toward reliable and useful objective metrics for the field.

7.2 Limitations

Our crowdsourced evaluation had a few limitations: First, in measuring appropriateness of gestures (i.e., the link between gestures and speech), semantic and rhythmic appropriateness were considered together, and there is no way to determine which aspect of appropriateness the participants rated. In addition, our appropriateness ratings were likely been affected by motion quality to some extent, as discussed in Section 6.4, despite the fact that participants were instructed participants to disregard 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., dyadic conversation versus monologue), which this challenge did not explore. We believe the models and the challenge can be extended to other languages if proper datasets are available, as audio processing is essentially language agnostic and pretrained word vectors are available for a multitude of languages [15].

A third limitation is that we considered only upper-body gestures, even though whole-body gestures (including posture, stepping motion and stance, facial expression, and hand motion) also are important in social interactions. Three teams stated that the most desirable extension of the challenge would be to include whole-body and/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

Conducting the gesture generation challenge has highlighted several take-away messages and lessons learned:

- Being human-like does not mean being appropriate for gestures of a virtual avatar. The challenge evaluation found some systems performed better than others in terms of human-likeness but worse in terms of appropriateness, highlighting that one does not imply the other. Any evaluation or comparison of synthetic gestures should keep this distinction in mind.

[5] See zenodo.org/record/4080919 and github.com/jonepatr/genea_visualizer for the motion stimuli and the visualiser, respectively.
[6] See zenodo.org/record/4088250.

- Providing carefully pre-processed data and good infrastructure (code for feature extraction, motion visualisation, baseline systems, etc.) enables challenge participants to focus on developing their system, instead of solving unrelated issues.
- A MUSHRA-like evaluation scheme can successfully benchmark numerous gesture-generation models in parallel.
- There is a need for future challenges, since there remains a big gap between natural and synthesised motion and variation across speakers, languages, and scenarios has yet to be explored in a challenge format.

We additionally think the following points are worth considering for anyone running a similar challenge in the future:

- Include some of the best systems used in the current challenge to provide continuity and assess whether the field keeps moving forward. This is facilitated by the fact that the baselines and several challenge entries have made their code publicly available.
- Evaluate gesture appropriateness in a more granular and precise way, for example having separate questions and studies for semantic and rhythmic appropriateness, and by also evaluating contrasts between matched and mismatched motion from all challenge entries. Since the link between speech and motion is important yet difficult to evaluate, challenges and their data may be used to explore how to better measure gesture appropriateness.
- Use a different speech-gesture dataset. As previously discussed, the dataset used in this challenge has limitations, e.g., it has already been used extensively and contains just a single actor speaking in isolation, while gesture generation systems usually are intended to be used in an interaction. More data may be necessary to better learn semantically meaningful gestures.

8 CONCLUSIONS

We have hosted the GENEA Challenge 2020 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 measurably, since most submissions performed significantly better than the baselines published the year before. Different systems were also found to be good at different things on the two scales (human-likeness 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), the visualisation code, and the associated library of rated motion clips from the challenge will be useful for future benchmarking and research in gesture generation. 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.

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 André Tiago Abelho Pereira, Bram Willemsen, Dmytro Kalpakchi, Jonas Beskow, Kevin El Haddad, and Ulme Wennberg 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. 2020. The StyleGestures entry to the GENEA Challenge 2020. In *Proc. GENEA Workshop*. https://doi.org/10.5281/zenodo.4088600

[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.

[3] Kirsten Bergmann and Stefan Kopp. 2009. GNetIc – Using Bayesian decision networks for iconic gesture generation. In *Proc. IVA*. 76–89.

[4] Kirsten Bergmann, Stefan Kopp, and Friederike Eyssel. 2010. Individualized gesturing outperforms average gesturing–evaluating gesture production in virtual humans. In *Proc. IVA*. 104–117.

[5] Alan W. Black and Keiichi Tokuda. 2005. The Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets. In *Proc. Interspeech*. 77–80.

[6] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* 5 (2017), 135–146.

[7] Robert M. Boynton. 1989. Eleven colors that are almost never confused. In *Proc. SPIE*, Vol. 1077. 322–332.

[8] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2001. BEAT: The behavior expression animation toolkit. In *Proc. SIGGRAPH*. 477–486.

[9] Marcela Charfuelan and Ingmar Steiner. 2013. Expressive speech synthesis in MARY TTS using audiobook data and EmotionML. In *Proc. Interspeech*. 1564–1568.

[10] Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella. 2015. Predicting co-verbal gestures: A deep and temporal modeling approach. In *Proc. IVA*. 152–166.

[11] Ylva Ferstl and Rachel McDonnell. 2018. Investigating the use of recurrent motion modelling for speech gesture generation. In *Proc. IVA*. 93–98.

[12] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning individual styles of conversational gesture. In *Proc. CVPR*. 3497–3506.

[13] Avashna Govender, Anita E. Wagner, and Simon King. 2019. Using pupil dilation to measure cognitive load when listening to text-to-speech in quiet and in noise. In *Proc. Interspeech*, Vol. 20. 1551–1555.

[14] F. Sebastian Grassia. 1998. Practical parameterization of rotations using the exponential map. *J. Graph. Tools* 3, 3 (1998), 29–48.

[15] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proc. LREC*. 3483–3487.

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

[17] 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.

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

[19] International Telecommunication Union, Radiocommunication Sector. 2015. *Method for the subjective assessment of intermediate quality levels of audio systems.*

Recommendation ITU-R BS.1534-3. https://www.itu.int/rec/R-REC-BS.1534-3-201510-I

[20] 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

[21] Carlos T. Ishi, Daichi Machiyashiki, Ryusuke Mikata, and Hiroshi Ishiguro. 2018. A speech-driven hand gesture generation method and evaluation in android robots. *IEEE Robot. Autom. Lett.* 3, 4 (2018), 3757–3764.

[22] Ryo Ishii, Taichi Katayama, Ryuichiro Higashinaka, and Junji Tomita. 2018. Generating body motions using spoken language in dialogue. In *Proc. IVA*. 87–92.

[23] Patrik Jonell, Taras Kucherenko, Gustav Eje Henter, and Jonas Beskow. 2020. Let's face it: Probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings. In *Proc. IVA*. Article 31, 8 pages.

[24] Patrik Jonell, Taras Kucherenko, Ilaria Torre, and Jonas Beskow. 2020. Can we trust online crowdworkers? Comparing online and offline participants in a preference test of virtual agents. In *Proc. IVA*. Article 30, 8 pages.

[25] Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, and Gustav Eje Henter. 2021. HEMVIP: Human evaluation of multiple videos in parallel. arXiv:2101.11898

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

[27] Vladislav Korzun, Ilya Dimov, and Andrey Zharkov. 2020. The FineMotion entry to the GENEA Challenge 2020. In *Proc. GENEA Workshop*. https://doi.org/10.5281/zenodo.4088609

[28] Taras Kucherenko. 2018. Data driven non-verbal behavior generation for humanoid robots. In *Proc. ICMI Doctoral Consortium*. 520–523.

[29] Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. 2019. Analyzing input and output representations for speech-driven gesture generation. In *Proc. IVA*. 97–104.

[30] Taras Kucherenko, Dai Hasegawa, Naoshi Kaneko, Gustav Eje Henter, and Hedvig Kjellström. 2021. Moving fast and slow: Analysis of representations and post-processing in speech-driven automatic gesture generation. *Int. J. Hum. Comput. Interact.* (2021). https://doi.org/10.1080/10447318.2021.1883883

[31] Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexanderson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proc. ICMI*. 242–250.

[32] Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. 2020. The GENEA Challenge 2020: Benchmarking gesture-generation systems on common data. (2020). https://doi.org/10.5281/zenodo.4094697

[33] Quoc Anh Le and Catherine Pelachaud. 2012. Evaluating an expressive gesture model for a humanoid robot: Experimental results. https://www.researchgate.net/publication/268257868_Evaluating_an_Expressive_Gesture_Model_for_a_Humanoid_Robot_Experimental_Results

[34] Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun. 2010. Gesture controllers. *ACM Trans. Graph.* 29, 4, Article 124 (2010), 11 pages.

[35] JinHong Lu, TianHang Liu, ShuZhuang Xu, and Hiroshi Shimodaira. 2020. Double-DCCCAE: Estimation of sequential body motion using wave-form - AlltheSmooth. In *Proc. GENEA Workshop*. https://doi.org/10.5281/zenodo.4088376

[36] David McNeill. 1992. *Hand and Mind: What Gestures Reveal about Thought.* University of Chicago Press.

[37] Gabriel Mittag and Sebastian Möller. 2020. Deep learning based assessment of synthetic speech naturalness. In *Proc. Interspeech*. 1748–1752.

[38] Sebastian Möller, Florian Hinterleitner, Tiago H. Falk, and Tim Polzehl. 2010. Comparison of approaches for instrumentally predicting the quality of text-to-speech systems. In *Proc. Interspeech*. 1325–1328.

[39] Pietro Morasso. 1981. Spatial control of arm movements. *Exp. Brain Res.* 42, 2 (1981), 223–227.

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

[41] NTIRE Challenge organisers. 2020. NTIRE 2020: Perceptual extreme super-resolution challenge. https://competitions.codalab.org/competitions/22217 Accessed: 2021-01-18.

[42] Kunkun Pang, Taku Komura, Hanbyul Joo, and Takaaki Shiratori. 2020. CGVU: Semantics-guided 3D body gesture synthesis. In *Proc. GENEA Workshop*. https://doi.org/10.5281/zenodo.4090879

[43] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proc. EMNLP*. 1532–1543.

[44] Manuel Sam Ribeiro, Junichi Yamagishi, and Robert A. J. Clark. 2015. A perceptual investigation of wavelet-based decomposition of $f0$ for text-to-speech synthesis. In *Proc. Interspeech*. 1586–1590.

[45] Najmeh Sadoughi and Carlos Busso. 2019. Speech-driven animation with meaningful behaviors. *Speech Commun.* 110 (2019), 90–100.

[46] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2013. To err is human(-like): Effects of robot gesture on perceived anthropomorphism and likability. *Int. J. Soc. Robot.* 5, 3 (2013), 313–323.

[47] Maha Salem, Stefan Kopp, Ipke Wachsmuth, Katharina Rohlfing, and Frank Joublin. 2012. Generation and evaluation of communicative robot gesture. *Int. J. Soc. Robot.* 4, 2 (2012), 201–217.

[48] 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 *Proc. Ro-MAN*. 247–252.

[49] 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.

[50] Ibon Saratxaga, Jon Sanchez, Zhizheng Wu, Inma Hernaez, and Eva Navas. 2016. Synthetic speech detection using phase information. *Speech Commun.* 81 (2016), 30–41.

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

[52] Éva Székely, João P. Cabral, Mohamed Abou-Zleikha, Peter Cahill, and Julie Carson-Berndsen. 2012. Evaluating expressive speech synthesis from audiobooks in conversational phrases. In *Proc. LREC*. 3335–3339.

[53] Ausdang Thangthai, Kwanchiva Thangthai, Arnon Namsanit, Sumonmas Thatphithakkul, and Sittipong Saychum. 2020. The Nectec gesture generation system entry to the GENEA Challenge 2020. In *Proc. GENEA Workshop*. https://doi.org/10.5281/zenodo.4088629

[54] 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.

[55] Yoji Uno, Mitsuo Kawato, and Rika Suzuki. 1989. Formation and control of optimal trajectory in human multijoint arm movement. *Biol. Cybern.* 61, 2 (1989), 89–101.

[56] Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview. *Speech Commun.* 57 (2014), 209–232.

[57] Mirjam Wester, Zhizheng Wu, and Junichi Yamagishi. 2016. Analysis of the Voice Conversion Challenge 2016 evaluation results. In *Proc. Interspeech*. 1637–1641.

[58] Pieter Wolfert, Taras Kucherenko, Hedvig Kjelström, and Tony Belpaeme. 2019. Should beat gestures be learned or designed? A benchmarking user study. In *Proc. ICDL-EPIROB Workshop*. 4 pages. http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-255998

[59] Pieter Wolfert, Nicole Robinson, and Tony Belpaeme. 2021. A review of evaluation practices of gesture generation in embodied conversational agents. arXiv:2101.03769

[60] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Trans. Graph.* 39, 6, Article 222 (2020), 16 pages.

[61] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *Proc. ICRA*. 4303–4309.

[62] Takenori Yoshimura, Gustav Eje Henter, Oliver Watts, Mirjam Wester, Junichi Yamagishi, and Keiichi Tokuda. 2016. A hierarchical predictor of synthetic speech naturalness using neural networks. In *Proc. Interspeech*. 342–346.