A Review of Evaluation Practices of Gesture Generation in Embodied Conversational Agents

Pieter Wolfert, *Graduate Student Member, IEEE*, Nicole Robinson, *Member, IEEE*, and Tony Belpaeme, *Member, IEEE*

***Abstract*—Embodied conversational agents (ECAs) are often designed to produce nonverbal behavior to complement or enhance their verbal communication. One such form of the nonverbal behavior is co-speech gesturing, which involves movements that the agent makes with its arms and hands that are paired with verbal communication. Co-speech gestures for ECAs can be created using different generation methods, divided into rule-based and data-driven processes, with the latter, gaining traction because of the increasing interest from the applied machine learning community. However, reports on gesture generation methods use a variety of evaluation measures, which hinders comparison. To address this, we present a systematic review on co-speech gesture generation methods for iconic, metaphoric, deictic, and beat gestures, including reported evaluation methods. We review 22 studies that have an ECA with a human-like upper body that uses co-speech gesturing in social human-agent interaction. This includes studies that use human participants to evaluate performance. We found most studies use a within-subject design and rely on a form of subjective evaluation, but without a systematic approach. We argue that the field requires more rigorous and uniform tools for co-speech gesture evaluation, and formulate recommendations for empirical evaluation, including standardized phrases and example scenarios to help systematically test generative models across studies. Furthermore, we also propose a checklist that can be used to report relevant information for the evaluation of generative models, as well as to evaluate co-speech gesture use.**

***Index Terms*—Human–computer interface, human–robot interaction, social robotics, virtual interaction.**

I. Introduction

HUMAN communication involves a large nonverbal component, with some suggesting that a large portion of communicative semantics is drawn from nonlinguistic elements of face-to-face interaction [1]. Nonverbal behavior can be broken down into several elements, such as posture, gestures, facial expressions, gaze, proxemics, and haptics (i.e., touch during communicative interactions). All these elements convey different types of meaning, which can complement or alter the semantic component of communication. Even minimal elements can provide a marked contribution to the interaction. For example, eye blinking with head nodding has been found to influence the duration of a response in a Q&A session between human subjects and a robot [2].

A significant component involved in nonverbal communication is the use of gestures—movements of the hands, arms, or body—to emphasize a message, communicate an idea, or express a sentiment [1]. Humans often use gestures in daily life, such as to point at objects in our visual space or to signal the size of an object. Co-speech gestures are gestures that accompany speech. McNeill [3] categorized four kinds of co-speech gestures: 1) iconic gestures; 2) metaphorical gestures; 3) beat gestures; and 4) deictic gestures. Iconic and metaphorical gestures both carry meaning and are used to visually enrich our communication [4]. An iconic gesture can be an up and down movement to indicate, for example, the action of slicing a tomato. Instead, a metaphoric gesture can involve an empty palm of a hand, which is used to symbolize "presenting a problem." In other words, metaphoric gestures have an arbitrary relation to the concept they communicate, and iconic gestures have a form, which is visually related to the concept being communicated. Iconic and metaphoric gestures not only differ in terms of content and presentation, but are also processed differently in the brain [5]. Beat gestures do not carry semantic meaning, and they are often used to emphasize the rhythm of speech. Beat gestures have been shown to both facilitate speech and word recall [6], [7] and are the most frequent type of gesture [3], [8], [9]. Finally, deictic gestures are used to point out elements of interest or to communicate directions. Not only do they enhance spoken communication, they also facilitate learning [10]. The rest of this introduction covers gesture research in ECAs, evaluation methods, review aim, and objectives.

A. Gesture Use in Human–Machine Interaction

As nonverbal behavior plays an important role in human-human interaction, researchers put substantial efforts into the generation of nonverbal behavior for ECAs (Fig. 1). ECAs, such as social robots today, can display a range of nonverbal behaviors, including the ability to make gesture-like movements [13]–[15]. The use of co-speech gestures in communication with humans by ECAs can influence the

Manuscript received February 22, 2021; revised September 24, 2021, December 23, 2021, and January 25, 2022; accepted January 28, 2022. Date of publication February 25, 2022; date of current version May 17, 2022. This work was supported in part by the Flemish Government through AI Research Program and in part by the Flemish Research Foundation under Grant 1S95020N. This article was recommended by Associate Editor Emilia I. Barakova. *(Corresponding author: Pieter Wolfert.)*

Pieter Wolfert and Tony Belpaeme are with the IDLab, Ghent University, 9052 Gent, Belgium (e-mail: pieter.wolfert@ugent.be; tony.belpaeme@ugent.be).

Nicole Robinson is with the Department of Electrical and Computer Systems Engineering, Turner Institute for Brain and Mental Health, Faculty of Engineering, Faculty of Medicine, Nursing and Health Sciences, Monash University, Clayton, Victoria 3800, Australia (e-mail: nicole.robinson@monash.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/THMS.2022.3149173.

Digital Object Identifier 10.1109/THMS.2022.3149173

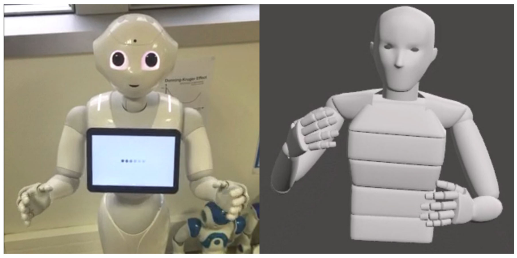

Fig. 1. Pepper robot (left-hand side) [11] and a virtual avatar (right-hand side) [12] using their arms, hands, and torso to complement their speech with co-speech gestures.

perception and understanding of the conveyed message [16], [17]. For example, participants recalled more facts from a narrative told by an ECA, when the ECA made use of deictic and beat gestures compared to when the ECA did not make use of gesticulation [18], [19]. As another example, humans are more willing to cooperate when an ECA showed appropriate gesturing (consisting of deictic, iconic, and metaphoric gestures) in comparison to when an ECA did not use gestures or when the gestures did not match the verbal utterances [20]. Gestures are particularly salient in humanoid robotics, i.e., when the ECA is physically embodied. Robots can be perceived to be more persuasive when they combine gestures with other interactive social behaviors, such as eye gaze, in comparison with when they do not use either of these techniques [21]–[24]. This demonstrates the impact of nonverbal behavior from ECAs can have on people and its importance for consideration in human-agent interactions.

Over the years, artificial intelligence powered systems have been used for the generation of communicative gestures. Gesture generation engines typically rely on matching language and gesture, given that the rhythm and semantic content signaled through gestures are highly correlated with the verbal utterance [3]. Early examples of ECA gesture generation relied on rule-based systems to generate gestures and nonverbal behavior, e.g., [25]. For example, the BEAT system for generating nonverbal behavior can autonomously analyze input text on a linguistic and contextual level, and the system assigns nonverbal behaviors, such as beat and iconic gestures, based on predefined rules [26]. A notable initiative was the behavior markup language (BML), which provided a unified multimodal behavior generation framework [27]. BML was used to describe physical behavior in an XML format and could be coupled with rule-based generation systems. To catch all aspects of nonverbal behavior generation, BML was aimed to not only integrate gesturing, but also other forms, such as body pose, head nodding, and gaze.

Instead of relying on hand-coding, gesture generation systems can also be created from human conversational data, known as the data-driven approach [28], [29]. These data-driven methods have predominantly relied on neural networks for synthesizing gestures. Paired with the rise of deep learning techniques, data-driven methods are capable of unprecedented generalization, an invaluable property when generating high-dimensional temporal output. Data-driven approaches using neural networks are capable of generating more dynamic and unique gestures, but this does heavily depend on the available training data and the type of neural networks that are used. Some approaches learn a mapping from acoustic features of speech signals to gesture [30], [31]. Audio signal-based methods are now much better at creating dynamic and fluent beat gestures, whereas text-based methods show an improved generation of iconic and metaphoric gestures. However, relying on only acoustic features of the speech audio means that semantic details are lost, hence these approaches often only generate beat gestures. Recent work by Kucherenko *et al.* [32] combined neural networks for beat gesture generation with sequential neural networks for generating iconic gestures, dispensing with the need for a rule-based hybrid approach. Yoon *et al.* [33] trained an encoder–decoder neural network on combinations of subtitles and human poses extracted from public TED(x) videos. This allowed the network to learn a relationship between written language, extracted from the video's subtitles, and gesture and was used to generate beat and iconic gestures for a humanoid robot. However, an in-depth evaluation of the different categories of gestures generated by the system was not part of the study. This method was a notable advance in gesture generation, given that videos contain a wealth of human conversational data and are abundantly available. The data used to build data-driven gesture generation can vary, where some use data collected from many individuals [33], others make use of datasets containing a single actor [34].

B. Objective and Subjective Methods for Gesture Evaluation

A central component for any method that can generate human-like behavior is the ability to evaluate the quality of the generated signals. To date, researchers make use of a variety of different methods to evaluate gesture generation systems. One way is to use objective evaluations, often consisting of metrics for the joint speed, joint trajectories, jerk, or the Frechet gesture distance [35]. The objective metrics that are often reported are not necessarily the same metrics that are used to train neural networks. Loss functions only tell how close the generated stimuli are to the ground truth, and they do not provide information on whether the generated motion is dynamic or natural enough. Others include subjective evaluations, which consist of a user study, where human participants evaluate the performance of the gestures used by the ECA. Examples of dimensions on which the performance is evaluated, are the perceived naturalness of the generated motion, the perceived appropriateness of the gestures' timing, "speech-gesture correlation," or "naturalness" [28], [36]. These are often evaluated using several items in one Likert scale. In the human–robot interaction [33], researchers have used questionnaires for general robot evaluation, such as the Godspeed questionnaire, or a selected subselection of items from such instruments. The Godspeed questionnaire can evaluate the perception of ECAs in a nondomain-specific measurement, and quantifies the human likeness, animacy, likability, and perceived intelligence of ECAs [37]. Other methods measure the effect that the gesticulation of an ECA has on the user, such as listener's comprehension and recall of spoken material [18], [19]. In recent work by Ferstl *et al.* [38], study designs and strategies for mitigating the impact of hand tracking loss in virtual reality were compared. In their experiments, they showed the importance of asking the "right" question through comparing several evaluation strategies. However, for the evaluation of generated co-speech gestures in ECAs, a standardized and validated evaluation methodology does not exist.

As objective and subjective measures are central to assessing the quality of the generated communicative behavior, standardized evaluation methods and a uniform way of reporting measures will help to improve the quality of the field.

C. Review Aim and Objectives

Given the importance that gestures can have on human–machine interaction, the ability to effectively identify and evaluate the appropriateness of gestures is vital. However, there is no standardized generation and evaluation protocol available for the field of co-speech gesture generation for ECAs. A standardized questionnaire, measure, or protocol would make comparing work drawn from different sources more effective and would allow for more reliable reporting of results to demonstrate improvement over time. The completion of a comprehensive review and analysis of previous work in the field will support in understanding what has been accomplished so far and help establish a proposed protocol with systematic reporting methods that can be used for more robust evaluation of gesture generation methods and their resulting gestures.

In this article, we present a systematic review that followed the preferred reporting items for systematic reviews and meta-analyses (PRISMA) protocol [39] to identify and assess evaluation methods used in co-speech gestures. We consider this review, timely given that work in co-speech gesture generation, is expanding, new techniques are emerging for creating novel gesture sets, and no systematic evaluation method has been provided to date. Central to this review, we have following three research questions.

1) What methods are used to evaluate co-speech gesture generation?
2) Which methods can be considered the most effective for assessing co-speech gestures?
3) What methods and related metrics should be adapted to create a standardized evaluation or reporting protocol?

These research questions will be used to formulate advice on how to make use of objective and subjective metrics to evaluate the co-speech gesture performance of ECAs, including creating a standardized testing and reporting method.

II. Methods

A. Search Strategy

This review focuses on evaluation studies of co-speech gesture generation methods for embodied conversational agents. Three databases were consulted for data extraction: 1) IEEE Explore; 2) Web of Science; and 3) Google Scholar. IEEE Explore was selected, given that it captures a substantial number of publications in computer science and engineering. Web of Science and Google Scholar were used because they provide access to multiple databases with a wide coverage extending beyond computer science and engineering. Data and record extraction occurred on April 8, 2020 and on June 25, 2020, to collect new records. Two authors conducted independent data extraction steps to reduce the chance of relevant papers being missed from the review, which included inter-rater checks on the included records. The databases were queried using four different keyword combinations, where the search engine would add "AND" between keywords: 1) "gesture generation for social robots;" 2) "co-speech gesture generation;" 3) "nonverbal gesture generation;" and 4) "nonverbal behavior generation."

B. Eligibility: Inclusion and Exclusion

The following inclusion criteria were used.

1) The ECA paper must report on gesture generation on either a robot or an embodied agent.
2) The ECA system must be humanoid in nature, with one or two human-like arms and/or hands that can be used to gesture information or messages to the human.
3) The ECA system must display multiple gestures (i.e., a minimum of two different gestures, one of which must be a beat, iconic, metaphoric, or deictic gesture).
4) Gestures created by the ECA system must be those that would be seen during a multimodal social interaction.
5) The ECA paper must report on a user study (i.e., not evaluated using technical collaborators or authors) in a laboratory, in the wild, or performed remotely through online platforms.
6) The ECA system must be evaluated by a human rater on its performance (either directly or indirectly).

To narrow down our search results, we used the following exclusion criteria.

1) The paper contains a nonhumanoid agent that lacks a typical human-like hand for making a gesture.
2) The paper does not have a clear focus on evaluation of co-speech gestures, i.e., secondary measures, which is less than 50% of the paper.
3) The paper only covers beat gesture generation.
4) The paper is either unpublished, a doctoral dissertation, a review, a technical paper, or a preprint.
5) The paper is not written in English.

Extracted records that only included beat gesture generation were recorded but excluded from the main analysis, as these records rely on audio inputs for the generation of beat gestures. Hence, these beat gesture generation systems do not take semantic information into account. Instead, a separate analysis outside the PRISMA protocol is provided to consider work on beat gestures only, as we do consider the work on beat gesture generation important.

III. Results

In this section, we discuss the results of our literature search. First, we discuss the found articles, followed by a discussion on the usage of different ECAs. Then, we discuss the characteristics of participant samples in experiments, the design of the experiments, and the use of objective and subjective evaluations. At the end, we present the results of our analysis of papers that only incorporated beat gesture generation.

A. Selected Articles

The initial search conducted across three separate databases resulted in 295 papers, which contained 92 duplicate records.

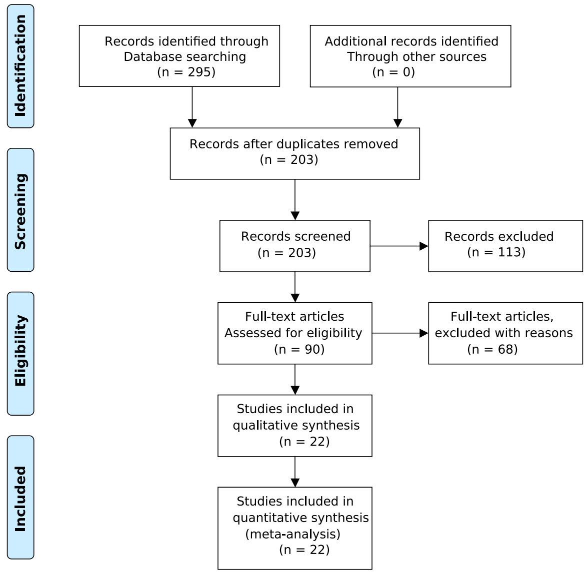

Fig. 2. PRISMA flow chart.

A total of 203 papers were screened for their titles and abstracts for an initial exclusion step, resulting in 113 papers being omitted for not meeting all the inclusion criteria. The 90 remaining papers were assessed in detail by reviewing the main text for eligibility. The 68 noneligible papers met one or more exclusion criteria, and were therefore discarded. This resulted in 22 papers that met all inclusion criteria and none of the exclusion criteria. Fig. 2 shows the PRISMA flow chart with the results of this process. Extracted information from the manuscripts included publication year, venue, design and conditions, method of generation, objective metrics, subjective metrics, type of ECA, evaluation type (online, in the wild, or in a laboratory), participants, characteristics of participants, and other important notes related to the experiment.

B. Embodied Conversational Agents

In the 22 included studies, 16 studies (73%) used different human-like robots, such as NAO (n = 3, 14%), ASIMO (n = 3, 14%), or Wakamaru (an = 2, 9%). Only six (27%) reported the use of a virtual agent (*viz.* [40]–[45]). All the virtual agents were modeled in 3-D as a virtual human, and there were no consistent features across the agents between studies. Of the six studies, four used female avatars [40], [42], [43], [45], one used a male avatar [41], and one study used both [44]. Half of the studies that used avatars, showed only the upper body [41], [43], [45], whereas the other half showed full-body avatars [36], [42], [44]. Specific descriptions of the hands were not provided in all the studies that used avatars. In 19 (87%) studies, the ECA performed iconic gestures, combined with other gestures [18]–[20], [33], [36], [40], [43], [45]–[55].

Metaphoric gestures, with other gestures, are used in 17 (77%) studies [18]–[20], [33], [36], [40]–[43], [45]–[48], [51]–[54]. Deictic gestures, with other gesture types, play a key role in 13 (59%) of the reviewed studies [18]–[20], [33], [36], [40], [43], [45]–[47], [51]–[56]. Last, 17 (77%) studies included iconic, metaphoric, *and* beat gestures [18], [19], [33], [36], [40]–[42], [44], [45], [48]–[53], [55], [56]. Half of the studies had the ECA perform "random gestures" that were included in the evaluation (i.e., gestures that had no alignment between gestures and speech). Other studies (n = 4) had the ECA present, the user with a variety of different nonverbal behavior schemes, such as gestures that were based on text, speech, or a combination of the two [20], [40], [49], [50].

C. Participants

The number of participants per study ranged from 13 to 250 in total (mean = 50, SD = 50, and median = 35). In these papers, 19 (86%) were conducted in the laboratory and three (14%) were conducted either online through Amazon mechanical turk (AMT) (n = 2) and one during an exhibition (i.e., "in the wild"). For the 12 (54%) studies, which did report the mean age of the participants, the mean reported age across all studies was 30.10 years of age (SD = 6.6). The remaining 11 (46%) did not provide demographic data for gender and age. Relating to trial location, 16 (73%) of the studies were performed outside English-speaking countries, with the top three countries being Germany (n = 5), Japan (n = 3), and France (n = 3). For participant recruitment, six (27%) of the studies reported the use of university students—a so-called *convenience sample*—to evaluate gesture generation. Table I provides a more detailed overview of the different studies, countries of origin, and characteristics.

D. Research Experiment and Assessment

In research design, 16 (68%) of the studies used a within-subject design and seven (32%) used a between-subject design. Most (n = 18, 82%) studies invited participants to a university research laboratory to have an interaction with an ECA. Other methods used AMT (n = 2, 9%). With the use in nine (41%) studies, "naturalness" was the most common metric for evaluation in generated gestures. This was followed by synchronization (n = 6, 27%), likability (n = 4, 18%), and human-likeness (n = 2, 9%). Two studies (9%) [42], [47] asked participants to choose which audio track matched best with a given generated gesture sequence. Nine (41%) studies made use of models that learn to generate co-speech gestures. When assessing generated gestures, 16 (73%) studies used questionnaires as a tool to evaluate ECA gesture performance. Only one study [47] included a previous iteration of their gesture model for evaluation. Four studies (18%) used a ground truth as part of the gesture generation evaluation. Three studies (13%) relied on pairwise comparisons, such as two or more videos put side by side with the user selecting the video that best matches with the speech audio, e.g., [44], [50], and [52]. Other evaluation methods involved robot performance, e.g., [18] and [19].

E. Objective and Subjective Evaluation

Table II provides a summary of studies that involved objective evaluation. It also includes the type of agents that were used, as well as the number of speakers in a dataset (when applicable) and the setting of the speakers in the conversation. Only five studies (23%) involved some form of objective evaluation metrics as a

TABLE I
PARTICIPANTS IN STUDIES

Study	Country	Gender	Mean Age (SD)	*N*	Characteristics	Lab/Remote Evaluation
[33]	South Korea	23M/23F	37 (-)	46	45 USA, 1 Australia	AMT
[50]	Spain	-	-	50	Non-native English Speakers	In Lab
[40]	Japan	-	-	10	Age + Gender not specified	In Lab
[36]	Japan	-	-	20	Age + Gender not specified	In Lab
[49]	Japan	-	-	13	-	In Lab
[45]	Slovenia	22M/8F	-	30	-	In Lab
[42]	U.S.A.	-	-	250	One 'worker' per comparison	AMT
[19]	U.S.A.	16M/13F	22.62 (4.35)	29	Convenience Sample	In Lab
[47]	Germany	10M/10F	28.5 (4.53)	20	Native German Speakers	In Lab
[48]	France	14M/7F	21-30	21	Convenience Sample	In Lab
[43]	Slovenia	23M/7F	26.73 (4.88)	30	Convenience Sample	In Lab
[18]	U.S.A.	16M/16F	24.34 (8.64)	32	Convenience Sample	In Lab
[20]	Germany	30M/32F	30.90 (9.82)	62	Convenience Sample	In Lab
[46]	Germany	30M/30F	31 (10.21)	60	Native German Speakers	In Lab
[56]	South Korea	-	-	65		In Lab
[51]	France	36M/27F	37 (12.14)	63	Convenience Sample	In Lab
[53]	France	-	-	63	French Speakers	In Lab
[54]	Germany	20M/20F - 20M/20F	31.31 (10.55)/31.54(10.96)	81	Two Studies	In Lab
[44]	U.S.A.	21M/14F	23 (-)	35	Convenience Sample	In Lab
[52]	U.S.A.	-	-	54	-	In Lab
[41]	U.S.A.	20M/6F	24-26 (-)	26	Non-experts	In Lab
[55]	Germany	-	-	-	-	Exhibition

TABLE II
OBJECTIVE EVALUATION METHODS

Study	Generation Method	Objective Metrics	Agent	#N Speakers	Setting
[33]	Data Driven	Variation on Mean Squared Error	NAO	1295	Single
[50]	Rule Based	-	REEM-C	2	Single
[40]	Data Driven	-	Virtual Agent (3D)	24	Conversation (two)
[36]	Data Driven	-	Android Erica	8	Conversation (three)
[49]	Data Driven	Log-likelihood of generated motion	Pepper	119	Single
[45]	Hybrid	-	Virtual Agent (3D)	5	Multiple
[42]	Rule Based	-	Virtual Agent (3D)	5	Single
[19]	Data Driven	-	Wakamaru	16	Conversation (two)
[47]	Rule Based	Qualitative Analysis of Joint Positions	ASIMO	-	-
[48]	Rule Based	-	NAO	-	-
[43]	Data Driven	-	Virtual Agent (3D)	4	Multiple
[18]	Rule Based	-	Wakamaru	8	Conversation (two)
[20]	Rule Based	-	ASIMO	-	-
[46]	Rule Based	Qualitative Analysis of Joint Positions	ASIMO	-	-
[56]	Rule Based	-	Industrial Service Robot	1	Single
[51]	Rule Based	-	NAO	-	-
[53]	Rule Based	-	NAO	-	-
[54]	Rule Based	-	ASIMO	-	-
[44]	Data Driven	Cost Function on Kinematic Parameters	Virtual Agent (3D)	1	Conversation (two)
[52]	Rule Based	-	ASIMO	4	Single
[41]	Data Driven	-	Virtual Agent (3D)	2	Single
[55]	Rule Based	-	Fritz	-	-

key method in their evaluation. Other metrics included variations on the mean squared error (MSE) ($n = 1$, 4.5%) between the generated and ground truth gestures, and qualitative analyses of joint velocities and positions ($n = 2$, 9%). In total, ten (45%) studies used a data-driven generation method, but only three studies (14%) reported outcomes of their objective metrics used for tuning their models. Only three (14%) studies reported the results of their objective metrics relating to their model performance. Seven studies (32%) relied on data featuring single speakers. In addition to this, seven studies (32%) relied on the data showing two or more speakers. The remainder did not report on the setting of the data or the number of speakers in their dataset.

Table III provides a detailed overview of study design, conditions, and subjective evaluation methods. Fewer studies used between-group design ($n = 6$, 27%) compared to within-group design ($n = 16$, 73%). Most were evaluated using questionnaires ($n = 16$, 73%) followed by pairwise comparisons ($n = 3$, 14%) and other methods ($n = 4$, 18%), such as preference matching (matching audio with video) and recalling facts from a story told by the agent.

F. Additional Results: Beat Gestures

Research work that focused on *only* beat gesture generation was excluded from the main analysis. Methods used to evaluate the performance of beat gesture generation systems in ECAs

TABLE III
SUBJECTIVE EVALUATION METHODS

Study	Design	Conditions	Gesture Types	Evaluation	Questionnaire items
[33]	Within-subject	Ground truth, proposed method, nearest neighbors, random or manual	Iconic, Beat, Deictic, Metaphoric	Questionnaire	Anthropomorphism, Likability, Speech-gesture correlation
[50]	Within-subject	Part-of-Speech-Based, Prosody-Based, Combined	Iconic, Beat	Pairwise + Questionnaire	Timing, Appropriateness, Naturalness
[40]	Within-subject	None, Random, Proposed Method	Iconic, Beat, Deictic, Metaphoric	Questionnaire	Naturalness of Movement, Consistency in utterance and movement, likability, humanness
[36]	Within-subject	No hand motion, Direct Human mapping, Text-based gestures, Text-based + prosody-based gestures	Iconic, Beat, Deictic, Metaphoric	Questionnaire	Human-likeness, Gesture-speech suitability, Gesture-Naturalness, Gesture-Frequency, Gesture-timing
[49]	Within-subject	Ground truth, seq2seq, seq2seq(model) + semantic, seq2seq_tts + semantic	Iconic, Beat	Questionnaire	Naturalness, Skill of presentation, Utilization of gesture, Vividness, Enthusiasm
[45]	Within-subject	Text+Speech (no avatar), Gestures	Iconic, Beat, Deictic, Metaphoric	Questionnaire	Content Match, Synchronization, Fluidity, Dynamics, Density, Understanding, Vividness
[42]	Within-subject	Hands never go into relax position, hands always go into rest position	Beat, Metaphoric	Match preference	N.A.
[19]	Between-subject	Learning-based, unimodal, random, conventional	Iconic, Beat, Deictic, Metaphoric	Questionnaire + Retelling Performance	Immediacy, Naturalness, Effectiveness, Likability, Credibility
[47]	Within-subject	Old version, new version of model	Iconic, Deictic, Metaphoric	Match preference	N.A.
[48]	Within-subject	Introverted versus Extraverted Robot, Adapted Speech and Behavior versus Adapted Speech	Iconic, Beat, Metaphoric	Questionnaire	24 questions on personality, interaction with the robot, speech, and gesture synchronization and matching
[43]	Between-subject	Virtual avatar versus iCub robot	Iconic, Deictic, Metaphoric	Questionnaire	Content Matching, Synchronization, Fluidness, Speech-Gesture Matching, Execution Speed, Amount of Gesticulation
[18]	Between-subject	Number of gestures, randomly selected	Iconic, Beat, Deictic, Metaphoric	Questionnaire + Retelling Performance	Naturalness, Competence, Effective use of Gestures
[20]	Between-subject	Unimodal (speech only), congruent multimodal, incongruent multimodal	Iconic, Deictic, Metaphoric	Questionnaire	Human likeness, Likability, Shared Reality, Future Contact Intentions
[46]	Between-subject	Unimodal versus multimodal (speech + gestures) in a kitchen task	Iconic, Deictic, Metaphoric	Questionnaire	Gesture Quantity, Gesture Speed, Gesture Fluidity, Speech-Gesture Content, Speech-Gesture Timing, Naturalness
[56]	Within-subject	-	Deictic, Beat	Questionnaire	Suitability of Gestures, Synchronization, Scheduling
[51]	Within-subject	Synchronized Gestures, not Synchronized Gestures, Gestures with Expressivity, Gestures without Expressivity	Iconic, Beat, Deictic, Metaphoric	Questionnaire	Synchronization, Naturalness, Expressiveness, Contradictiveness, Gestures are complementary, Gesture-speech Redundancy
[53]	Within-subject	One Condition	Iconic, Beat, Deictic, Metaphoric	Questionnaire	Speech-Gesture Synchronization, Expressiveness, Naturalness
[54]	Between-subject	Study 1: Unimodal versus Multimodal; Study 2: Same	Iconic, Deictic, Metaphoric	Questionnaire	Appearance, Naturalness, Liveliness, Friendliness
[44]	Within-subject	Generated versus Ground Truth	Iconic, Beat	Pairwise	-
[52]	Within-subject	4 studies: Audio vs Wrong Audio; Excited vs Calm Gestures; Low Expressivity, Medium Expressivity, High Expressivity; Slow Gesticulation, Medium Gesticulation, Fast Gesticulation	Iconic, Beat, Deictic, Metaphoric	Pairwise	-
[41]	Within-subject	Speaker 1, speaker 2	Beat, Metaphoric	Match style to speaker	-
[55]	-	-	Iconic, Beat, Deictic	Public Exhibition	-

were similar to those used in work on semantic gesture generation. Ten papers were selected that met the criteria [28], [30], [57]–[64]. A total of seven (70%) studies mentioned the number of participants, with a total of 236 participants. Only four (40%) mentioned statistics on age and gender. Of the ten studies, four (40%) were performed in a lab, and five (50%) online or via AMT. One study was evaluated in an exhibition. As beat gesture generation mostly relied on prosody information, eight (80%) studies used a data-driven approach. Only four of the eight studies that relied on data-driven methods reported their metrics used for an objective evaluation, with either the average position error or the MSE. Seven (70%) of the papers ran their evaluation on a virtual avatar or stick figure with no discernible face. The subjective evaluations performed in these studies were similar to studies that included more gesture categories. Six (60%) used a postexperiment questionnaire to assess the quality of the generated gestures by the ECA. 30% relied on pairwise comparisons and one (10%) relied on the time spent with focused attention on an ECA [59]. All studies ($n = 10$) relied on a within-subject evaluation. The questionnaire items that were used the most: 1) "naturalness" ($n = 4$, 40%); and 2) "time consistency" ($n = 4$, 40%).

IV. Principal Findings and Implications

In this section, we examine the abovementioned observations in more detail and discuss implications for gesture generation methods. Due to the high variation and diversity in the experiments presented in the main analysis, a meta-analysis of the experiments' results will not be provided.

A. Participant Sample

More than half of the studies involved in the main analysis did not report details on the raters, such as the average age, gender, or cultural background. This is a challenge for knowing the generalizability of the findings to larger samples, or its appropriateness for a particular cultural and geographical context. Many studies (30%) used participants that were readily available, for example, from a higher education campus. However, such a *convenience sample* of students is not representative of the general population and may result in a sample of a predominant young adult cohort from higher socioeconomic backgrounds, which might bias the results [65]. Subsequently, the evaluation of gestures generated from models represents a narrower cultural and social viewpoint, and some gestures that are acceptable and natural in other cultures may have been misrepresented or rated poorly in the evaluation process from the use of a more restricted sample.

B. Recruitment and Trial Location

The use of online workers, through services, such as AMT or prolific, does have its merits. Large amounts of data can be collected for a modest budget and in a very short period of time, and it can reach participants from different global regions with very diverse backgrounds. In addition, studies have shown that crowd-sourced data can be of comparable quality to lab-based studies [66]. Given that the majority of users on AMT are U.S.-based, and it is important that studies report the cultural background and country of residence for their participants [67]. Although a recent study showed that there might be no difference between studies for the evaluation of gesture generation in ECAs in the lab and on AMT, it is important to include attention checks and response quality control mechanisms, and to report on these [68].

C. Experimental Set-Up and Assessment

In the main analysis, 14 (65%) studies relied on a within-subject design, which helps to evaluate iterations of gestures over multiple exposures, introduces less variation in the participant scores, and requires fewer participants to achieve sufficient statistical power. It is, however, somewhat problematic that not all studies relied on ground truth comparisons. A ground truth condition typically is a recording of gestures by a human with corresponding speech audio, which are then compared to computer-generated gestures. Human ground truth can serve as a concrete baseline, and this should score the highest on scales for appropriateness and naturalness, providing a clear comparison with other evaluation scores. Several studies also involved random movement generation as a control condition. Random movement is interpreted in different ways, some take random samples from their dataset, which are then put on the top of original speech [33], or insert random parameters for generating gestures [19]. Random gestures are an important control condition for this type of work, ensuring that people are not simply attributing meaning to every gesture seen in the experiment, whether it was a relevant co-speech gesture or not. Overall, we note that the quality of the experimental set-up for gesture generation and evaluation was moderate.

D. Evaluation Methods

The reviewed literature did not show a consistent use of evaluation metrics for gestures, with different research groups focusing on features of interest to them specifically. In most cases, evaluation methods, such as questionnaires, were used for assessing the quality of co-speech gestures in ECAs [33], [40], [49], [53]. Different questionnaires did extract information around similar outcomes, but there was no gold standard for questionnaires, or agreement on a single questionnaire to evaluate the perception of generated gestures. Many items were conflated in a single dimension, which causes an evaluation to miss detail. Questionnaires often involved the use of Likert scales, which sometimes are incorrectly used [69], such as failing to report internal consistency, except for [18] and [19]. Objective evaluations were also highly varied, from using MSE to reporting on histograms with joint velocities and positions.

V. Recommendations for Gesture Evaluation

In the previous section, we discussed the principal findings of our literature review on evaluation used in co-speech gesture generation. Following our findings and our experience, we provide recommendations for researchers working in this field. First, we give more general recommendations, coupled with examples from other, relevant fields. Second, we propose an additional method of evaluation, for which we provide sentences and scenarios. Last, we introduce a checklist that researchers can incorporate in their future work, to improve the level of reporting on datasets, methodology, and results.

A. Participant Sample

As mentioned in the previous section, many studies fail to report on the details of the participant samples. Additionally, not all participant samples reflect the data on which models or systems are trained. We recommend subjective evaluations with participants from diverse populations and backgrounds, reflecting the data on which models or systems are trained.

Some work is more focused on equipping virtual agents with gesticulation, whereas others take it a step further and use their methodology to drive nonverbal behavior in social robots. Often, intermediate evaluation is overlooked, which can potentially lead to unwanted results when these engines are used in an interactive scenario. We recommend that participant evaluation is conducted—when feasible—before putting the model in production or when using the model on a new dataset, ensuring better validity and relevance when deployed for human social interaction.

B. Experimental Setup

In this section, we cover recommendations relating to the conditions, design of studies, and measurements.

The cornerstone of each subjective evaluation is to compare the output of a system to the ground truth. This ground truth condition must contain both motion and audio. Another condition that can shed light on a system's performance is a random or mismatched condition, in which real motion is put on top of a different audio track. An interesting example of this is the subjective evaluation that was part of the GENEA 2020 Challenge, part of the International Conference on Intelligent Virtual Agents, and to our knowledge, the first of its kind in this field [70]. In this challenge, multiple data-driven co-speech generators were compared to two baseline systems. A crowd-sourced subjective evaluation was part of this challenge, for which the results on "appropriateness" and "human-likeness" are displayed in Fig. 3. Here, we see that the ground truth scored higher than the submitted systems on both dimensions and can function as a proper baseline. As for human-likeness, the mismatched condition offers an intriguing result, it does still look as human-like as the ground truth, yet it scored much lower on appropriateness. Both a ground truth condition and a mismatched condition can function as a sanity check when being compared to the output of a system.

Most studies analyzed for this review, ask participants to rate individual stimuli. This can be substantiated with more rigor using the contrastive approach, also known as A/B testing or side-by-side testing [71]. With such an approach, two or more stimuli are presented at the same moment, and a user is asked to either rate both stimuli or to select the preferred stimulus. In a recent study by Wolfert *et al.* [72], these two types of contrastive approaches were tested, as we wanted to find out whether one of the two contrastive approaches should be preferred. In one condition, participants were asked to make a choice between the two videos (pairwise comparison) or to rate both videos. They found that when evaluating many conditions, an approach that makes use of rating scales is to be preferred over using pairwise comparisons. However, pairwise comparisons are a lot faster and less cognitively demanding on participants [73].

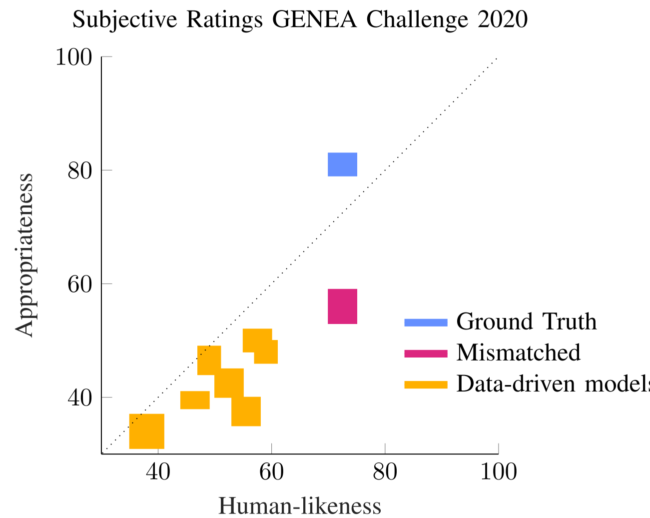

Fig. 3. Human-likeness and appropriateness subjective measurements comparisons between data-driven models and the ground truth from the GENEA 2020 Challenge. Adapted from [70].

Many studies evaluate the performance of their approach in a one-way fashion, videos are put online and participants are asked to evaluate individual videos. However, the need for proper gesticulation in ECAs is often tied to how humans communicate with each other. We recommend (when feasible) evaluating these systems in an interactive scenario, given that it is often the aim of researchers to eventually use ECAs in interactive scenarios. This might require additional engineering, such as creating systems that can also deal with synthetic speech (and thus, with entirely new input), and creating dialogues to be used in an interactive scenario. However, by using an interactive scenario to evaluate an ECAs performance, it becomes possible to record and annotate interactions for indirect measurements, which we will discuss in the next paragraph.

A common way of evaluating stimuli is to ask for ratings on certain dimensions on a 5 or 7 point scale. Table III gives us the richness in terms of questionnaire items used for subjective evaluations. These items can also be seen as "direct" items since they are used for direct measurement on a certain dimension. Frequently used items for this are "naturalness," "human-likeness," "appropriateness," or "likability." Our recommendation here, when one wants to rely on direct measurements only, is that subjective evaluations cover specific dimensions: 1) naturalness; 2) human-likeness; 3) fluency; 4) appropriateness; or 5) intelligibility. Ideally, these dimensions are scored on a 5 or 7 point scale (as these tend to provide more reliable results than larger scales [74]). In addition to direct measurements, we would like to make the case for using a more indirect way of measuring. Examples of indirect measurements are the time it takes to complete a task (task completion), recall rate (recall of facts when letting an ECA tell a story), eye contact and gaze, or response duration (in a question–answering session). For example, task completion is an often-used proxy to estimate

TABLE IV
PREFERRED REPORTING ITEMS FOR CO-SPEECH GESTURE EVALUATION

Embodied Conversational Agent:
☐ ECA: Avatar/robot
☐ DOF (shoulder, elbow, wrist, hand, neck)
☐ Level of articulation of hands
Demographics:
☐ Recruitment method
☐ Sample size
☐ Age
☐ Gender distribution
☐ Geographical distribution
☐ Prior exposure with ECAs
☐ Language(s) spoken
Gesture Generation Model:
☐ Included generated gestures: [iconic, metaphorical, beat, deictic]
☐ Gesture generation model: [rule based, data driven, both, other]
☐ Gesture generation model link/repository
☐ (If not included – why not?)
Gesture Generation Evaluation:
☐ Context / application
☐ Evaluation method/questionnaire set
☐ Gestures annotated by human raters? [Yes/No]
☐ How many human raters were used?
☐ Inter-rater agreement
Metrics:
☐ Objective metrics [average jerk, distance between velocity histograms]
☐ Subjective metrics [human likeness, gesture appropriateness, quality, other]
Training dataset:
☐ Domain of dataset
☐ Length/size of dataset
☐ Gesture types annotated in the dataset
☐ Details on the actors in the dataset (N, language, conversation topic)
Statistical analysis scripts:
☐ Link to scripts

effectiveness in a human–computer interaction [75], and might serve a similar role in our domain. The recall rate has already been used to evaluate gestures [18], [19], but could play a more important role in future interactive evaluations. Eye contact, gaze, or response duration are good proxies to estimate a user's engagement, and taking engagement into account has worked well for other domains [76], [77]. The level of engagement could in turn be a good predictor of how effective an ECA's gesticulation is. However, the drawback of using indirect ways of measuring is that some of these approaches require annotating video recordings of experimental sessions with multiple raters.

C. *Qualitative Analysis of Model Output*

Data-driven models are often trained on a combination of speech–audio and text. Whereas, some systems rely on one speaker (as is the case with systems submitted for the GENEA 2020 Challenge), others rely on multiple speakers. When data-driven systems are capable of generating gestures independent of a specific input voice, it becomes possible to use synthetic text-to-speech as input. This in turn makes it possible to present new data and to qualitatively analyze the performance of models on this new data. We propose a new task that takes entirely new sentences (and text-to-speech output when necessary) as input for gesture generation models. The output then needs to be analyzed for the occurrence of gesture categories. For example, for the sentence "I was throwing a ball," a model might generate an iconic gesture for the word "ball." We have crowdsourced a set of sentences and scenarios that can be used for this task[1]. We propose that researchers take a subset of these as input and that they annotate the model's output for the occurrence of gesture categories. This approach can provide an insight into the richness and diversity of the output of these models. However, this task only works for systems that can work with either only input text or a combination of input text and synthetic speech audio.

D. *Preferred Reporting Items for Gesture Generation Researchers*

To supplement the recommendations made in the previous sections, we offer a nonexhaustive list with preferred reporting items. These draw upon our observations of reporting and our research experiences (see [62], [70], and [72]). Considering the items in the proposed list, researchers could further enhance the quality of their reporting. Our proposed list with items that would be worth including in the future work is summarized in Table IV. It contains items we deem important to report in a scientific publication when working on gesture generation for both physical and nonphysical agents. We hope that the use of this list will make it easier in the future to allow for more systematic evaluation and benchmarking.

VI. CONCLUSION

We reviewed 22 studies on the generation and evaluation of co-speech gestures for ECAs, with a specific focus on evaluation methods. Three questions guided our review, with the first asking what methods are used to evaluate co-speech gesture generation. We found a large diversity of different methods, both objective and subjective, that were applied to the evaluation of generated co-speech gestures. Our main analysis found that many studies did not mention basic statistics on participant characteristics, few studies reported detailed evaluation methods, and there were no systematic reporting methods used for gesture generation and evaluation steps. Our second question asked which methodology is most effective for assessing co-speech gestures. From our review, we cannot conclude that one way of evaluating is to be preferred over another, and recommend making use of both objective and subjective methods. Our third and final question asked what methods and metrics should be adapted to create a standardized evaluation or reporting protocol. Our findings indicate that the field of gesture generation and evaluation would benefit from more experimental rigor and a shared methodology for conducting systematic evaluations, see, e.g., [78] and [79]. We offer questionnaire dimensions, a list with preferred items for designing and reporting studies, and new evaluation tasks, and call on the community to work towards a standardized protocol and questionnaire for the evaluation of systems that produce co-speech gestures. We hope that this work can contribute to further development of the field, and that it will contribute to further advancements in terms of co-speech gesture generation in ECAs.

[1]Online available at https://github.com/pieterwolfert/gesturegeneration-checklist

REFERENCES

[1] M. L. Knapp, J. A. Hall, and T. G. Horgan, *Nonverbal Communication in Human Interaction*. Boston, MA, USA: Cengage Learning, 2013.

[2] P. Hömke, J. Holler, and S. C. Levinson, "Eye blinks are perceived as communicative signals in human face-to-face interaction," *PLoS One*, vol. 13, no. 12, 2018, Art. no. e0208030.

[3] D. McNeill, *Hand and Mind: What Gestures Reveal About Thought*. Chicago, IL, USA: Univ. Chicago Press, 1992.

[4] A. Kendon, "Gesticulation and speech: Two aspects of the process of utterance," in *The Relationship of Verbal and Nonverbal Communication*, Hague, The Netherlands: Mouton Publishers, 1980.

[5] B. Straube, A. Green, B. Bromberger, and T. Kircher, "The differentiation of iconic and metaphoric gestures: Common and unique integration processes," *Hum. Brain Mapping*, vol. 32, no. 4, pp. 520–533, 2011.

[6] C. Lucero, H. Zaharchuk, and D. Casasanto, "Beat gestures facilitate speech production," in *Proc. 36th Annu. Conf. Cogn. Sci. Soc.*, vol. 36, 2014, pp. 898–903.

[7] A. Igualada, N. Esteve-Gibert, and P. Prieto, "Beat gestures improve word recall in 3-to 5-year-old children," *J. Exp. Child Psychol.*, vol. 156, pp. 99–112, 2017.

[8] K. Chui, "Topicality and gesture in chinese conversational discourse," *Lang. Linguistics-Taipei*, vol. 6, no. 4, pp. 635–654, 2005.

[9] A. P.-H. Kong, S.-P. Law, C. C.-Y. Kwan, C. Lai, and V. Lam, "A coding system with independent annotations of gesture forms and functions during verbal communication: Development of a database of speech and gesture (dosage)," *J. Nonverbal Behav.*, vol. 39, no. 1, pp. 93–111, 2015.

[10] K. Lucca and M. P. Wilbourn, "Communicating to learn: Infants' pointing gestures result in optimal learning," *Child Develop.*, vol. 89, no. 3, pp. 941–960, 2018.

[11] A. K. Pandey and R. Gelin, "A mass-produced sociable humanoid robot: Pepper: The first machine of its kind," *IEEE Robot. Automat. Mag.*, vol. 25, no. 3, pp. 40–48, Sep. 2018.

[12] S. Alexanderson, G. E. Henter, T. Kucherenko, and J. Beskow, "Style-controllable speech-driven gesture synthesis using normalising flows," in *Computer Graphics Forum*, vol. 39, no. 2. Hoboken, NJ, USA: Wiley, 2020, pp. 487–496.

[13] C. Bartneck, T. Belpaeme, F. Eyssel, T. Kanda, M. Keijsers, and S. Šabanović, *Human–Robot Interaction: An Introduction*. Cambridge, U.K.: Cambridge Univ. Press, 2020.

[14] C. Breazeal, C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin, "Effects of nonverbal communication on efficiency and robustness in human–robot teamwork," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2005, pp. 708–713.

[15] S. Saunderson and G. Nejat, "How robots influence humans: A survey of nonverbal communication in social human–robot interaction," *Int. J. Social Robot.*, vol. 11, pp. 575–608, 2019.

[16] P. Bremner, A. Pipe, C. Melhuish, M. Fraser, and S. Subramanian, "Conversational gestures in human-robot interaction," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, 2009, pp. 1645–1649.

[17] K. Allmendinger, "Social presence in synchronous virtual learning situations: The role of nonverbal signals displayed by avatars," *Educ. Psychol. Rev.*, vol. 22, no. 1, pp. 41–56, 2010.

[18] C.-M. Huang and B. Mutlu, "Modeling and evaluating narrative gestures for humanlike robots," in *Proc. Robot.: Sci. Syst.*, 2013, pp. 57–64.

[19] C.-M. Huang and B. Mutlu, "Learning-based modeling of multimodal behaviors for humanlike robots," in *Proc. 9th ACM/IEEE Int. Conf. Hum.-Robot Interact.*, 2014, pp. 57–64.

[20] M. Salem, F. Eyssel, K. Rohlfing, S. Kopp, and F. Joublin, "To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability," *Int. J. Social Robot.*, vol. 5, no. 3, pp. 313–323, 2013.

[21] J. Ham, R. H. Cuijpers, and J.-J. Cabibihan, "Combining robotic persuasive strategies: The persuasive power of a storytelling robot that uses gazing and gestures," *Int. J. Social Robot.*, vol. 7, no. 4, pp. 479–487, 2015.

[22] V. Chidambaram, Y.-H. Chiang, and B. Mutlu, "Designing persuasive robots: How robots might persuade people using vocal and nonverbal cues," in *Proc. 7th Annu. ACM/IEEE Int. Conf. Hum.-Robot Interact.*, 2012, pp. 293–300.

[23] A. S. Ghazali, J. Ham, E. Barakova, and P. Markopoulos, "The influence of social cues in persuasive social robots on psychological reactance and compliance," *Comput. Hum. Behav.*, vol. 87, pp. 58–65, 2018.

[24] A. S. Ghazali, J. Ham, E. Barakova, and P. Markopoulos, "Assessing the effect of persuasive robots interactive social cues on users' psychological reactance, liking, trusting beliefs and compliance," *Adv. Robot.*, vol. 33, no. 7-8, pp. 325–337, 2019.

[25] J. Cassell *et al.*, "Animated conversation: Rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents," in *Proc. 21st Annu. Conf. Comput. Graph. Interactive Techn.*, 1994, pp. 413–420.

[26] J. Cassell, H. H. Vilhjálmsson, and T. Bickmore, "Beat: The behavior expression animation toolkit," in *Life-Like Characters*. Berlin, Germany: Springer, 2004, pp. 163–185.

[27] S. Kopp *et al.*, "Towards a common framework for multimodal generation: The behavior markup language," in *Proc. Int. Workshop Intell. Virtual Agents*, 2006, pp. 205–217.

[28] S. Levine, C. Theobalt, and V. Koltun, "Real-time prosody-driven synthesis of body language," in *Proc. ACM SIGGRAPH Asia*, 2009 pp. 1–10.

[29] K. Bergmann and S. Kopp, "Gnetic-using Bayesian decision networks for iconic gesture generation," in *Proc. Int. Workshop Intell. Virtual Agents*, 2009, pp. 76–89.

[30] T. Kucherenko, D. Hasegawa, G. E. Henter, N. Kaneko, and H. Kjellström, "Analyzing input and output representations for speech-driven gesture generation," in *Proc. 19th ACM Int. Conf. Intell. Virtual Agents*, 2019, pp. 97–104.

[31] D. Hasegawa, N. Kaneko, S. Shirakawa, H. Sakuta, and K. Sumi, "Evaluation of speech-to-gesture generation using bi-directional LSTM network," in *Proc. 18th Int. Conf. Intell. Virtual Agents*, 2018, pp. 79–86.

[32] T. Kucherenko *et al.*, "Gesticulator: A framework for semantically-aware speech-driven gesture generation," in *Proc. ACM Int. Conf. Multimodal Interaction*, 2020, pp. 242–250.

[33] Y. Yoon, W.-R. Ko, M. Jang, J. Lee, J. Kim, and G. Lee, "Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots," in *Proc. Int. Conf. Robot. Automat.*, 2019, pp. 4303–4309.

[34] Y. Ferstl, M. Neff, and R. McDonnell, "Multi-objective adversarial gesture generation," in *Proc. Motion, Interact. Games*, 2019, pp. 1–10.

[35] Y. Yoon *et al.*, "Speech gesture generation from the trimodal context of text, audio, and speaker identity," *ACM Trans. Graph.*, vol. 39, no. 6, pp. 1–16, 2020.

[36] C. T. Ishi, D. Machiyashiki, R. Mikata, and H. Ishiguro, "A speech-driven hand gesture generation method and evaluation in android robots," *IEEE Robot. Automat. Lett.*, vol. 3, no. 4, pp. 3757–3764, Oct. 2018.

[37] C. Bartneck, D. Kulic, and E. Croft, "Measuring the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," in *Proc. 3rd ACM/IEEE Int. Conf. Hum. Robot Interact., Annu. ACM/IEEE Int. Conf. Hum.-Robot Interact.*, Tech. Rep. 471, Amsterdam, 2008, pp. 37–44.

[38] Y. Ferstl, R. McDonnell, and M. Neff, "Evaluating study design and strategies for mitigating the impact of hand tracking loss," in *Proc. ACM Symp. Appl. Percep.*2021, pp. 1–12.

[39] D. Moher *et al.*, "Preferred reporting items for systematic reviews and meta-analyses: The prisma statement," *Int J. Surg*, vol. 8, no. 5, pp. 336–341, 2010.

[40] R. Ishii, T. Katayama, R. Higashinaka, and J. Tomita, "Generating body motions using spoken language in dialogue," in *Proc. 18th Int. Conf. Intell. Virtual Agents*, 2018, pp. 87–92.

[41] M. Neff, M. Kipp, I. Albrecht, and H.-P. Seidel, "Gesture modeling and animation based on a probabilistic re-creation of speaker style," *ACM Trans. Graph.*, vol. 27, no. 1, pp. 1–24, 2008.

[42] Y. Xu, C. Pelachaud, and S. Marsella, "Compound gesture generation: A model based on ideational units," in *Proc. Int. Conf. Intell. Virtual Agents*, 2014, pp. 477–491.

[43] I. Mlakar, Z. Kačič, and M. Rojc, "TTS-driven synthetic behaviour-generation model for artificial bodies," *Int. J. Adv. Robot. Syst.*, vol. 10, no. 10, 2013, doi: 10.5772/56870.

[44] S. Levine, P. Krähenbühl, S. Thrun, and V. Koltun, "Gesture controllers," in *Proc. ACM SIGGRAPH Papers*, 2010, pp. 1–11.

[45] M. Rojc, I. Mlakar, and Z. Kačič, "The TTS-driven affective embodied conversational agent EVA, based on a novel conversational-behavior generation algorithm," *Eng. Appl. Artif. Intell.*, vol. 57, pp. 80–104, 2017.

[46] M. Salem, S. Kopp, I. Wachsmuth, K. Rohlfing, and F. Joublin, "Generation and evaluation of communicative robot gesture," *Int. J. Social Robot.*, vol. 4, no. 2, pp. 201–217, 2012.

[47] M. Salem, S. Kopp, and F. Joublin, "Closing the loop: Towards tightly synchronized robot gesture and speech," in *Proc. Int. Conf. Social Robot.*, 2013, pp. 381–391.

[48] A. Aly and A. Tapus, "A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human–robot interaction," in *Proc. 8th ACM/IEEE Int. Conf. Hum.-Robot Interact.*, 2013, pp. 325–332.

[49] A. Shimazu *et al.*, "Generation of gestures during presentation for humanoid robots," in *Proc. 27th IEEE Int. Symp. Robot Hum. Interactive Commun.*, 2018, pp. 961–968.

[50] L. Pérez-Mayos, M. Farrús, and J. Adell, "Part-of-speech and prosody-based approaches for robot speech and gesture synchronization," *J. Intell. Robot. Syst.*, vol. 99, no. 2, pp. 277–287, 2020.

[51] Q. A. Le and C. Pelachaud, "Evaluating an expressive gesture model for a humanoid robot: Experimental results," in *Proc. 8th ACM/IEEE Int. Conf. Hum.-Robot Interact.*, 2012, pp. 134–140.

[52] V. Ng-Thow-Hing, P. Luo, and S. Okita, "Synchronized gesture and speech production for humanoid robots," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2010, pp. 4617–4624.

[53] Q. Le, J. Huang, and C. Pelachaud, "A Common gesture and speech production for humanoid robots," in *Proc. ICMI Workshop Speech Gesture Prod. Virtually Physically Embodied Conversational Agents*, Santa Monica, CA, USA, Oct. 2012.

[54] M. Salem, K. Rohlfing, S. Kopp, and F. Joublin, "A friendly gesture: Investigating the effect of multimodal robot behavior in human–robot interaction," in *Proc. RO-MAN*, 2011, pp. 247–252.

[55] M. Bennewitz, F. Faber, D. Joho, and S. Behnke, "Fritz-A humanoid communication robot," in *Proc. RO-MAN 16th IEEE Int. Symp. Robot Hum. Interactive Commun.*, 2007, pp. 1072–1077.

[56] H.-H. Kim, Y.-S. Ha, Z. Bien, and K.-H. Park, "Gesture encoding and reproduction for human–robot interaction in text-to-gesture systems," *Ind. Robot: Int. J.*, vol. 39, no. 6, pp. 551–563, 2012.

[57] J. Ondras, O. Celiktutan, P. Bremner, and H. Gunes, "Audio-driven robot upper-body motion synthesis," *IEEE Trans. Cybern.*, vol. 51, no. 11, pp. 5445–5454, Nov. 2021.

[58] J. Kim, W. H. Kim, W. H. Lee, J.-H. Seo, M. J. Chung, and D.-S. Kwon, "Automated robot speech gesture generation system based on dialog sentence punctuation mark extraction," in *Proc. IEEE/SICE Int. Symp. Syst. Integrat.*, 2012, pp. 645–647.

[59] P. Bremner, A. G. Pipe, M. Fraser, S. Subramanian, and C. Melhuish, "Beat gesture generation rules for human–robot interaction," in *Proc. RO-MAN 18th IEEE Int. Symp. Robot Hum. Interactive Commun.*, 2009, pp. 1029–1034.

[60] C.-C. Chiu and S. Marsella, "Gesture generation with low-dimensional embeddings," in *Proc. Int. Conf. Auton. Agents Multi-Agent Syst.*, 2014, pp. 781–788.

[61] A. Fernández-Baena, R. Montaño, M. Antonijoan, A. Roversi, D. Miralles, and F. Alías, "Gesture synthesis adapted to speech emphasis," *Speech Commun.*, vol. 57, pp. 331–350, 2014.

[62] P. Wolfert, T. Kucherenko, H. Kjelström, and T. Belpaeme, "Should beat gestures be learned or designed? A benchmarking user study," in *Proc. ICDL-EPIROB Workshop Naturalistic Non-Verbal Affect. Hum.-Robot Interact.*, 2019, pp. 1–4.

[63] K. Takeuchi, D. Hasegawa, S. Shirakawa, N. Kaneko, H. Sakuta, and K. Sumi, "Speech-to-gesture generation: A challenge in deep learning approach with bi-directional LSTM," in *Proc. 5th Int. Conf. Hum. Agent Interact.*, 2017, pp. 365–369.

[64] M. Kipp, M. Neff, K. H. Kipp, and I. Albrecht, "Towards natural gesture synthesis: Evaluating gesture units in a data-driven approach to gesture synthesis," in *Proc. Int. Workshop Intell. Virtual Agents*, 2007, pp. 15–28.

[65] R. A. Peterson and D. R. Merunka, "Convenience samples of college students and research reproducibility," *J. Bus. Res.*, vol. 67, no. 5, pp. 1035–1041, 2014.

[66] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data?" *Perspectives Psychol. Sci.*, vol. 6, no. 1, pp. 3–5, 2011.

[67] A. J. Moss, C. Rosenzweig, J. Robinson, and L. Litman, "Is it ethical to use mechanical turk for behavioral research? Relevant data from a representative survey of MTurk participants and wages," 2020.

[68] P. Jonell, T. Kucherenko, I. Torre, and J. Beskow, "Can we trust online crowdworkers? comparing online and offline participants in a preference test of virtual agents," in *Proc. 20th ACM Int. Conf. Intell. Virtual Agents*, 2020, pp. 1–8.

[69] M. L. Schrum, M. Johnson, M. Ghuy, and M. C. Gombolay, "Four years in review: Statistical practices of Likert scales in human–robot interaction studies," in *Proc. Companion ACM/IEEE Int. Conf. Hum.-Robot Interact.*, 2020, pp. 43–52.

[70] T. Kucherenko, P. Jonell, Y. Yoon, P. Wolfert, and G. E. Henter, "A large, crowdsourced evaluation of gesture generation systems on common data: The Genea challenge 2020," in *Proc. 26th Int. Conf. Intell. User Interfaces*, 2021, pp. 11–21.

[71] R. Kohavi and R. Longbotham, "Online controlled experiments and a/b testing," *Encyclopedia Mach. Learn. Data Mining*, vol. 7, no. 8, pp. 922–929, 2017.

[72] P. Wolfert, J. M. Girard, T. Kucherenko, and T. Belpaeme, "To rate or not to rate: Investigating evaluation methods for generated co-speech gestures," *Proc. ACM Int. Conf. Multimodal Interact.*, 2021, pp. 494–502.

[73] B. Weijters, E. Cabooter, and N. Schillewaert, "The effect of rating scale format on response styles: The number of response categories and response category labels," *Int. J. Res. Marketing*, vol. 27, no. 3, pp. 236–247, 2010.

[74] J. Dawes, "Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales," *Int. J. Market Res.*, vol. 50, no. 1, pp. 61–104, 2008.

[75] P. W. Jordan, *An Introduction to Usability*. Boca Raton, FL, USA: CRC Press, 2020.

[76] Y. I. Nakano and R. Ishii, "Estimating user's engagement from eye-gaze behaviors in human-agent conversations," in *Proc. 15th Int. Conf. Intell. User Interfaces*, 2010, pp. 139–148.

[77] S. Lemaignan, F. Garcia, A. Jacq, and P. Dillenbourg, "From real-time attention assessment to "with-me-ness" in human–robot interaction," in *Proc. 11th ACM/IEEE Int. Conf. Hum.-Robot Interact.*, 2016, pp. 157–164.

[78] G. Hoffman and X. Zhao, "A primer for conducting experiments in human-robot interaction," *ACM Trans. Hum.-Robot Interact.*, vol. 10, no. 1, pp. 1–31, 2020.

[79] Y. Liu, G. Mohammadi, Y. Song, and W. Johal, "Speech-based gesture generation for robots and embodied agents: A scoping review," in *Proc. 9th Int. Conf. Hum.-Agent Interact.*, 2021, pp. 31–38.

Pieter Wolfert (Graduate Student Member, IEEE) received the M.Sc. degree in artificial intelligence from Radboud University, Nijmegen, The Netherlands, in 2018. He is currently working toward the Ph.D. degree with IDLab, Interuniversity Microelectronics Centre, Ghent University, Ghent, Belgium.

His research interests include nonverbal behavior generation and evaluation for embodied conversational agents such as social robots and virtual agents.

Mr. Wolfert co-organized the GENEA Challenge 2020 at the International Conference on Virtual Agents (IVA 2020) and the GENEA Workshop 2021 at the International Conference on Multimodal Interaction (ICMI'21).

Nicole Robinson (Member, IEEE) received the Ph.D. degree from Queensland University of Technology, Brisbane, QLD, Australia, in 2018.

She is currently a Research Fellow with Monash University, Melbourne, VIC, Australia, with a cross appointment between the Faculty of Engineering and the Faculty of Medicine, Nursing and Health Sciences, and an Adjunct Lecturer with the School of Electrical Engineering and Robotics, Queensland University of Technology. She is also affiliated with the Department of Electrical and Computer Systems Engineering, Turner Institute for Brain and Mental Health, Melbourne, VIC, Australia, and the School of Psychological Sciences, Melbourne, VIC, Australia. She uses an interdisciplinary research approach across behavioral science, human-computer interaction, and robotics to evaluate and improve the quality and success of human–robot interaction.

Tony Belpaeme (Member, IEEE) received the Ph.D. degree in computer science from the Vrije Universiteit Brussel, Ixelles, Belgium, in 2002.

He is currently a Professor and a Member of IDLab, Interuniversity Microelectronics Centre, Ghent University, Ghent, Belgium, a Visiting Professor of Robotics and Cognitive Systems with the University of Plymouth, Plymouth, U.K., and also leading a team studying cognitive robotics and human–robot interaction. Starting from the premise that intelligence is rooted in social interaction, he and his research team try to further the science and technology behind artificial intelligence and social human–robot interaction. This results in a spectrum of results, from theoretical insights to practical applications.