HEMVIP: Human Evaluation of Multiple Videos in Parallel

Patrik Jonell
pjjonell@kth.se
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

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

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

ABSTRACT

In many research areas, for example motion and gesture generation, objective measures alone do not provide an accurate impression of key stimulus traits such as perceived quality or appropriateness. The gold standard is instead to evaluate these aspects through user studies, especially subjective evaluations of video stimuli. Common evaluation paradigms either present individual stimuli to be scored on Likert-type scales, or ask users to compare and rate videos in a pairwise fashion. However, the time and resources required for such evaluations scale poorly as the number of conditions to be compared increases. Building on standards used for evaluating the quality of multimedia codecs, this paper instead introduces a framework for granular rating of multiple comparable videos in parallel. This methodology essentially analyses all condition pairs at once. Our contributions are 1) a proposed framework, called HEMVIP, for parallel and granular evaluation of multiple video stimuli and 2) a validation study confirming that results obtained using the tool are in close agreement with results of prior studies using conventional multiple pairwise comparisons.

CCS CONCEPTS

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

KEYWORDS

evaluation paradigms, video evaluation, conversational agents, gesture generation

1 INTRODUCTION

Owing to the difficulties of objectively quantifying human perception and preference, user studies have become the canonical way to evaluate stimuli in many fields. This includes, for example, aspects of human-computer interaction such as video stimuli of synthetic gesture motion for avatars and social robots [1, 10, 12, 21]. There exist several methods for performing such evaluations, particularly Likert scales [15] and pairwise preference tests. These approaches do however not scale well when comparing many different conditions, for example when performing ablation studies or benchmarking a new system against other available approaches.

This paper proposes a novel method for evaluating comparable video stimuli from multiple conditions. Our method is inspired by an evaluation standard called MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) [7], which is widely used for identifying subtle differences in audio quality. We present HEMVIP (Human Evaluation of Multiple Videos in Parallel), which adapts MUSHRA to the evaluation of video stimuli, and validate our proposal by comparing results obtained from HEMVIP against a previous evaluation of videos of generated non-verbal behavior motion for a virtual agent [13]. Our code is available at: https://github.com/jonepatr/hemvip/ and the data and analysis code for the validation study we describe can be found at: https://doi.org/10.5281/zenodo.5196581.

2 RELATED WORK

Subjective evaluations of video stimuli are often carried out using Mean Opinion Scores (MOS) [9], Likert-type scales [15, 18], or relative preference tests. In this paper, we use the problem of evaluating nonverbal-behavior-generation systems for embodied conversational agents (ECAs) as a running example, where both MOS (cf. [6, 21]) and pairwise preference tests (cf. [3, 13]) are commonplace; see [20] for a comprehensive review. In MOS tests, participants rate individual stimuli (in our case, videos) on a discrete scale, e.g., 1 through 5. A proper Likert-scale evaluation requires many such judgments [18]. Since videos are rated in isolation, raters may struggle to notice minor differences between stimuli and to apply a consistent standard to stimuli presented at different points during the test. Relative preference tests are easier than MOS for participants as selection is easier than scoring [4], and such tests are usually better at identifying subtle differences. However, relative preference tests introduce additional design choices, e.g., how

many pairs to be compared and in which combinations. The binary nature of many preference tests means that responses are relatively information-poor and makes it harder to verify that two conditions are statistically different. Moreover, neither of these two evaluation schemes scales well with the number of systems to be evaluated.

The parallel field of audio-quality evaluation has long used both MOS and pairwise preference tests [19], but also a more recent standard for comparing multiple audio systems called MUSHRA [7]. While originally proposed for comparing audio coding systems, MUSHRA-like tests have also been shown to be more efficient than comparable MOS tests for evaluating speech synthesizers [16]. In a MUSHRA test, multiple comparable stimuli (e.g., stimuli generated from the same input text or audio) are presented to the listener on the same page, and listeners rate each stimulus individually in the context of the other stimuli. The main benefit of MUSHRA over the MOS setup is that related stimuli are assessed together, which makes differences easier to spot. Unlike pairwise tests, many systems can be compared at once. In contrast to both MOS and preference tests, a high-resolution scale is used, which can resolve smaller differences and use more sensitive statistical analyses. Recent software for conducting MUSHRA tests online [17] makes it easy to evaluate synthetic audio using crowdsourcing platforms.

Our proposal also has elements in common with the MUSHRA-derived ITU SAMVIQ (Subjective Assessment Methodology for Video Quality) standard [8], which, however, has been withdrawn in early 2020. SAMVIQ was proposed for evaluation of video codec quality and functions similarly to MUSHRA, using hidden references and anchors. Braude also used a MUSHRA-derived setup for videos, specifically to evaluate head-motion synthesis, however, only one video was shown at a time together with a reference [2]. Our proposal instead adapts the MUSHRA approach to the parallel evaluation of video stimuli, to meet the needs of user studies and subjective evaluations in, e.g., human-computer interaction.

3 HEMVIP

The HEMVIP framework is an extension of the MUSHRA standard [7], but instead of audio recordings we assess multiple comparable videos together using the same setup of parallel rating sliders. The core aspects of MUSHRA that HEMVIP inherits are 1) joint, parallel scoring of comparable stimuli, e.g., ones intended for the same context or corresponding to the same system input, which makes HEMVIP scale well for multiple comparisons; 2) blind judgments, in that the systems being compared are unlabeled and the order typically is random; and 3) the use of highly granular ratings entered via a slider GUI (0–100 being the default). There are also some differences from MUSHRA that go beyond the fact that HEMVIP uses video instead of audio: Unlike audio-codec evaluations, where signal bandwidth can be reduced to degrade quality, there is no canonical way to define objectively poor and degraded stimuli to use as low-end anchor stimuli in, e.g., gesture motion. HEMVIP thus does not mandate a specific low-end anchor (one can easily be included if it exists), nor does it provide an explicit reference example by default. Absent this reference, there is no requirement to rate the perceived best stimulus on each page a perfect 100, since there is no assumption that any of the stimuli is perfect.

Fig. 1 shows an example of the HEMVIP evaluation interface (on the left). Each "Play" button corresponds to a video stimulus, which

is played upon clicking the button. Clicking any other play button will immediately start playing its corresponding video instead. Below the play buttons are sliders for rating the different stimuli in response to the question shown below the video. In its current incarnation, HEMVIP is best suited to evaluate one question per page. Text on the left-hand side, here the same labels as used by MOS tests [9], is used to anchor different intervals on the scale.

We implemented the HEMVIP framework based on webMUSHRA [17], but modified to support video material, and including a fully configured web server for the task. Configuration files, in JSON, are used to set up an experiment. These files for instance define the pages (including instructions and a post-test survey) and questions, and which stimuli that are shown. To allow counterbalancing there is one file per participant, that specifies the order of pages and the stimuli on each page. To weed out inattentive or non-serious test-takers, we also implemented a mechanism for attention checks where participants were required to input a specific value for a certain stimulus (within a tolerance of plus/minus 3 and accepting acoustically ambiguous numbers such as 13 vs. 30). An instruction to set this value was inserted into the attention-check stimulus, either through a text overlay or using a synthesized voice, which otherwise appeared similar to other videos.

During the test, every click in the interface is recorded in an interaction log together with timestamps and what element was clicked. This could potentially be used for interaction analysis or detecting cheating participants. The time a participant spends watching each video is also recorded. To make it easier to remember which slider goes with which video, we apply individual colors to each slider and a colored border around the video, matching the color of the slider of the currently playing stimulus (see Fig. 1). These colors were assigned randomly for each page, and test participants were informed that assignments were completely random.

In addition to extending webMUSHRA with additional JavaScript templates and functionality, our HEMVIP implementation incorporates tools to create configuration files and managing participants and a Python-based web server using FastAPI to provide the configuration files to the frontend, block users that fail attention checks, and save result data in an external MongoDB database, see Fig. 1.

4 VALIDATION STUDY

We validated the results obtained using HEMVIP by reproducing a previously conducted evaluation [13], where different speech gesture generation models were compared. This previous evaluation contained six individual ablation studies and a study comparing the best ablation to the ground truth. Our validation study compared those six ablations (named NoAR, NoPCA, NoFiLM, NoAudio, NoText, and NoVel), the "Full" model, and the ground truth side by side (see http://doi.org/10.6084/m9.figshare.13055609 for video examples and [13] for the details of the ablated and "Full" model).

Participants were asked to rate video stimuli produced by various gesture generation models. The stimuli from all conditions were presented together in parallel, and the participants were asked to rate them individually. The study was balanced such that each stimulus appeared on each page with approximately equal frequency (stimulus order), and each condition was associated with each slider with approximately equal frequency (condition order). For any given participant and study, each page would use different speech segments.
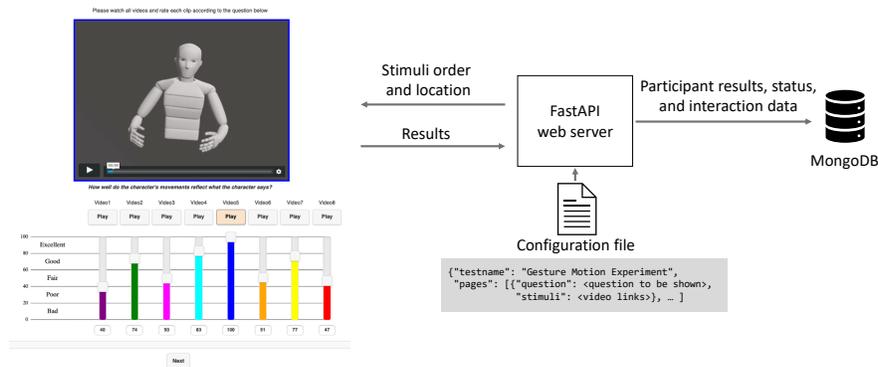
Figure 1: On the left: A screenshot of a page with stimuli from the HEMVIP evaluation interface. On the right: System overview showing how the web server interacts with the evaluation interface.

Every page would contain the "Full" condition, the ground truth condition, and additionally five or six of the ablated conditions, depending on whether an attention check was employed or not. Three attention checks were incorporated into the pages for each study participant as described in Section 3, using randomly selected numbers between 5 and 95. Which sliders on which pages that were used for attention checks was uniformly random, except that no page had more than one attention check, and condition "Full" and the ground truth never were replaced by attention checks.

Four separate studies were conducted, one for each question. Since the original questions were meant for pairwise comparisons, they were slightly modified to fit the parallel question context, but are still in close agreement with the questions asked in [13]:

Q1 "In which video are the character's movements most human-like?" became "How human-like are the character's movements?"
Q2 "In which video do the character's movements most reflect what the character says?" became "How well do the character's movements reflect what the character says?"
Q3 "In which video do the character's movements most help to understand what the character says?" became "How well do the character's movements help to understand what the character says?"
Q4 "In which video are the character's voice and movement more in sync?" became "How well are the character's voice and movements in sync?"

After completing the 10 stimuli pages, the participants answered a questionnaire asking demographic questions (age, gender, which continent they had lived on the most, whether English was their native language and how they perceived the task difficulty) together with more qualitative questions not covered in this paper.

The video stimuli were the same as in [13][1]. The clips were 10 s long and comprised generated motion (or ground truth motion capture) together with matching speech from a single actor originally recorded by [5]. There were 50 speech segments in total, each associated with one motion from each of the eight conditions.

We used the crowdsourcing platform Prolific to recruit 46 participants per study (totaling 184 participants). This resulted in approximately as many ratings as in the data from [13][2], i.e., 4 studies, 46 participants, 10 pages, 8 ratings per page, minus 552 attention checks, for a total of 14,168 ratings, to be compared against 143 par-

ticipants, 26 pages, 4 ratings per page, minus 858 attention checks, totaling 14,014 ratings. Demographics for each study are shown in Table 2, in the supplementary material. The participants were paid 6 GBP for completing the study.

Binary preference judgments as in [13] can be analyzed using, e.g., Clopper-Pearson (C-P) tests. The granular nature of HEMVIP responses, however, open up additional analysis options, and HEMVIP ratings can also be analyzed using pairwise t-tests (for differences in true mean rating) and pairwise Wilcoxon signed-rank tests (differences in true median), as used for MUSHRA tests in audio.

Detailed results of Clopper-Pearson tests (HEMVIP and [13]) and t-tests (HEMVIP only) are reported in Table 3, in the supplementary material, with an overview provided in Table 1. Pairwise Wilcoxon signed-rank tests found exactly the same contrasts to be significant as our t-tests did, except the Full→NoVel comparison in Q4.

5 DISCUSSION

If we ignore the magnitude of pairwise differences in HEVMIP responses for the same page, and consider only their sign, we can use C-P tests to analyze preference ignoring ties seen in the data from [13] and in our ratings. A comparison shows a high, but not perfect, correspondence between the two evaluations (23 out of 28 contrasts agree on whether or not a difference is significant). The 5 differing contrasts do however show the same direction of preference. Among analysis methods that also leverage the granular nature of HEMVIP ratings (and which cannot be applied to preference tests like [13]), both pairwise t-tests and Wilcoxon signed-rank tests produce results that match the original study for all contrasts, except one for the t-tests and two for the Wilcoxon signed-rank tests (see Table 1 for an overview or the supplementary materials for detailed results, confidence intervals, and p-values). This validates that the new methodology delivers conclusions highly similar to the conventional pairwise evaluation.

The fact that C-P tests identified fewer significant differences than t-tests and Wilcoxon signed-rank tests did on the HEMVIP responses is likely because C-P tests only consider the sign of the difference between paired system ratings. This uses less of the information available in the ratings, thus requiring more samples to identify a statistically significant difference. We therefore recommend the two more granular tests for HEMVIP (especially the Wilcoxon, which does not assume Gaussianity) as they leverage more of the information in the responses and are better at telling

---

[1]Obtained online from http://doi.org/10.6084/m9.figshare.13055609.
[2]Obtained online from https://doi.org/10.6084/m9.figshare.13055585.v6.

Table 1: Significant results found by HEMVIP against those found by a re-analysis of the data from the pairwise user study in [13]. "Significant in [13]?" marks which system contrasts that were found to be significantly different by Holm-Bonferroni-corrected Clopper-Pearson (C-P) tests (preference ignoring ties) applied to the data from [13]. Other rows report whether or not these findings of significance agree with the significant differences found by HEVMIP in our user studies, using either C-P tests or pairwise t-tests (mean rating), both with Holm-Bonferroni correction. Empty cells signify agreement (no difference), while + or − mean that HEMVIP found, respectively did not find, the system contrast significant, unlike "Significant in [13]?".

System contrast	Full→NoAR				Full→NoPCA				Full→NoFiLM				Full→NoAudio				Full→NoText				Full→NoVel				NoPCA→GT			
Study question	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Significant in [13]?	✓	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓
C-P test difference				+								−	−														−	−
t-test difference				+																								

conditions apart. HEMVIP additionally allows for comparing all systems against one another (28 pairs), and not only the contrasts in Table 1, at no additional cost. Analyzing our data in this way using the Wilcoxon test finds (after Holm-Bonferroni correction) 26, 23, 24, and 24 of the 28 pairs to be statistically significantly different from one another at $\alpha = 0.05$, for questions Q1, Q2, Q3, and Q4.

One of the key benefits of HEMVIP is its efficiency, i.e., the time it takes to provide a certain amount of ratings for a single question. Compared to pairwise comparisons like in, e.g., [11, 13], HEMVIP is more efficient as it evaluates multiple stimuli in parallel. To quantify this efficiency, it is perhaps more meaningful to compare against the partial replication study in [11]. That study only considered the contrast "NoPCA" vs. "NoText" but otherwise used the same evaluation methodology as [13] and reached the same conclusions. Although [13] evaluated a greater set of conditions than [11] (which made it more relevant for our validation study, by providing more data on how HEMVIP compares to a pairwise approach) it is not as suitable for comparing test-taking efficiency to HEMVIP. This is because [13] asked four questions simultaneously, whereas HEMVIP and [11] present one question at a time, and [13] was conducted on Amazon Mechanical Turk, while [11] also considered Prolific, like the validation study here. (Also, the attention checks in [13] were based on detecting heavily degraded videos and were more subtle than the attention checks employed here, but this is less relevant to timing information.) Prolific participants in [11] took approximately 32 s to complete one page in a pairwise comparison, while the average time our study participants spent per page was 176 s. This means that comparing one system to the other seven in a pairwise fashion takes 224 s (7×32 s), as opposed to 176 s using HEMVIP. Furthermore, if one wants to compare all systems with each other, HEMVIP still only needs 176 s (one page), as opposed to 896 s (8 choose 2 = 28 pages) using a pairwise method, where each system would be compared against a single other system on each page. HEMVIP thus scales much better than pairwise evaluations.

A limitation of HEMVIP is that it mainly focuses on one question at a time. A way of scoring multiple questions at once (e.g., for efficiently conducting proper Likert-scale evaluations; cf. [18]) could probably be added by, for example, adding a new set of sliders after one question has been rated, or putting in multiple sliders per stimulus. This – and related issues such as how many stimuli that can be evaluated in parallel without exhausting the participant – should be addressed in a future usability study, as it is currently unknown if adding an extra question or increasing the amount of stimuli would be too cognitively demanding for the users. The MUSHRA stan-

dard for audio [7] recommends no more than 12 stimuli per page. A majority of participants indicated repetitiveness from rating 8 videos with the same sound per page, which could indicate that a lower number than 8 could be preferable. Another limitation of this work is that we only evaluated video stimuli from one particular domain and task. While we do evaluate it using material that has been evaluated twice before, and found consistent results, future work should investigate different kinds of video stimuli to further study how well the method generalizes to different situations.

6 CONCLUSIONS AND FUTURE WORK

In this paper we proposed a framework for evaluating multiple comparable video stimuli in parallel, called HEMVIP. A validation experiment compared results obtained using the proposed method to results obtained using pairwise binary preference tests in earlier work in the domain of nonverbal behavior, finding high correspondence between the two experiments, but with greater efficiency and vastly better scaling properties for HEMVIP. It has for example already been used for a large-scale evaluation in the GENEA Challenge 2020 [14]. We believe HEMVIP can be of great benefit to researchers performing thorough evaluations across multiple video stimuli – not only as an alternative to pairwise video presentations but also as a replacement for many MOS tests [9]. Furthermore, the video stimuli need not show or compare gestures or feature ECAs: the key point is that stimuli from all conditions can be compared side by side, e.g., different voiceover audio for the same video clip, or comparisons of different signal processing methods. With some code modifications, stimuli may also be displayed on a separate device, e.g., for VR/AR evaluations.

Future work includes validating on stimuli from other domains, along with usability testing of various aspects of the interface, such as how many videos users are able to rate effectively on a single page without incurring excessive cognitive load.

REFERENCES

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

[2] David A. Braude. 2016. *Head motion synthesis: Evaluation and a template motion approach.* Ph.D. Dissertation. The University of Edinburgh, UK. http://hdl.handle.net/1842/20418

[3] Chung-Cheng Chiu and Stacy Marsella. 2014. Gesture generation with low-dimensional embeddings. In *Proceedings of the International Conference on Autonomous Agents and Multi-agent Systems.* ACM, 781–788. https://doi.org/10.5555/2615731.2615857

[4] Andrew P. Clark, Kate L. Howard, Andy T. Woods, Ian S. Penton-Voak, and Christof Neumann. 2018. Why rate when you could compare? Using the "Elo-Choice" package to assess pairwise comparisons of perceived physical strength. *PLOS One* 13, 1 (2018), 1–16. https://doi.org/10.1371/journal.pone.0190393

[5] Ylva Ferstl and Rachel McDonnell. 2018. Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the ACM International Conference on Intelligent Virtual Agents.* ACM, 93–98. https://doi.org/10.1145/3267851.3267898

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

[7] ITU-R BS.1534-3. 2015. *Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems.* Standard. ITU. https://www.itu.int/rec/R-REC-BS.1534-3-201510-I

[8] ITU-R BT.1788. 2007. *Methodology for the Subjective Assessment of Video Quality in Multimedia Applications.* Standard. ITU. https://www.itu.int/rec/R-REC-BT.1788

[9] ITU-T P.800. 1996. *Methods for Subjective Determination of Transmission Quality.* Standard. ITU. https://www.itu.int/rec/T-REC-P.800-199608-I

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

[11] Patrik Jonell, Taras Kucherenko, Ilaria Torre, and Jonas Beskow. 2020. Can we trust online crowdworkers? Comparing online and offline participants in a preference test of virtual agents.. In *Proceedings of the ACM International Conference on Intelligent Virtual Agents.* ACM, Article 31, 8 pages. https://doi.org/10.1145/3383652.3423860

[12] 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. *International Journal of Human–Computer Interaction* 37, 14 (2021), 1300–1316. https://doi.org/10.1080/10447318.2021.1883883

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

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

[15] Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology* 140 (1932), 1–55.

[16] Manuel Sam Ribeiro, Junichi Yamagishi, and Robert A. J. Clark. 2015. A perceptual investigation of wavelet-based decomposition of $f0$ for text-to-speech synthesis. In *Proceedings of the Annual Conference of the International Speech Communication Association.* ISCA, 1586–1590. https://www.isca-speech.org/archive/interspeech_2015/i15_1586.html

[17] Michael Schoeffler, Sarah Bartoschek, Fabian-Robert Stöter, Marlene Roess, Susanne Westphal, Bernd Edler, and Jürgen Herre. 2018. webMUSHRA — A comprehensive framework for web-based listening tests. *Journal of Open Research Software* 6, 1 (2018), 8:1–8:8. https://doi.org/10.5334/jors.187

[18] Mariah L. Schrum, Michael Johnson, Muyleng Ghuy, and Matthew C. Gombolay. 2020. Four years in review: Statistical practices of Likert scales in human-robot interaction studies. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction.* ACM, 43–52. https://doi.org/10.1145/3371382.3380739

[19] Mirjam Wester, Cassia Valentini-Botinhao, and Gustav Eje Henter. 2015. Are we using enough listeners? No! An empirically-supported critique of Interspeech 2014 TTS evaluations. In *Proceedings of the Annual Conference of the International Speech Communication Association*, Vol. 16. ISCA, 3476–3480. https://www.isca-speech.org/archive/interspeech_2015/i15_3476.html

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

[21] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *Proceedings of the IEEE International Conference on Robotics and Automation.* IEEE, 4303–4309. https://doi.org/10.1109/ICRA.2019.8793720