A summary of the GENEA Challenge 2022 on speech-to-gesture generation

Pieter Wolfert[*1], *Taras Kucherenko*[*2], *Youngwoo Yoon*[*3],
Teodor Nikolov[4], *Mihail Tsakov*[4], *Gustav Eje Henter*[5]
[1]*Ghent University – imec, Belgium* [2]*SEED – Electronic Arts, Sweden* [3]*ETRI, Republic of Korea* [4]*Umeå University, Sweden* [5]*KTH Royal Institute of Technology, Sweden*
[*]*Joint first authors*

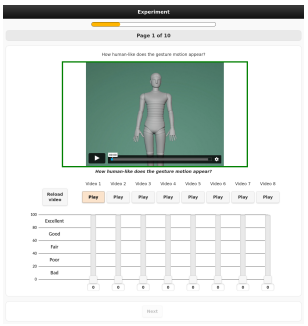
pieter.wolfert@ugent.be, tkucherenko@ea.com, youngwoo@etri.re.kr,
tnikolov@hotmail.com, tsakovm@gmail.com, ghe@kth.se

Verbal and nonverbal communication are important and complementary components of embodied human communication. In such communication, speech is typically accompanied by co-speech gestures or gesticulation, performed by the hands, head, and occasionally the body. Automatically generating such co-speech gestures is an important task to speed up character animation and improve human-agent interaction, seeing that a substantial fraction of human communication takes place through co-speech gestures McNeill (1992); Kendon (2004). In particular, gesticulation has been shown to enhance interactions with embodied conversational agents (ECAs) Bergmann and Macedonia (2013); Luo et al. (2013), e.g., helping with learning tasks Bergmann and Macedonia (2013) and leading to a higher sense of co-presence Wu et al. (2014).

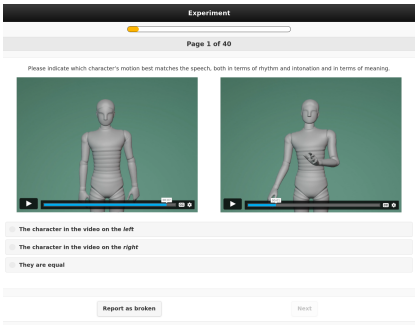
This paper is concerned with systems for automatic generation of nonverbal behaviour, and how these can be compared in a fair and systematic way in order to advance the state-of-the-art. Synthetic gestures used to be based on rule-based systems, e.g., Cassell et al. (2001); Salvi et al. (2009), but these are gradually being supplanted by data-driven approaches, e.g., Bergmann and Kopp (2009); Levine et al. (2010), with recent work Alexanderson et al. (2020); Yoon et al. (2020) showing improvements in gesticulation production for ECAs. Challenges like the Blizzard Challenges have been of great use for advancing text-to-speech technology King (2014), inspiring us to conduct the first and second challenge in speech-driven gesture generation, the GENEA (Generation and Evaluation of Non-verbal Behaviour for Embodied Agents) Challenges. In these challenges, participating teams built automatic gesture-generation systems using a common dataset. Motion produced by these systems was then evaluated in several large, crowdsourced user studies using the same motion-rendering pipeline. Differences in evaluation outcomes are then attributable only to differences between the motion-generation approaches.

The 2022 challenge dataset was based on Lee et al. (2019), 18 hours of full-body motion capture, including fingers, of different persons engaging in dyadic conversations. Only one side of the conversation was considered at a time. Ten teams participated in the evaluation across two tiers: full-body ('F') and upper-body gesticulation ('U'). Not all teams participated in all tiers, and (like in the Blizzard Challenges) results are reported using anonymous labels, prefixed 'FS' and 'US' for the tiers. The evaluation also included natural motion capture (FNA/UNA) and two baseline systems (UBA and FBT/UBT). For each tier we evaluated both the human-likeness (i.e., quality) of the gesture motion and its appropriateness for the specific speech. Figure 1 shows the two evaluation interfaces used. Human-likeness was evaluated by 121 and 150 crowdsourced test-takers (for F and U, respectively) using the HEMVIPmethodology Jonell et al. (2021), with audio removed from the videos, to control for the effect of the speech on gesture perception. Appropriateness was assessed by 247 and 304 test-takers based on a matched-mismatched paradigm. Here, two videos with the same speech audio but different motion were presented. The motion always came from the same system, but from different speech input: in one case, it came from the speech in the video, but in the other case, it was motion associated with some other, unrelated speech. How often subjects selected the matched video as more appropriate gives an idea of how specific the motion is to the speech. This methodology decouples gesture appropriateness from the human-likeness of the motion, which has been a major confounder in previous evaluations. We also compared several objective metrics of motion quality to the subjective ratings from our user studies. Unfortunately, the objective metrics were not well aligned with human perception. In the full-body tier, one of the least human-like systems, FSB, received some of the best scores in terms of average absolute jerk, acceleration, and Hellinger distance. At the same time, one of the most human-like systems, FSC, was not in the top three according to any of the objective metrics used. In the upper-body tier, one of the least human-like systems, USP, was in the top three systems according to average jerk, acceleration, and Hellinger distance, whilst one of the most human-like systems, USO, is not in the top three according to any of the objective metrics. The rank correlations in Table 1 make these observations more precise, by showing that most correlations are not statistically significantly different from zero. The one exception is the FGD metric Yoon et al. (2020). Although the correlations we found there are moderate (around −0.5) and system USN shows an outlying value, FGD might nonetheless have some potential for faster evaluation in the development phase, though how well it will resolve smaller differences between systems is unclear. The evaluation results, shown in Figure 2, are both surprising and revealing. In each tier, one synthetic condition is rated as significantly more human-like than human motion capture. We believe this has not been demonstrated before on a high-fidelity avatar. On the other hand, all synthetic motion was found to be vastly less appropriate for the speech than the original motion-capture recordings are, even though our new paradigm controls for the influence of motion quality on the evaluation. This mirrors the situation in text-to-speech, where synthetic speech has attained human-like quality, but contextual appropriateness remains to be solved. Data and materials are available through the challenge webpage at https://youngwoo-yoon.github.io/GENEAchallenge2022/

Index Terms: speech-gesture generation, human-computer interaction

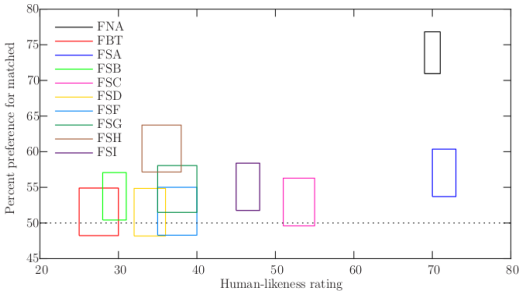


(a) Human-likeness interface and full-body video

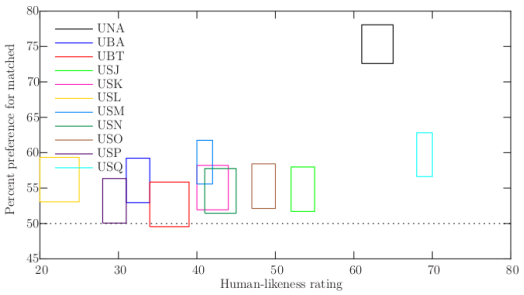


(b) Appropriateness interface and upper-body videos

Figure 1: Screenshots of the evaluation interface used in the studies, also showing the camera perspectives used by the tiers.



(a) Full-body results



(b) Upper-body results

Figure 2: Evaluation results for each tier. Each box is a system. Box widths show 95% confidence intervals for the median human-likeness rating. Box heights show appropriateness, as 95% confidence intervals for the preference for matched motion in percent

Metric	Average Jerk	Average accel.	Global CCA	Hellinger distance	FGD
τ	-0.11	-0.26	0.11	-0.40	-0.51
p-value	0.64	0.27	0.64	0.085	0.029

Table 1: Upper-body tier rank correlations (Kendall's τ) between objective dissimilarity metrics (how different generated motion is from recorded human mocap) and median human-likeness scores. Negative correlations are better.

References

Alexanderson, S., Henter, G. E., Kucherenko, T., and Beskow, J. (2020). Style-controllable speech-driven gesture synthesis using normalising flows. Comput. Graph. Forum, 39(2).

Bergmann, K. and Kopp, S. (2009). GNetIc – using Bayesian decision networks for iconic gesture generation. In Proc. IVA, pages 76–89.

Bergmann, K. and Macedonia, M. (2013). A virtual agent as vocabulary trainer: iconic gestures help to improve learners' memory performance. In Proc. IVA, pages 139–148.

Cassell, J., Vilhjálmsson, H. H., and Bickmore, T. (2001). BEAT: The behavior expression animation toolkit. In Proc. ACM SIGGRAPH, pages 477–486.

Jonell, P., Yoon, Y., Wolfert, P., Kucherenko, T., and Henter, G. E. (2021). HEMVIP: Human evaluation of multiple videos in parallel. In Proc. ICMI.

Kendon, A. (2004). Gesture: Visible Action as Utterance. Cambridge University Press.

King, S. (2014). Measuring a decade of progress in text-to-speech. Loquens, 1(1):e006.

Lee, G., Deng, Z., Ma, S., Shiratori, T., Srinivasa, S. S., and Sheikh, Y. (2019). Talking With Hands 16.2M: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In Proc. ICCV, pages 763–772.

Levine, S., Krähenbühl, P., Thrun, S., and Koltun, V. (2010). Gesture controllers. ACM Trans. Graph., 29(4).

Luo, P., Ng-Thow-Hing, V., and Neff, M. (2013). An examination of whether people prefer agents whose gestures mimic their own. In Proc. IVA, pages 229–238.

McNeill, D. (1992). Hand and Mind: What Gestures Reveal about Thought. University of Chicago Press.

Salvi, G., Beskow, J., Al Moubayed, S., and Granström, B. (2009). SynFace—speech-driven facial animation for virtual speech-reading support. EURASIP J. Audio Spee.

Wu, Y., Babu, S. V., Armstrong, R., Bertrand, J. W., Luo, J., Roy, T., Daily, S. B., Dukes, L. C., Hodges, L. F., and Fasolino, T. (2014). Effects of virtual human animation on emotion contagion in simulated inter-personal experiences. IEEE T. Vis. Comput. Gr., 20(4):626–635.

Yoon, Y., Cha, B., Lee, J.-H., Jang, M., Lee, J., Kim, J., and Lee, G. (2020). Speech gesture generation from the trimodal context of text, audio, and speaker identity. ACM Trans. Graph., 39(6