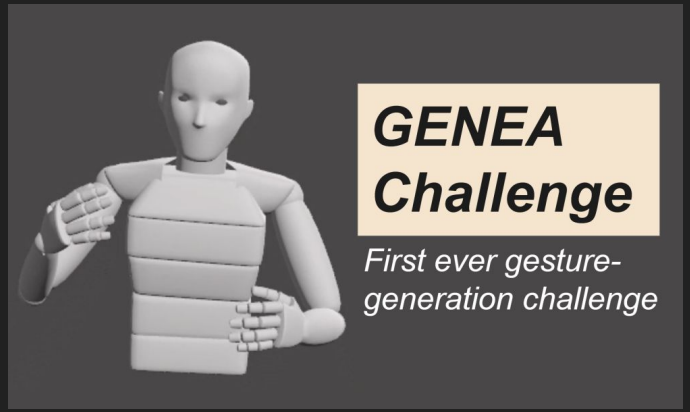
A large, crowdsourced evaluation of gesture generation systems on common data: The GENEA Challenge 2020



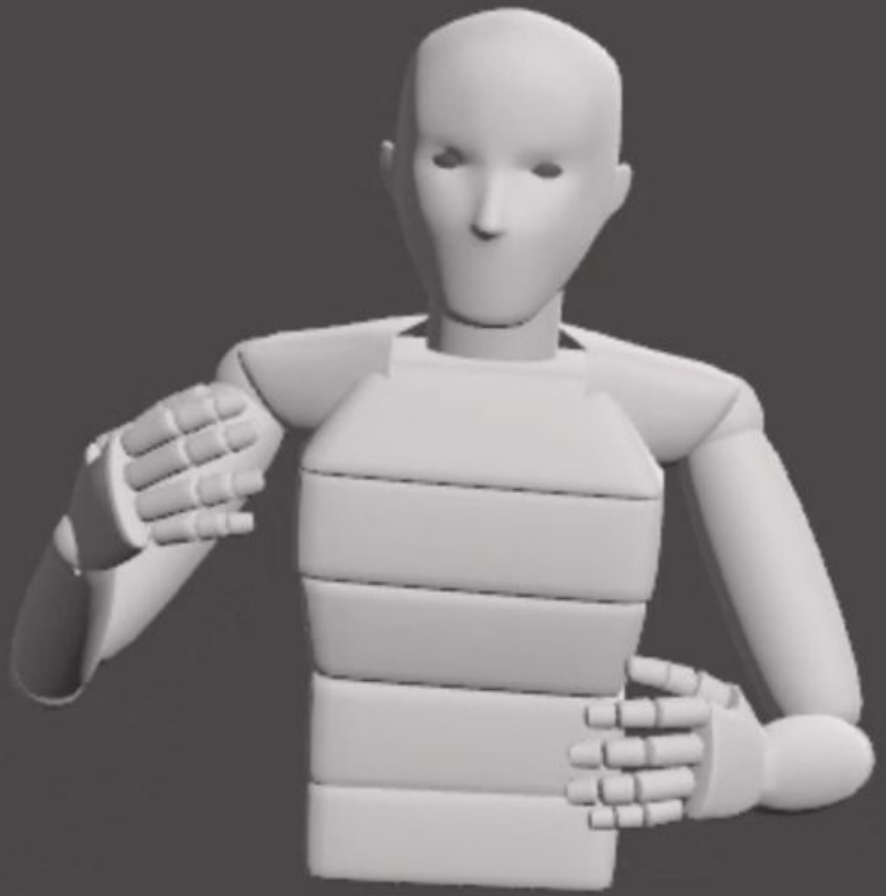
Taras Kucherenko*, Patrik Jonell*, Youngwoo Yoon*, Pieter Wolfert, and Gustav Eje Henter

KTH VETENSKAP OCH KONST

ETRI

GHENT UNIVERSITY

"*" indicates joint first authors



GENEA Challenge

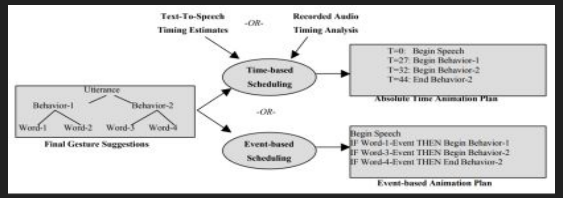
First ever gesture-generation challenge

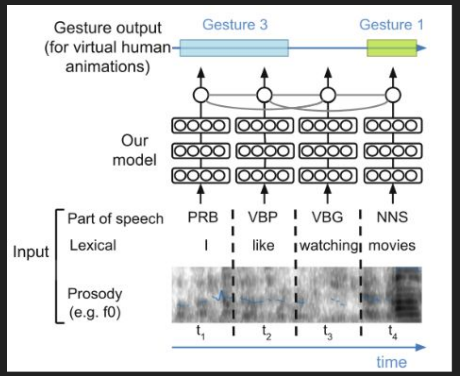
Take-home message

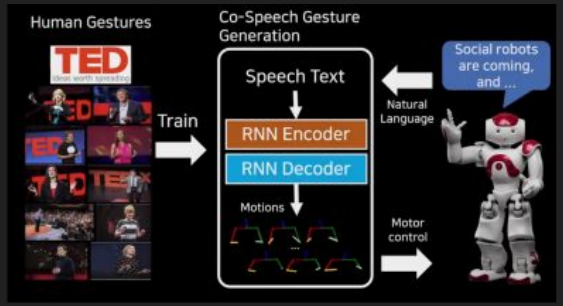
- We ran the first ever challenge in speech-driven gesture generation
  - To enable direct comparison between different gesture-generation methods
- Open science
  - Code, data, responses, analysis scripts, and system descriptions are publicly available
  - Can be used for future benchmarking and perception research

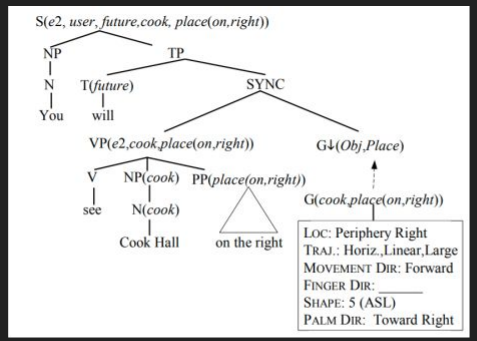
Importance of gestures

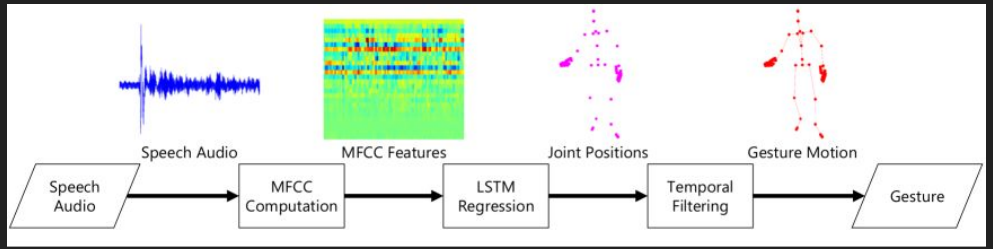
Gesture-generation field











Benchmarking: The weakest link

- Previous work often do not compare to other systems
- The comparisons to previous works are rare and small scale
- What is the state-of-the-art is not clear

How do people evaluate objectively?

Many of us don’t

CCA

Sadoughi & Busso (2019)
Bozkurt et al. (2016, 2020)

Motion Statistics

Ahuja et al. (2020)
Yoon et al. (2020)
Ferstl et al. (2020)
Kucherenko et al (2019, 2020)

FID, FGD - Fréchet Distance

Yoon et al. (2020),
Ahuja et al. (2020)

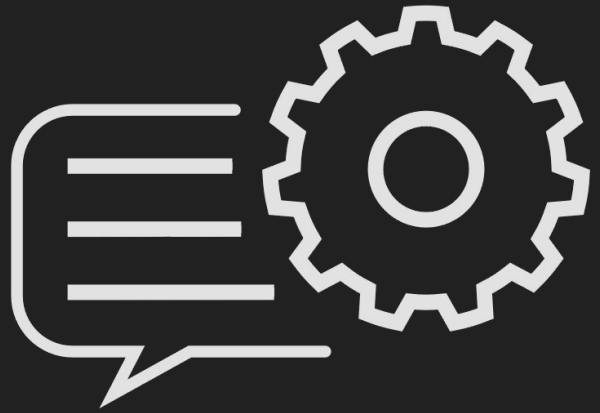
Likelihood

Alexanderson et al. (2020)

~~MAE, MRSE, PCK~~

Benchmarking is essential

- It has moved several other fields forward, such as :
  - Speech synthesis
  - Computer vision
  - Natural Language Processing
  - Machine learning
  - Etc ...

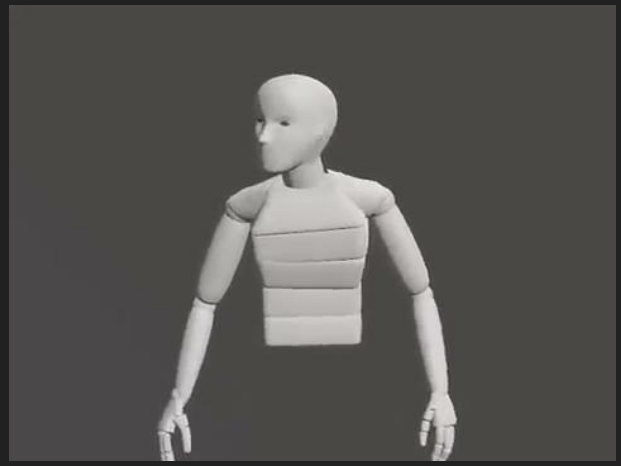


So... we run the first *gesture-generation challenge*

Dataset for the challenge

Trinity Speech-Gesture Dataset:

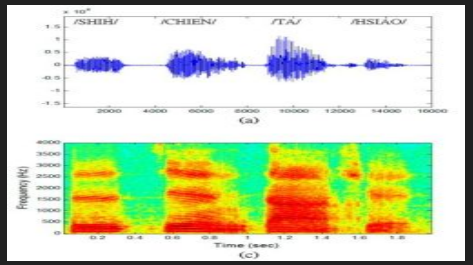
- 244 minutes of audio and 3D motion capture recordings of one male actor
- Only the 15 upper-body joints
- No finger data included



We extended the dataset with the audio transcriptions and made them publicly available alongside the original dataset

Ferstl, Ylva, and Rachel McDonnell. "Investigating the use of recurrent motion modelling for speech gesture generation." *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 2018.

Challenge pipeline



System

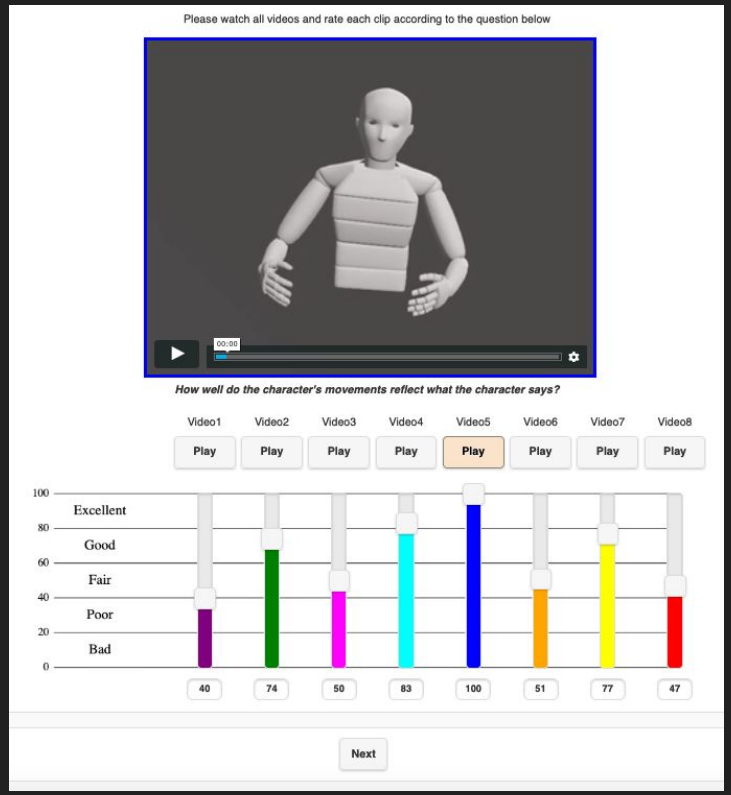
Joint angle rotations

Standardised

Varied

Standardised

Evaluation interface



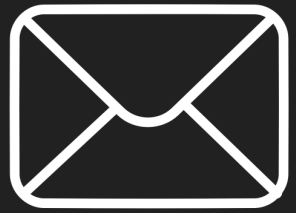
Jonell, Patrik, Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, and Gustav Eje Henter. "HEMVIP: Human Evaluation of Multiple Videos in Parallel." *arXiv preprint arXiv:2101.11898* (2021).

12


Statistics

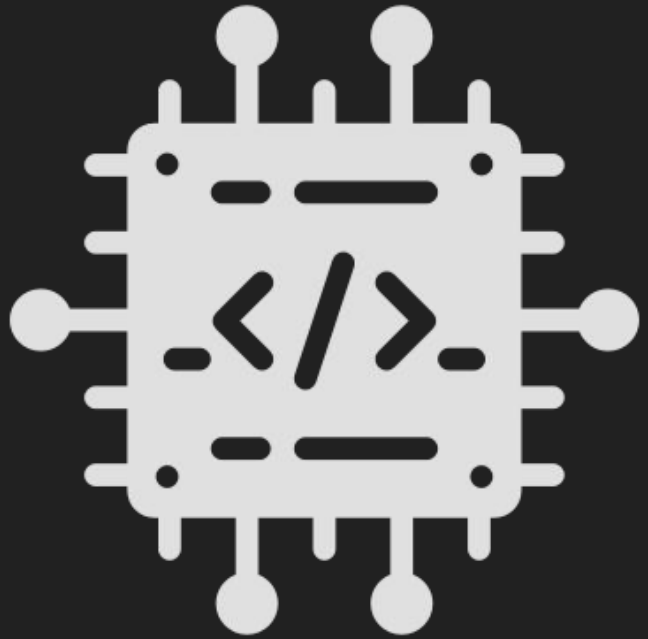
15 registrations

5 submissions

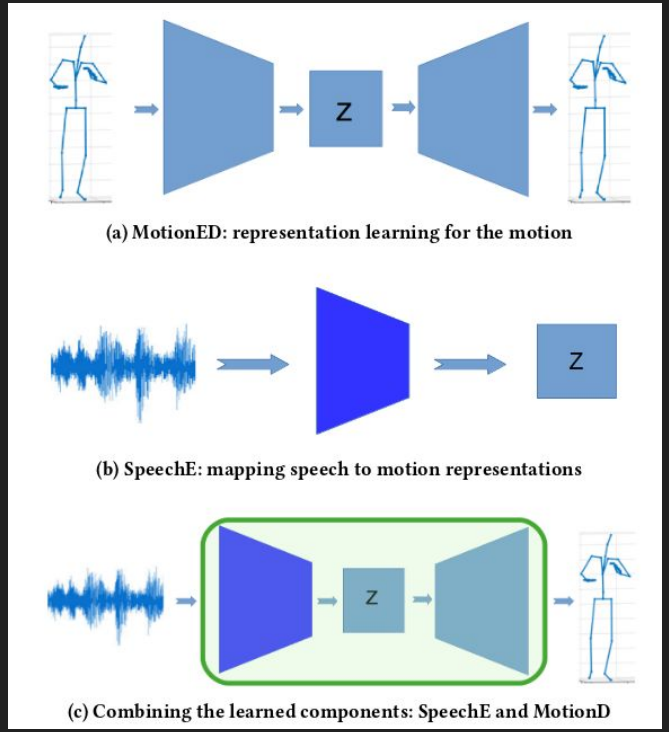


2 code repositories

Baselines and systems



Audio-only baseline (BA)



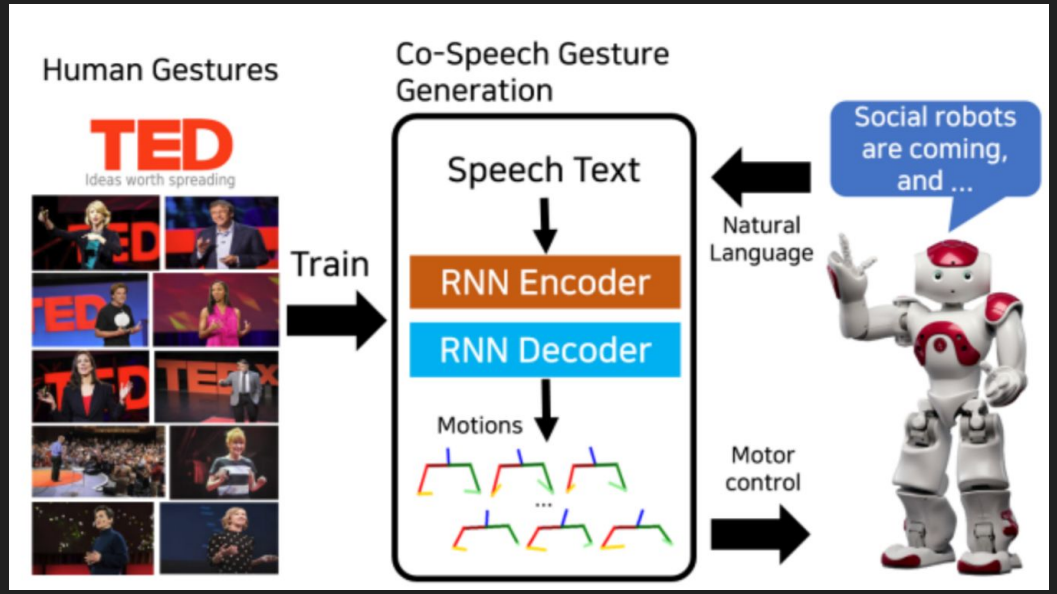
Kucherenko, Taras, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. "Analyzing input and output representations for speech-driven gesture generation." In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pp. 97–104. 2019.

Audio-only baseline (BA)

- Trained on the challenge dataset
- Synthesizes joint rotation values instead of joint positions
- Smoothed using Savitzky–Golay filter
- Hyper-parameters tuned for the new data

Kucherenko, Taras, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. "Analyzing input and output representations for speech-driven gesture generation." In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pp. 97–104. 2019.

Text-only baseline (BT)

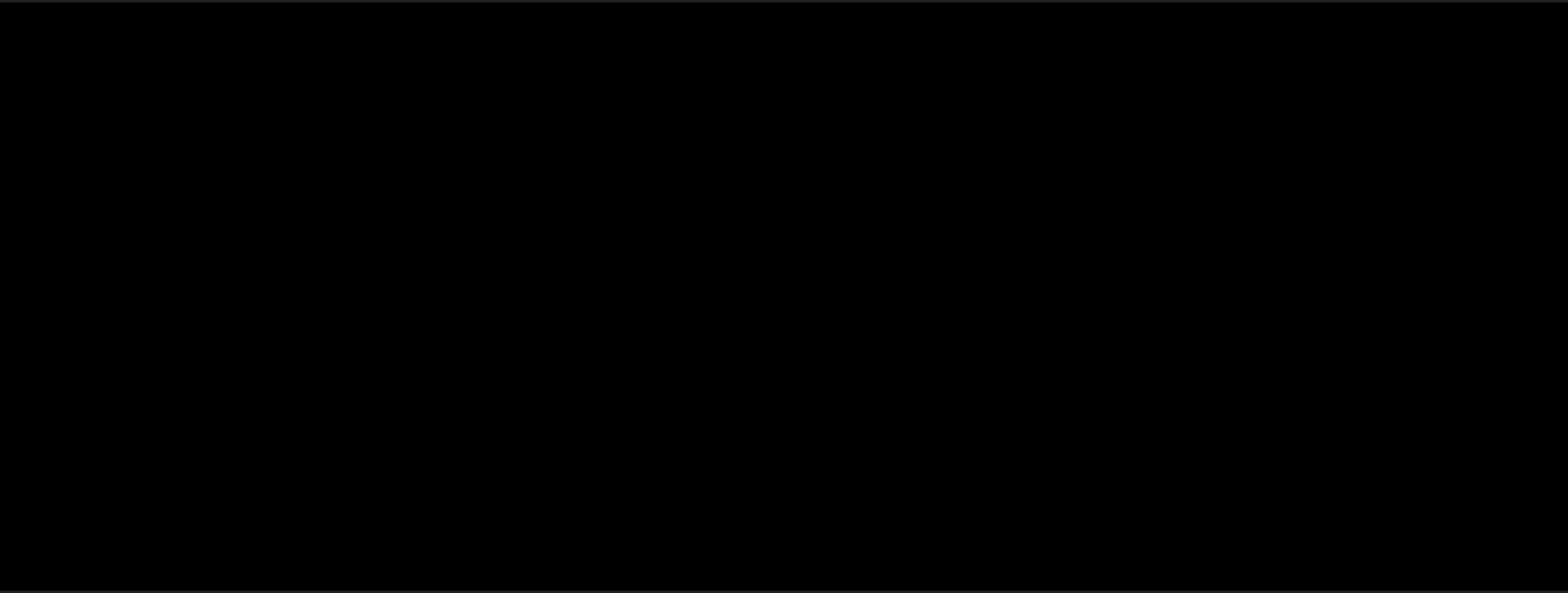


Yoon, Youngwoo, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. "Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots." In *Proceedings of the IEEE International Conference on Robotics and Automation*. IEEE, 2019.

Text-only baseline (BT)

- Trained on the challenge dataset
- Synthesizes joint rotation values instead of joint positions
- Pretrained FastText instead of GloVe

Yoon, Youngwoo, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. "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, 2019.



Ground Truth

Mismatched

User Study (x2!)

Crowdsourced

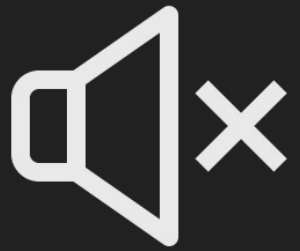
125

8

Questions asked in user studies

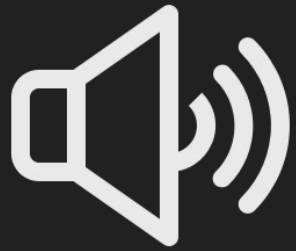
1

“How human-like does the gesture motion appear?”



2

“How appropriate are the gestures for the speech?”

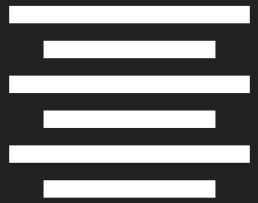


Conditions

N

M

BA

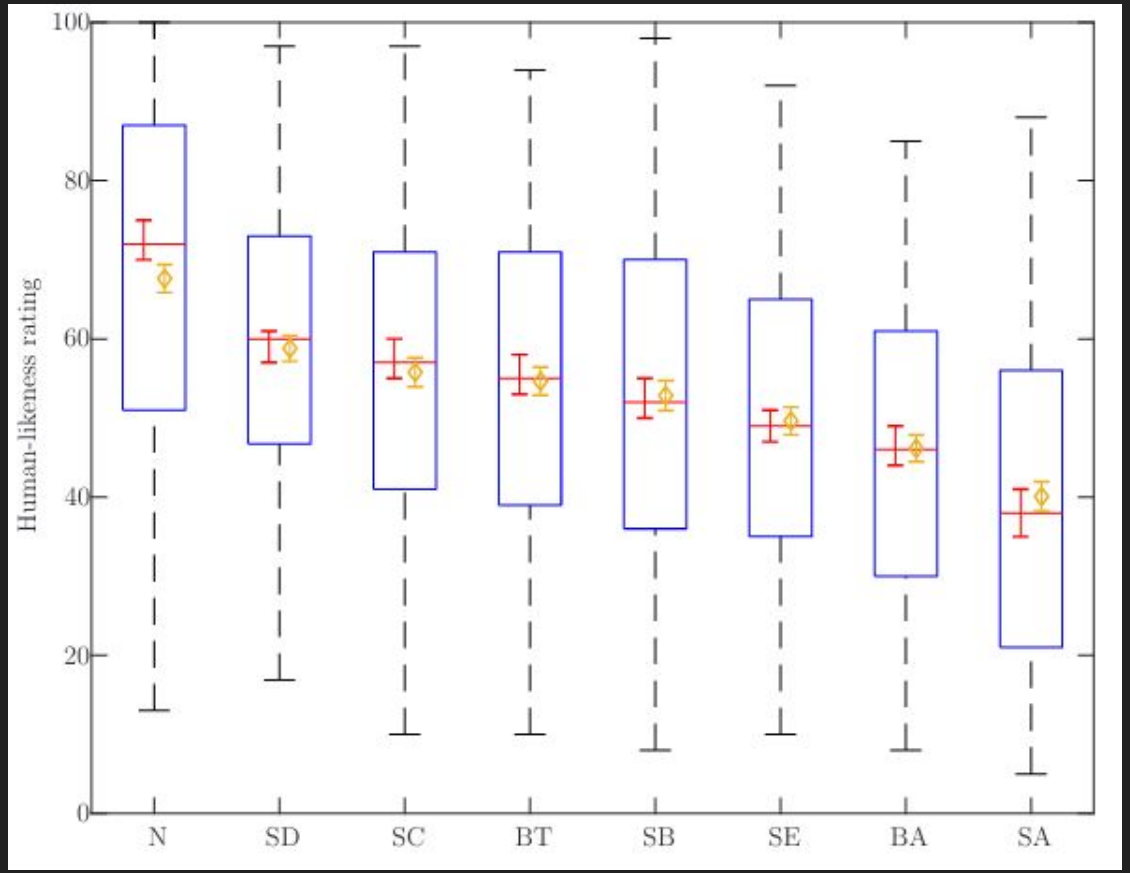


BT

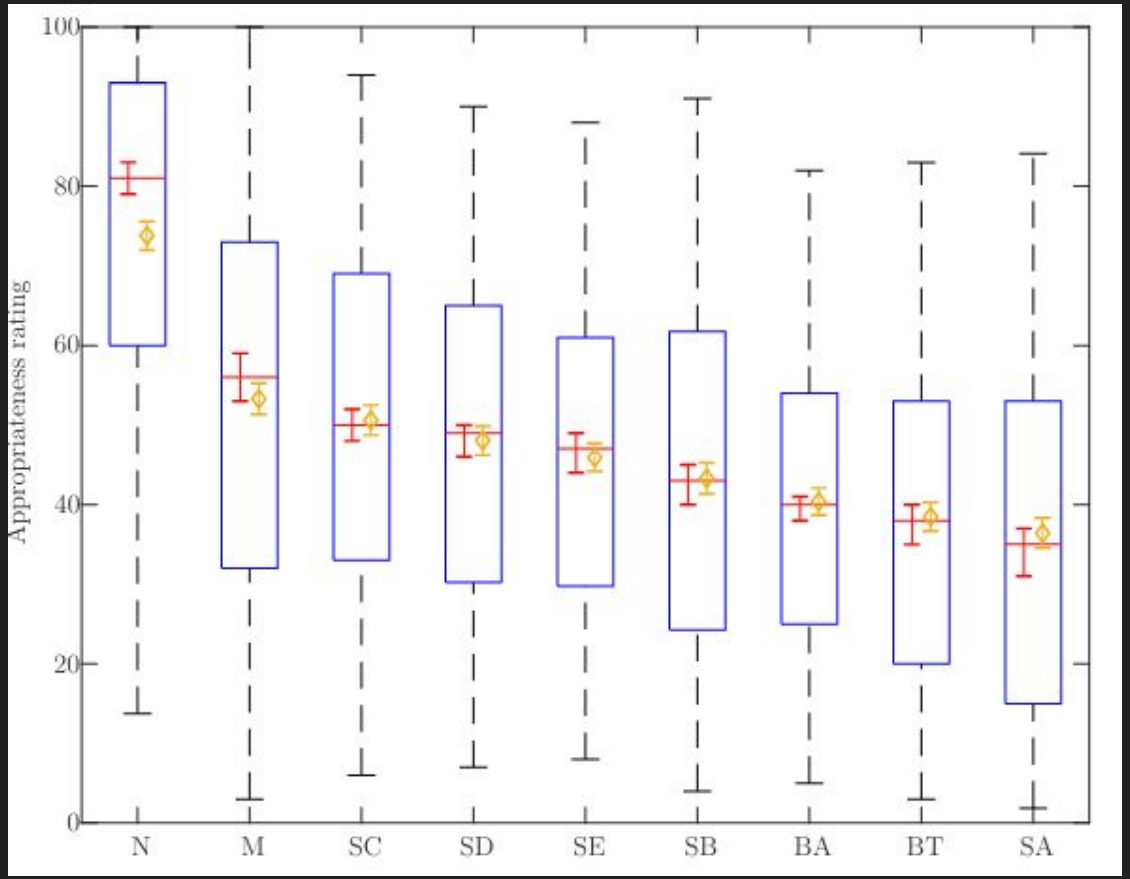
+5

SA – SE

Results



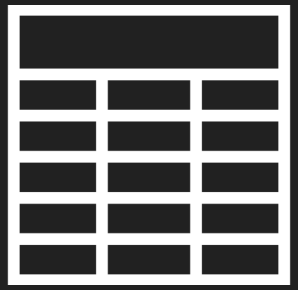
Human-likeness ratings



Appropriateness ratings

Open science

Motion & video



Subj. & obj.

Visual. & eval.

System descriptions

Lessons learned:
State of gesture generation

- Data-driven gesture-generation methods are improving

- The gap between human and generated gestures remains substantial

- Human-like does not mean being appropriate for gestures of a virtual avatar; systems can be good at different things

28

Lessons learned: Gesture-generation evaluation

- A MUSHRA-like evaluation (HEMVIP) can be successfully used to benchmark numerous gesture-generation models in parallel

- Established objective measures are not well correlated with subjective evaluations

- Appropriateness is not easy to disentangle from human-likeness

29

Lessons learned:
Arranging challenges

- Challenges address an unmet need in the community
- Finding a good dataset is a bottleneck in this field
- Actively reaching out to and engaging existing community members is important to foster the participation and relevance of a challenge

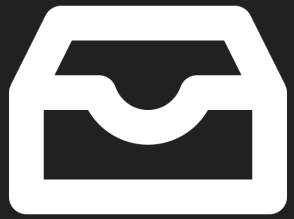
30

Questions?

Timeline

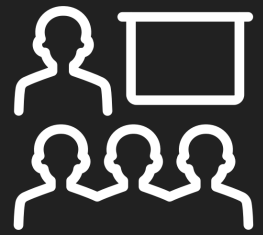
Rules

Rules



1 submission

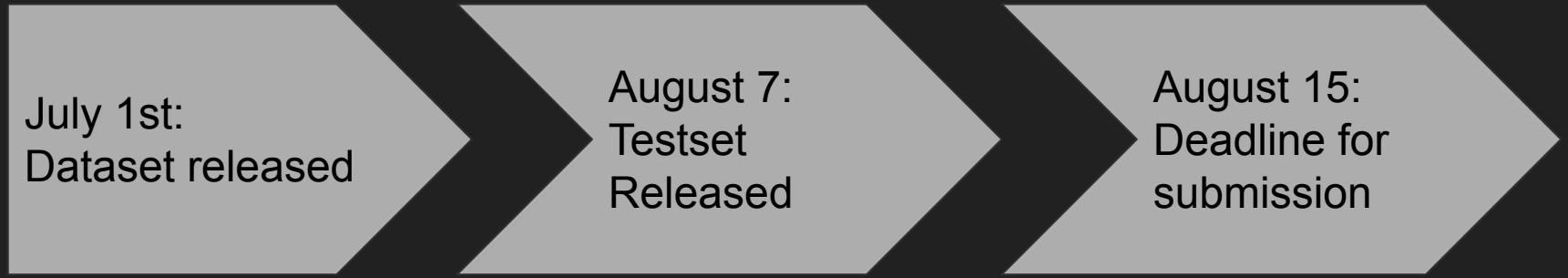
No post-processing



Presentation Obligatory

Limits on external data

Timeline


July 1st:
Dataset released
August 7:
Testset
Released
August 15:
Deadline for
submission

OBJECTIVE

36

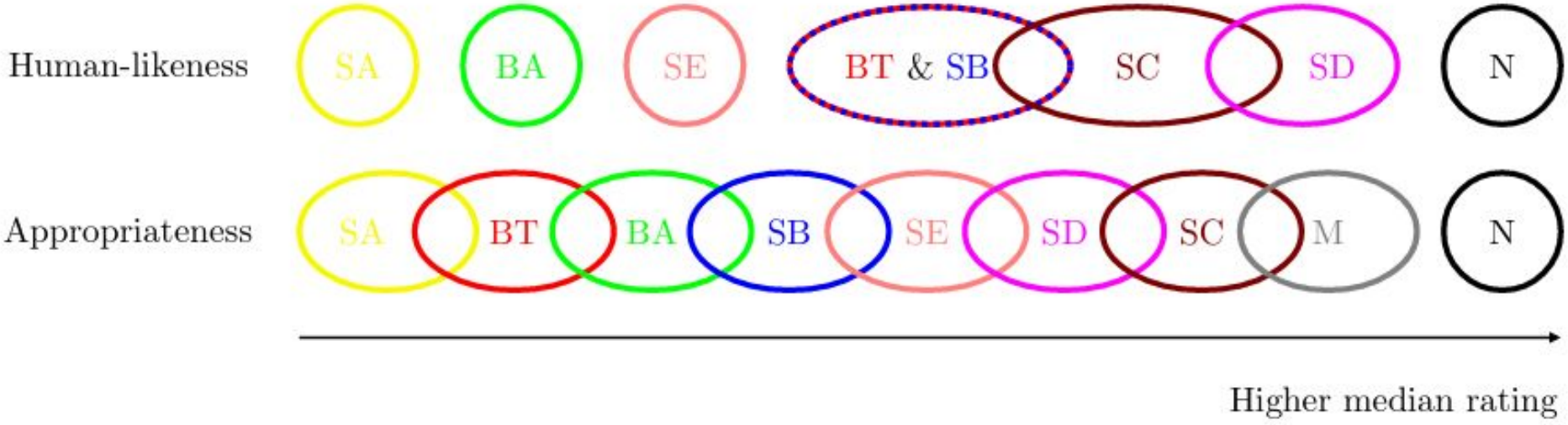
Average jerk

Distance between speed histograms

SUBJECTIVE

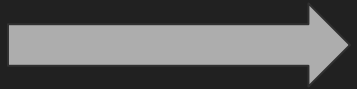
39

Partial ordering between systems



How do researchers evaluate subjectively?

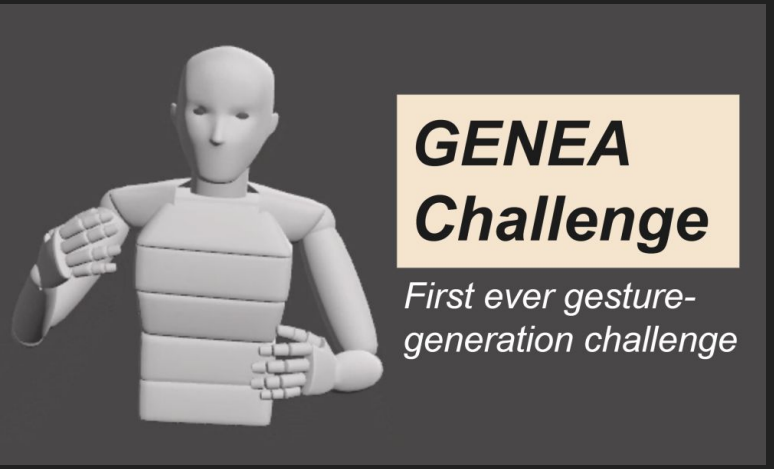
Likert Scales
Ishi et al. (18)
Hasegawa et al. (18)
Kucherenko et al (19)
Yoon et al. (19)



Preference Test
Ginosar et al. (19)
Kucherenko et al (20)
Ahuja et al. (20)
Yoon et al. (20)

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

Objective evaluation



A large, crowdsourced evaluation of gesture generation systems on common data: The GENEA Challenge 2020

Taras Kucherenko*, Patrik Jonell*, Youngwoo Yoon*, Pieter Wolfert, and Gustav Eje Henter

** indicates joint first authors*

KTH VETENSKAP OCH KONST

ETRI

GHENT UNIVERSITY