GENEA 2020 Challenge Introduction

Dataset for the Challenge

Trinity Speech-Gesture Dataset:

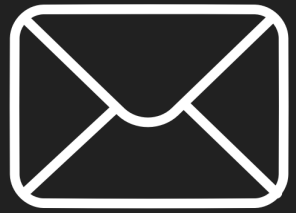
- 244 minutes of audio and 3D motion capture recordings of one male actor
- Only the 15 upper-body joints
- No finger data included
- Speech audio with time-aligned transcriptions

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.


Statistics

15 registrations

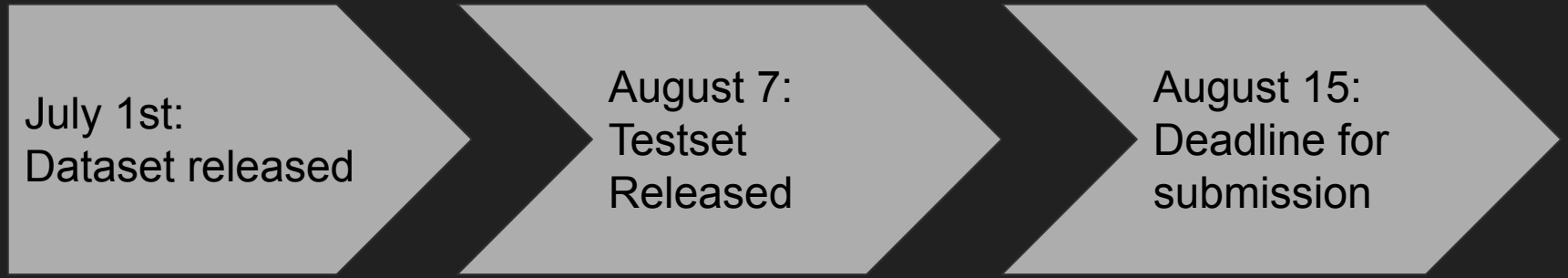
5 submissions



2 code repositories

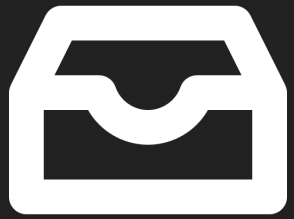
Timeline

Timeline


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

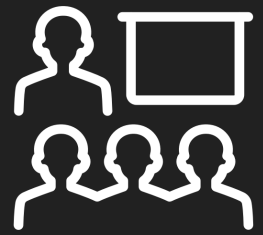
Rules

Rules



1 submission

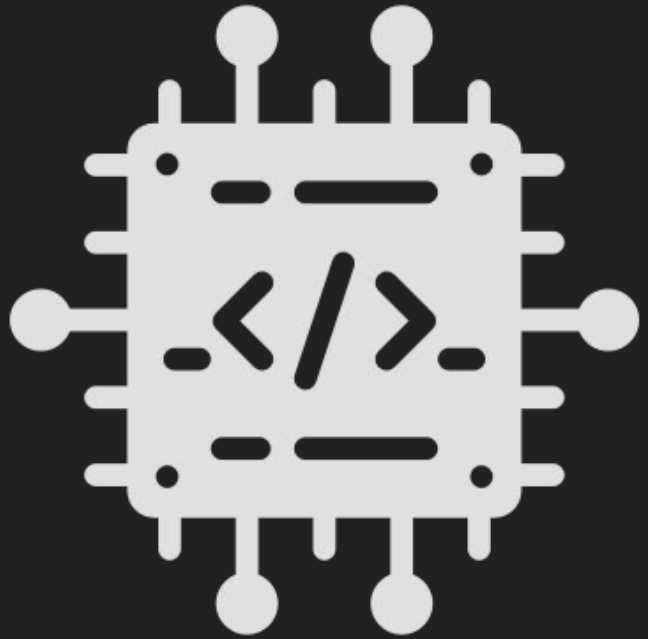
No post-processing



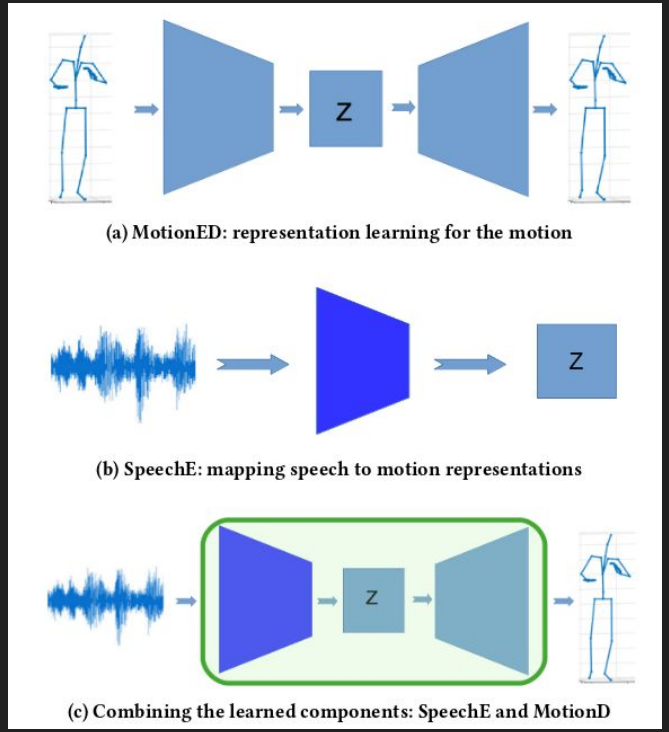
Presentation Obligatory

Limits on external data

Baselines and Systems



Audio-only baseline (BA)



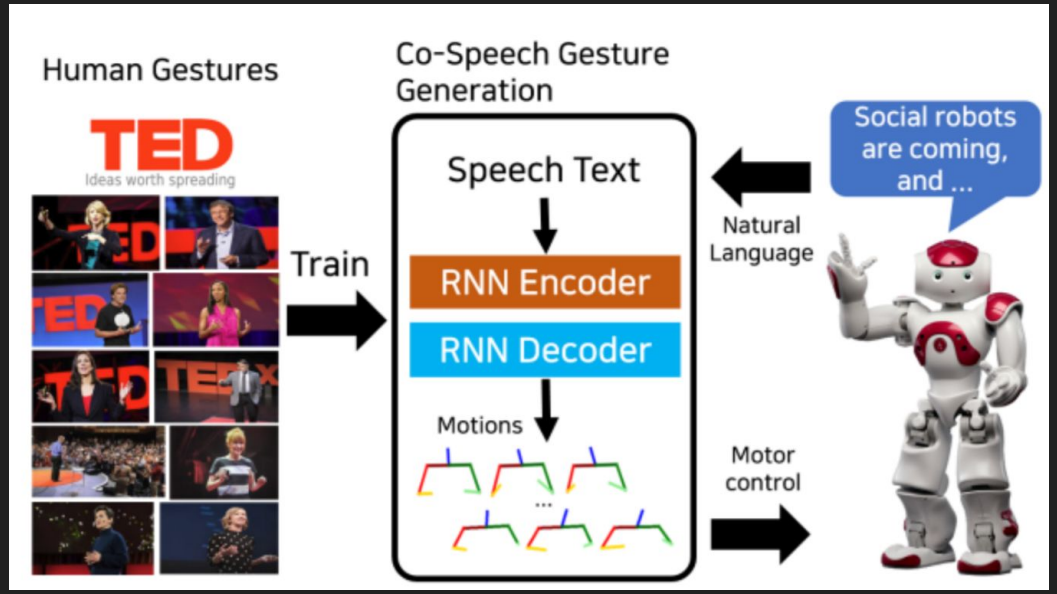
Kucherenko, Taras, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. "Analyzing input and output representations for speech-driven gesture generation." In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pp. 97-104. 2019.

Audio-only baseline (BA)

- Trained on the challenge dataset
- Synthesizes joint rotation values instead of joint positions
- Smoothed using Savitzky–Golay filter
- Different hyper-parameters

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.

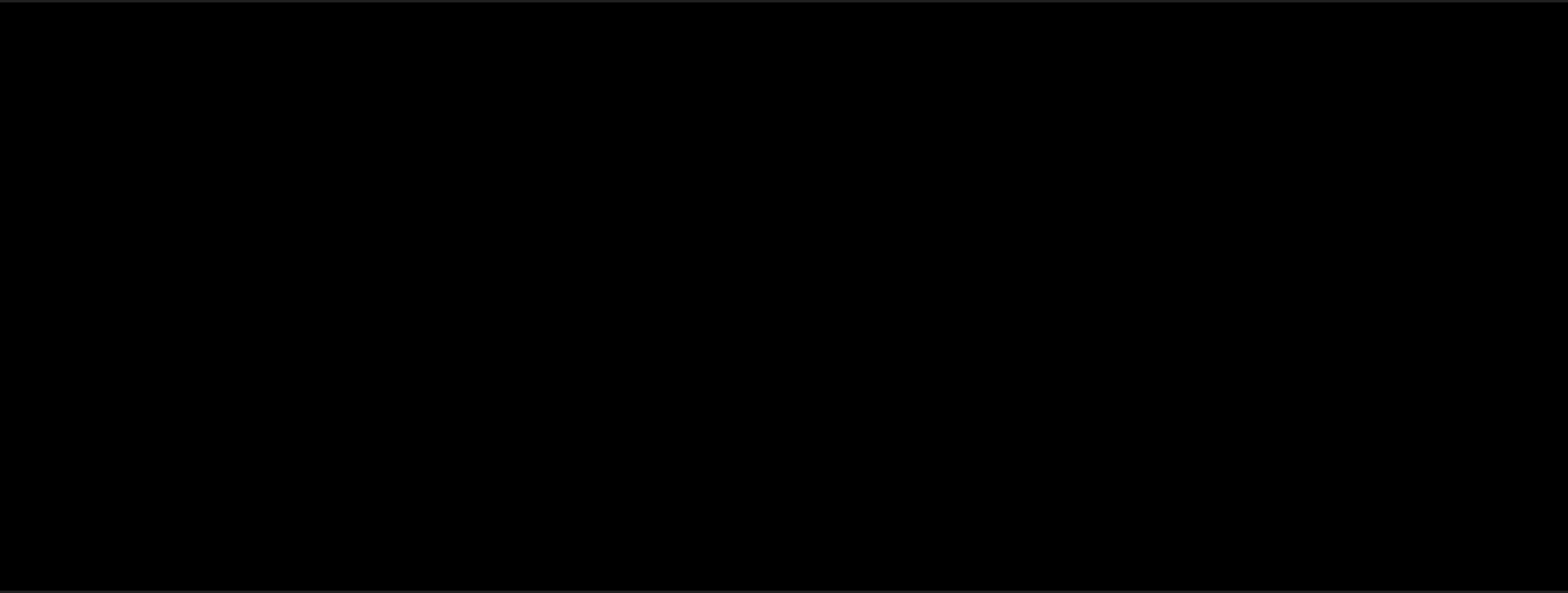
Text-only baseline (BT)



13 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 ICRA. IEEE, 2019.

Text-only baseline (BT)

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



Ground Truth

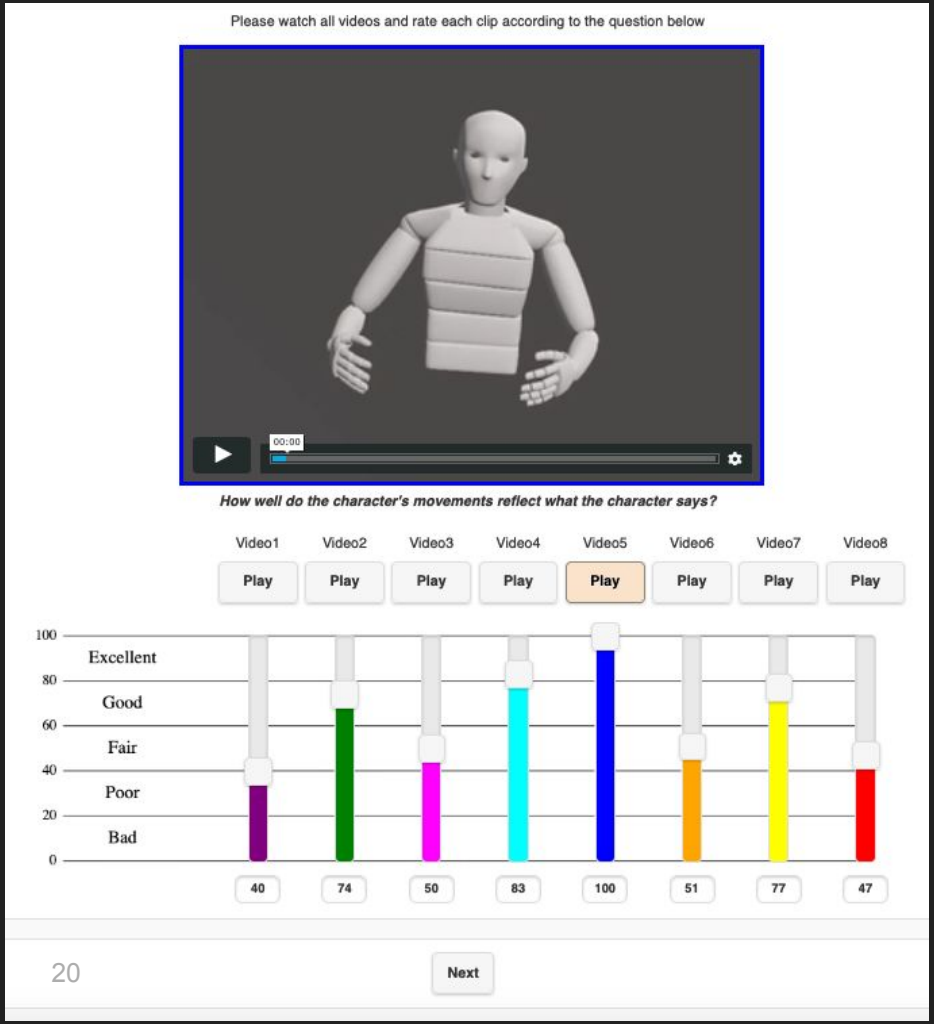
Mismatched

OBJECTIVE

Average jerk

Distance between speed histograms

SUBJECTIVE



Evaluation interface

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

“How appropriate are the gestures for the speech?”

User Study

Crowdsourced

125

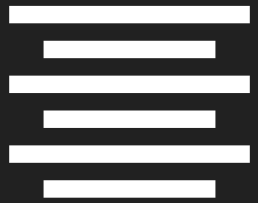
8

Conditions

N

M

BA



BT

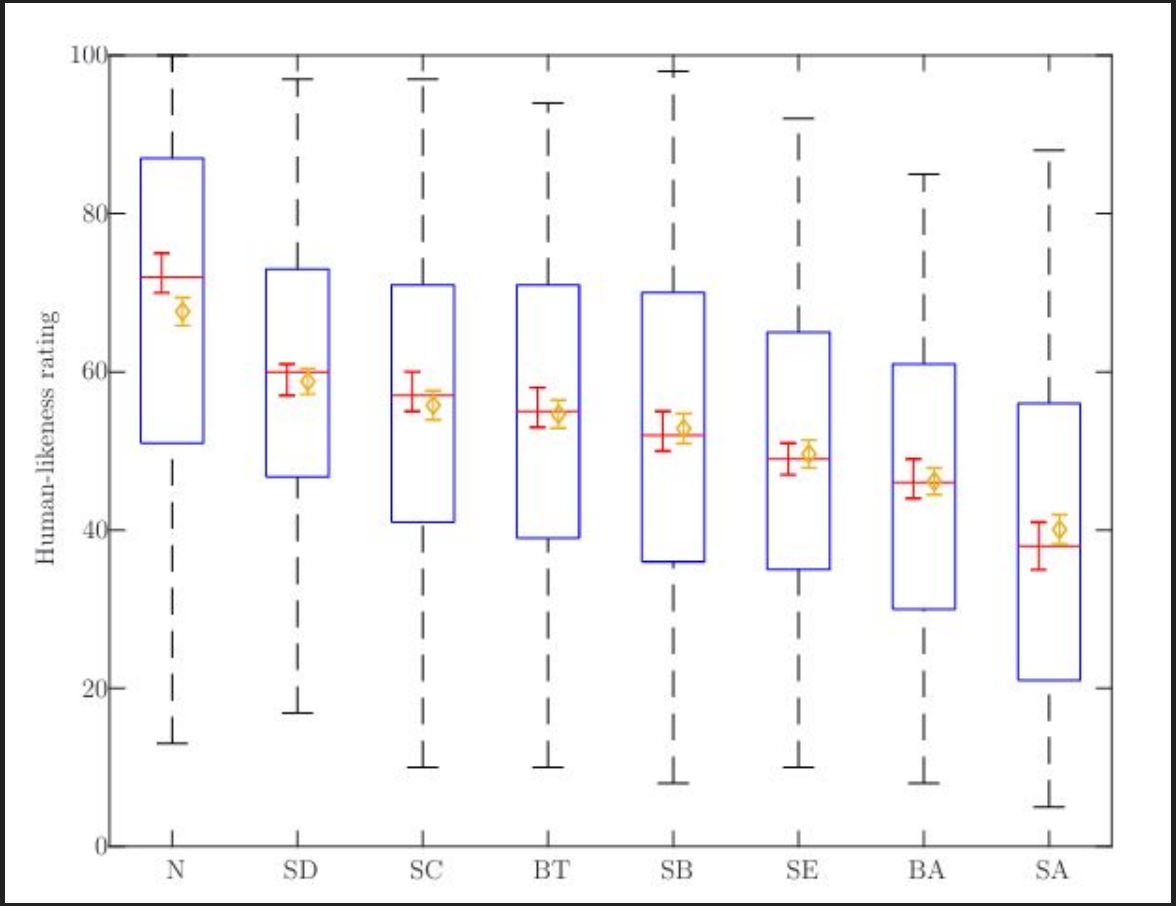
+5

SA - SE

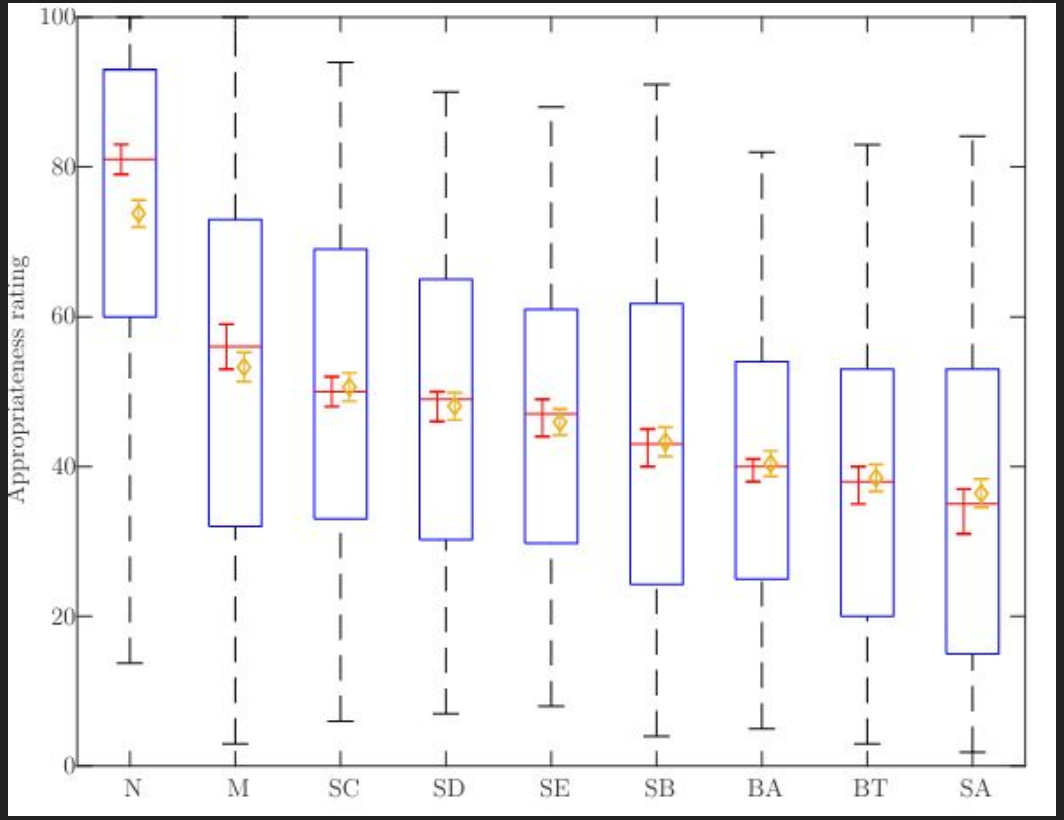
Results

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

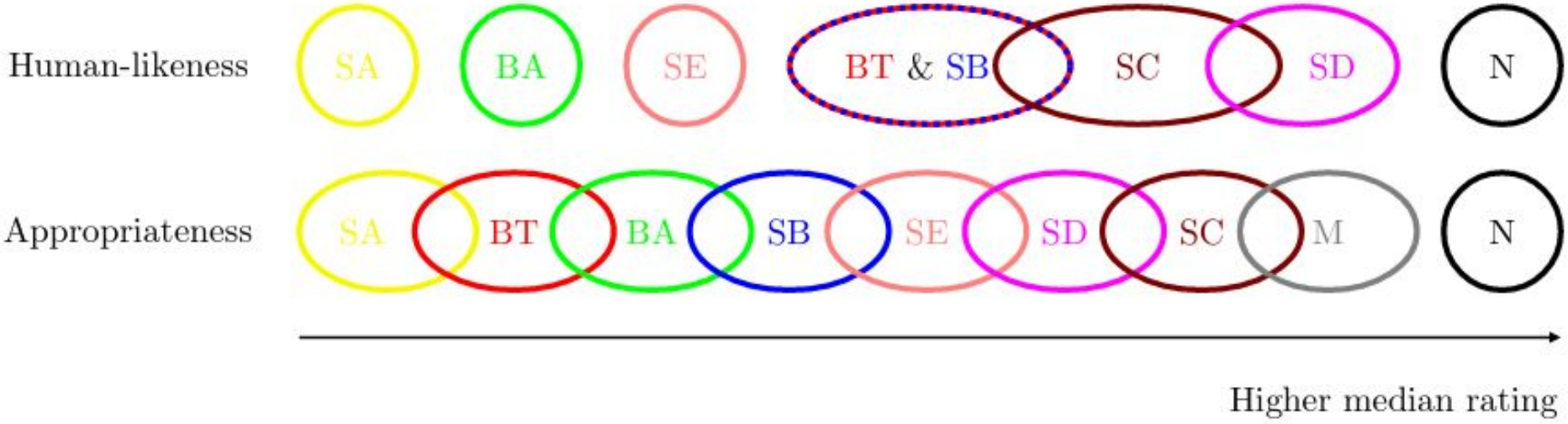


Human-likeness ratings



Appropriateness ratings

Partial ordering between systems



System Labels

SA -> Edinburgh CVGU

SB -> AlltheSmooth

SC -> StyleGestures

SD -> FineMotion

SE -> NecTec

Questions? Ask them through text!

How did they do it?