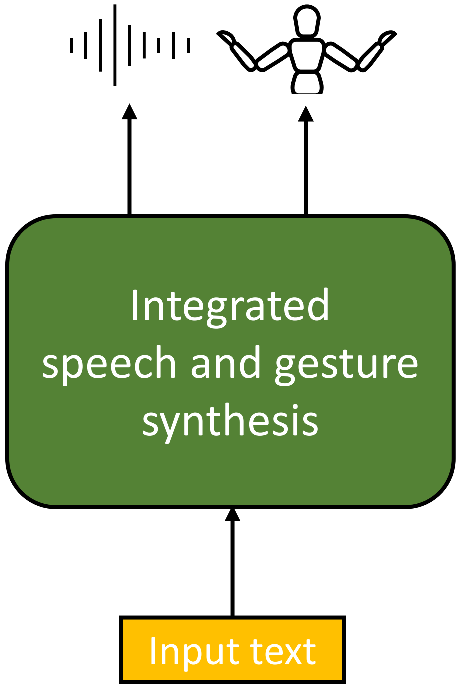
Integrated Speech and Gesture Synthesis

Siyang Wang, Simon Alexanderson, Joakim Gustafson,
Jonas Beskow, Gustav Eje Henter, Éva Székely
KTH Royal Institute of Technology, Stockholm

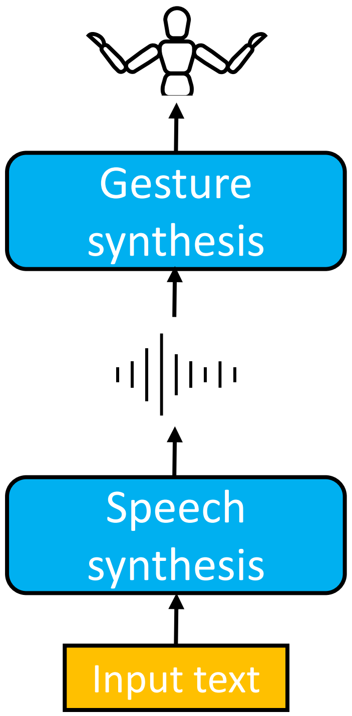


Current: speech and gesture synthesis

- Many applications need both speech and gesture
  - Embodied Conversational Agents (ECA), social robots
- Current solution: a simple pipeline of TTS and co-speech gesture synthesis
  - TTS: text to speech
  - Co-speech gesture synthesis: speech audio (sometimes with additional inputs) to gesture

Pipeline: TTS → co-speech gesture synthesis

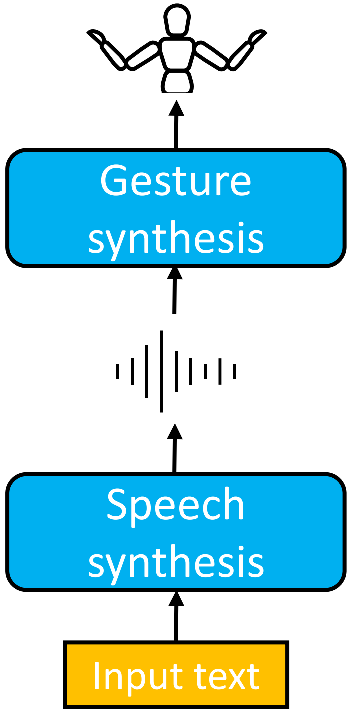
- TTS first → then co-speech gesture synthesis
- Advantages:
  - Two modules are developed and trained separately
  - Improvement in one (generally) improves whole system



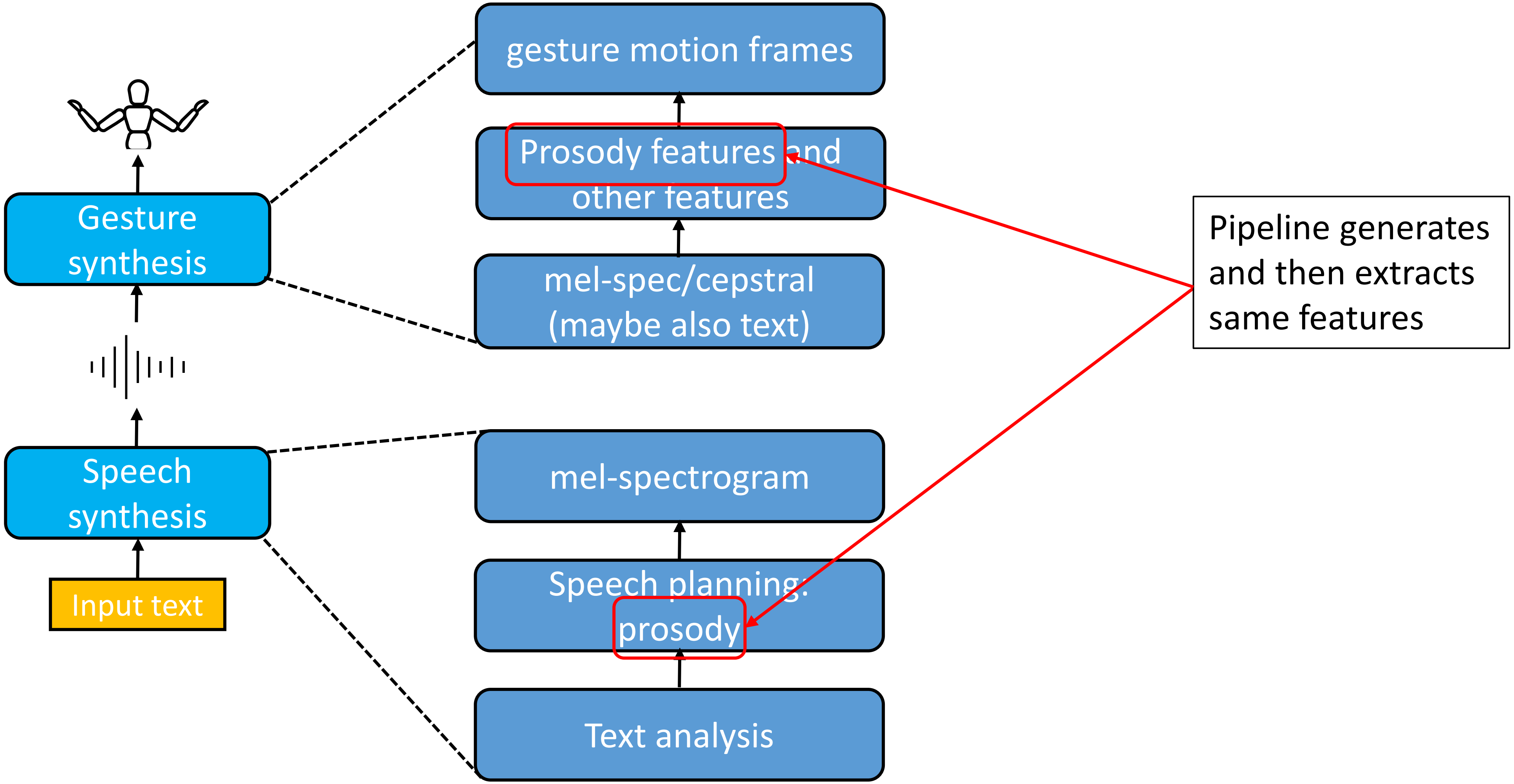
Pipeline: TTS → co-speech gesture synthesis

- Disadvantages:
  1. Domain mismatch
     - synthesized speech that is fed into gesture synthesis is not of GT quality
     - TTS may be trained on read speech while co-speech gesture is often trained on spontaneous corpus (no gesture when reading audio books)
     - Makes generated gesture worse
  2. Modeling inefficiency

     Pipeline systems generate prosody features in speech synthesis and then extract same features in gesture synthesis



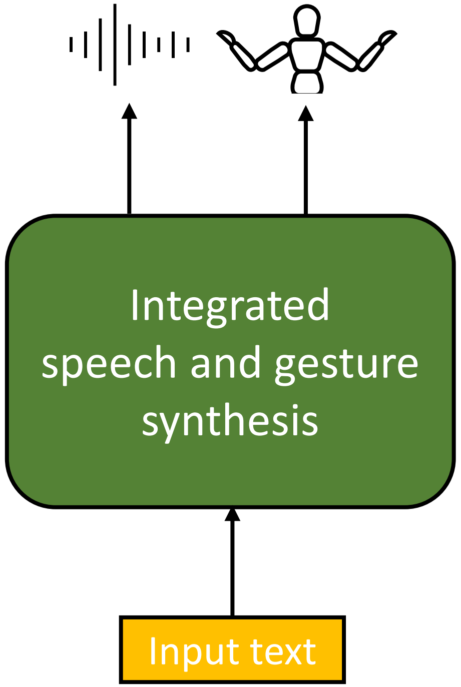
Modeling inefficiency in pipeline system



Pipeline generates and then extracts same features

Integrated Speech and Gesture Synthesis (ISG)

- A single, integrated model that generates speech AND gesture
- Alternative to the pipeline system that avoids the latter's disadvantages
- To the best of our knowledge, we are the first to study this problem



Proposed ISG models

Modify representative state-of-the-art TTS models

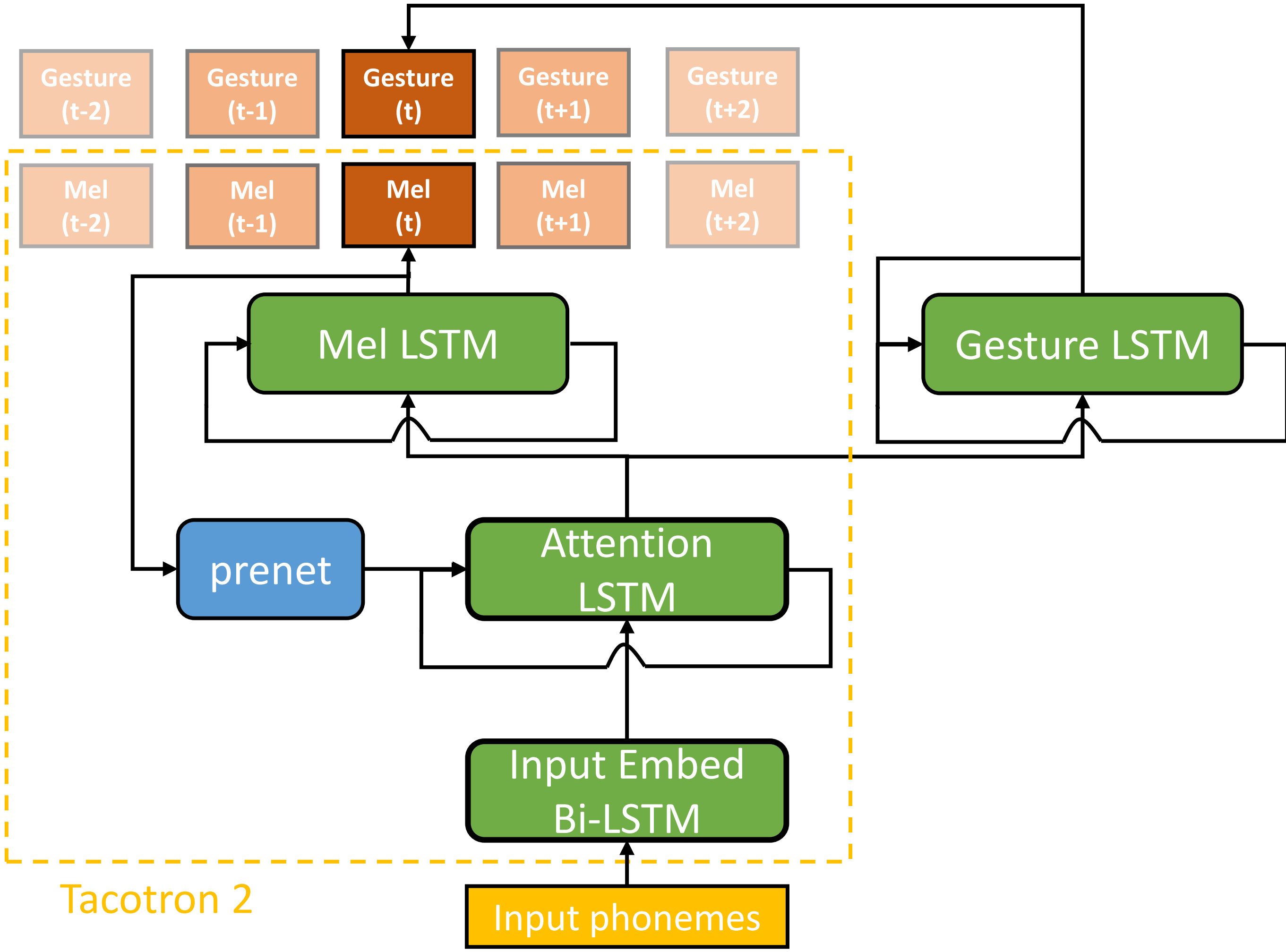
1. Tacotron 2: auto-regressive, non-probabilistic.
2. GlowTTS: parallel, probabilistic.

Reference:

1. Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, et al. 2018. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In Proc. ICASSP. 4779–4783.
2. Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. 2020. Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search. In Proc. NeurIPS. 8067–8077.

Proposed model: Tacotron2-ISG

- Modified Tacotron 2
  1. Not changing the original TTS architecture: early experiments show small changes make speech worse
  2. Using intermediate representation with speech planning information to generate gesture



Proposed model: Tacotron2-ISG

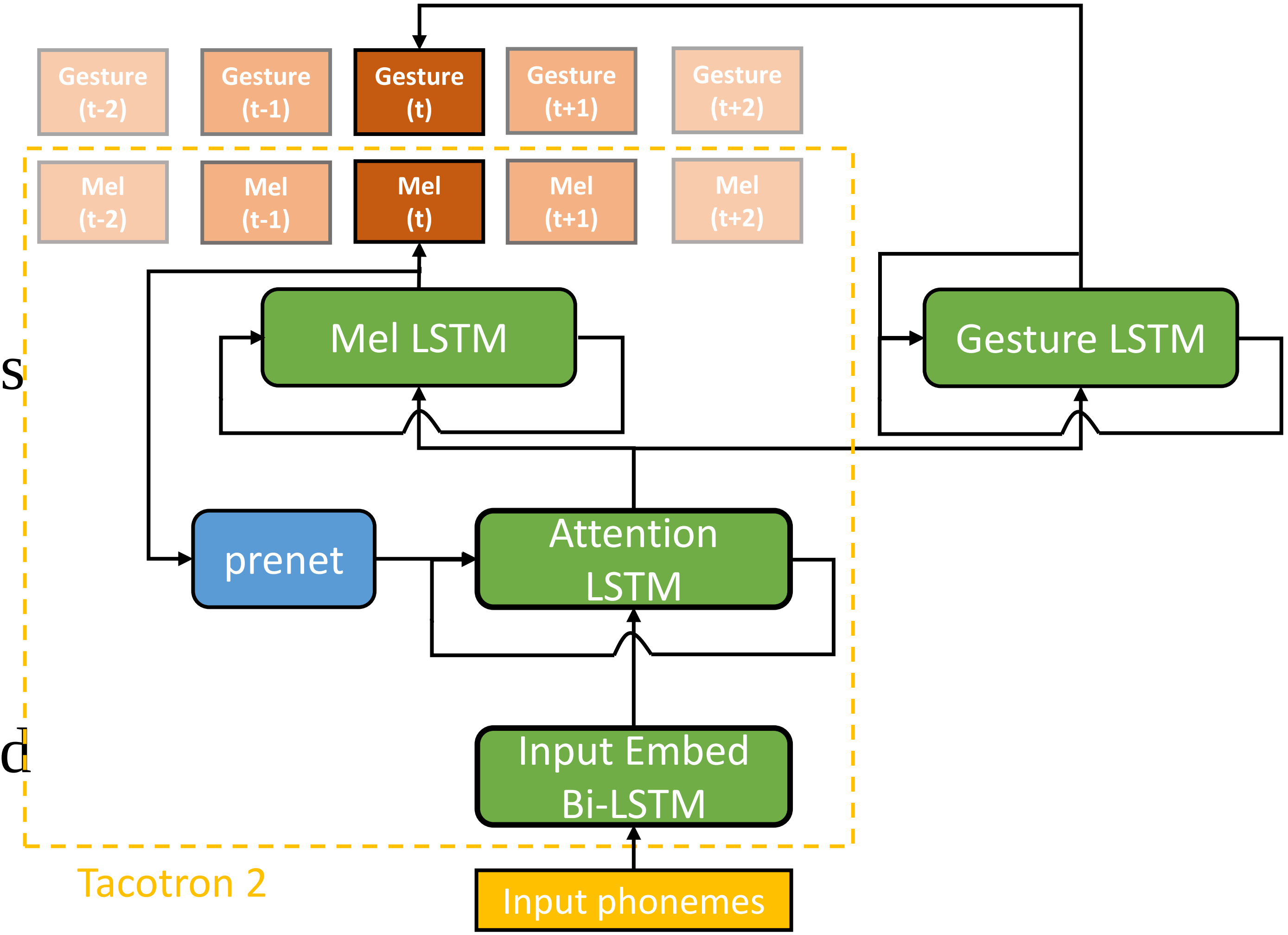
Training setup:

Step 0: Read-speech-pretrained

Step 1: Speech-only training

Step 2: ISG training in two different ways (compared in evaluation)

- Train speech and gesture sub-networks together with MSE: CT-Tacotron2-ISG
- Freeze speech sub-network, and train gesture sub-network with both MSE and speech-gesture GAN



Proposed Model: GlowTTS-ISG

- Expand the normalizing flow input from speech-only (GlowTTS) to also include gesture dimensions
- Hard to use intermediate representations from GlowTTS to generate gestures due to its layer-wise representation being entangled
- Same training setup as GlowTTS

Baseline: pipeline from Alexanderson et al. (IVA 2020)

- TTS: Tacotron2 pretrained on LJSpeech and finetuned on Trinity Speech-gesture Dataset
- Gesture generation: StyleGestures trained on Trinity Speech-gesture Dataset

Reference:

1. Simon Alexanderson, Éva Székely, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Generating coherent spontaneous speech and gesture from text. In Proc. IVA. 1–3
2. 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.

Data

Trinity Speech-Gesture Dataset

- 25 impromptu monologues with both speech and gesture recorded (~10 min each)
- Transcribed and manually corrected
- Segmented into ≤ 12s utterances for TTS-compatible training (using breathgroup bigram method)

Reference:

1. Ylva. Ferstl and Rachel. McDonnell. 2018. Investigating the use of recurrent motion modelling for speech gesture generation. In Proc. IVA. 93–98. https: //trinityspeechgesture.scss.tcd.ie
2. Éva Székely, Gustav Eje Henter, and Joakim Gustafson. 2019. Casting to corpus: Segmenting and selecting spontaneous dialogue for TTS with a CNN-LSTM speaker-dependent breath detector. In Proc. ICASSP. 6925–6929.

Test inputs

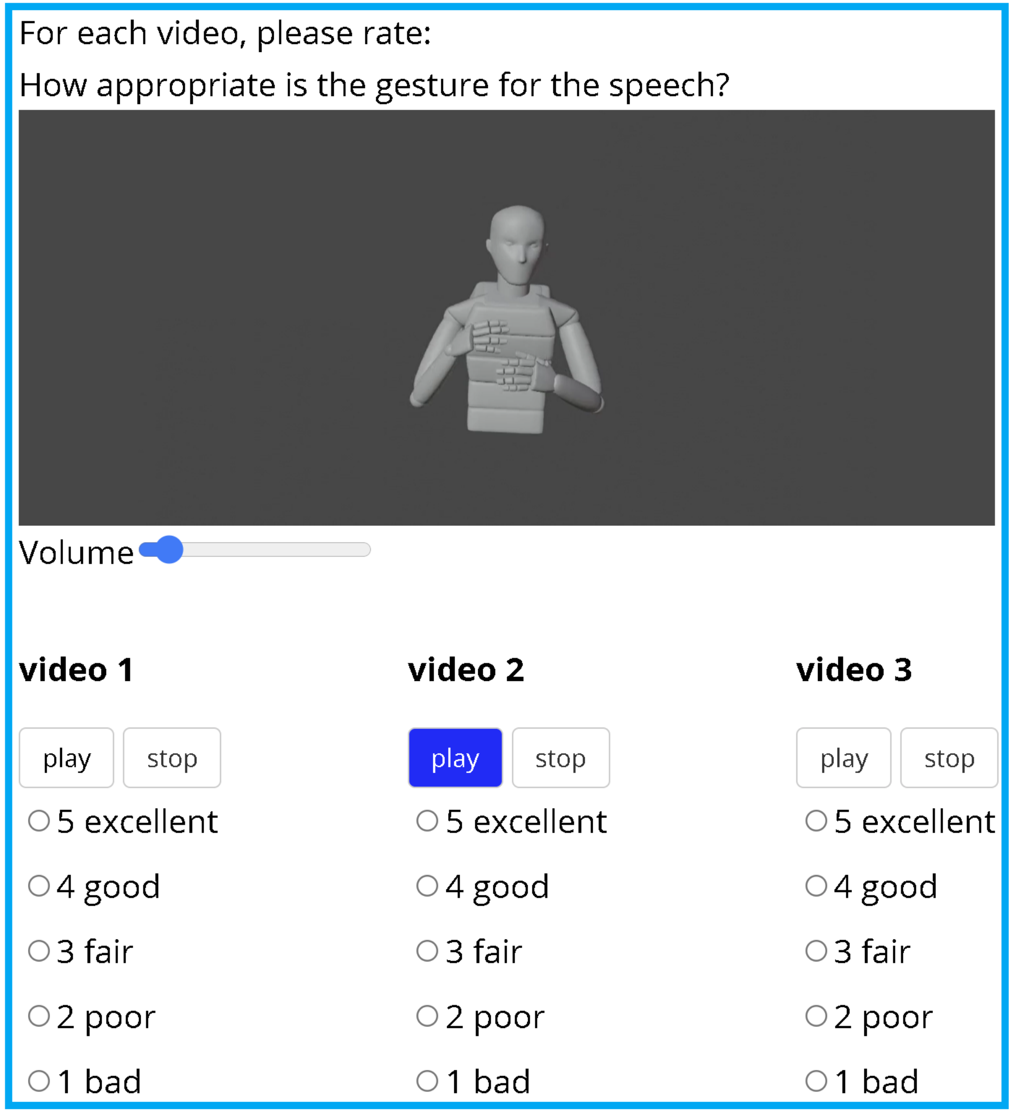
- The utterances from training dataset are largely incoherent and lack clear sentence structure
- Solution: use generated prompts from a GPT-2 model fine-tuned on the training corpus
  - Manually selected 17 that are coherent and relatively long
  - Longer inputs: distinguish models more

Evaluation: uni-modal and bi-modal

- Must evaluate speech and gesture together (bi-modal)
    - But what if a model excels in gesture which increases its bi-modal score despite generating much worse speech?
- We also evaluate speech and gesture separately (uni-modal)
- Overall we made 3 evaluations:
    - Speech-and-gesture
    - Gesture-only
    - Speech-only

Evaluation: MUSHRA-like interface

- MUSHRA
  - Widely used in speech eval
  - Increasing usage in gesture eval



For each video, please rate:

How appropriate is the gesture for the speech?

video 1

○ 5 excellent

○ 4 good

○ 3 fair

○ 2 poor

○ 1 bad

video 2

○ 5 excellent

○ 4 good

○ 3 fair

○ 2 poor

○ 1 bad

video 3

○ 5 excellent

○ 4 good

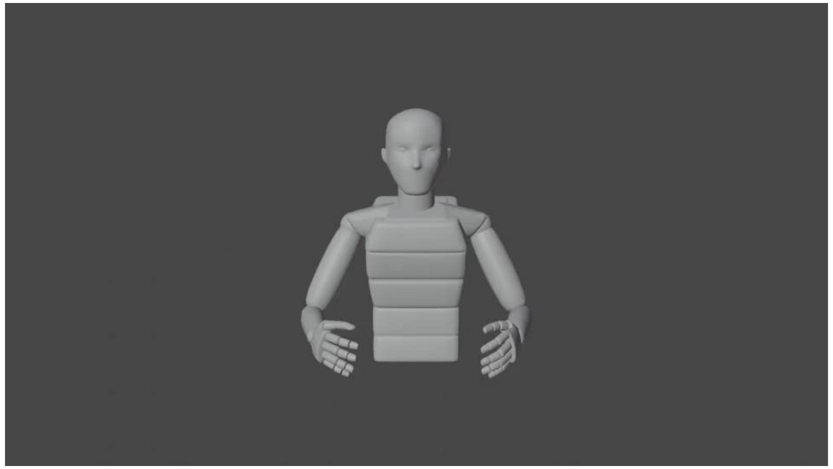
○ 3 fair

○ 2 poor

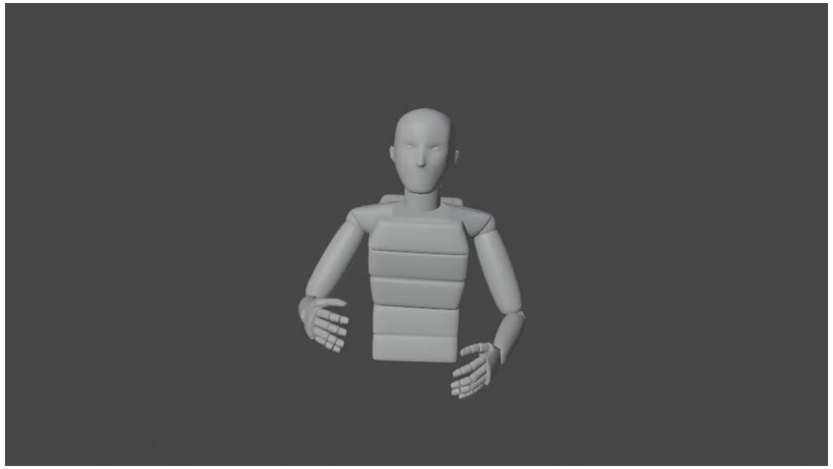
○ 1 bad

Reference:
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-RECBS.1534-3-201510-I

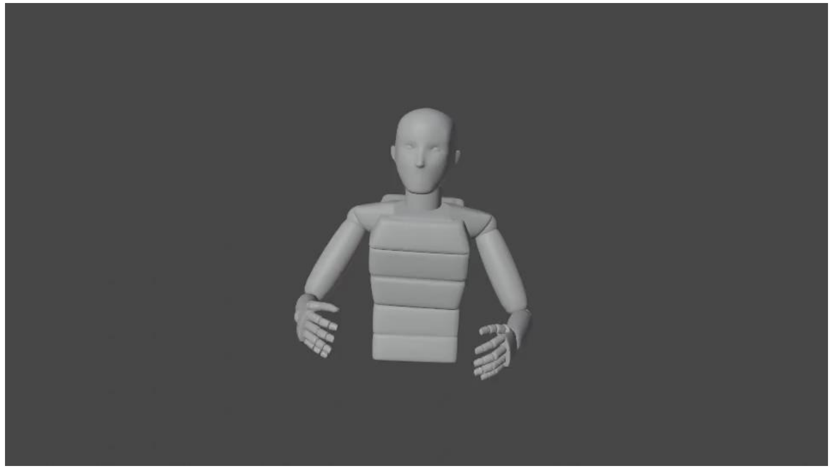
Video samples (input sentence 1)



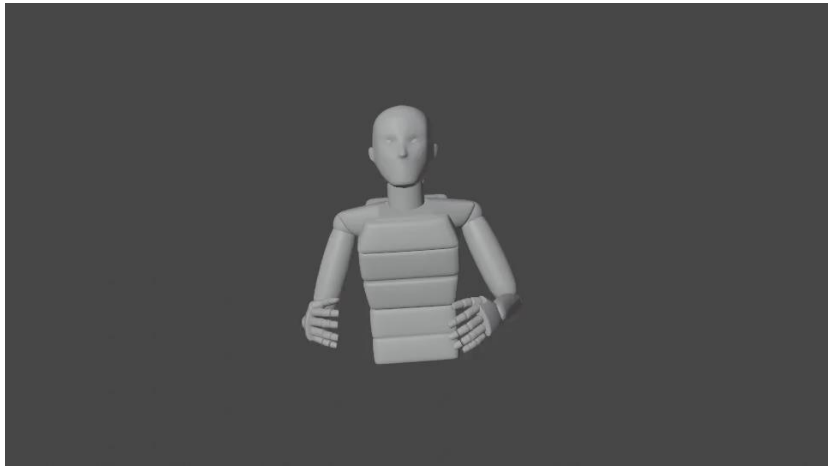
Pipeline



ST-Tacotron2-ISG

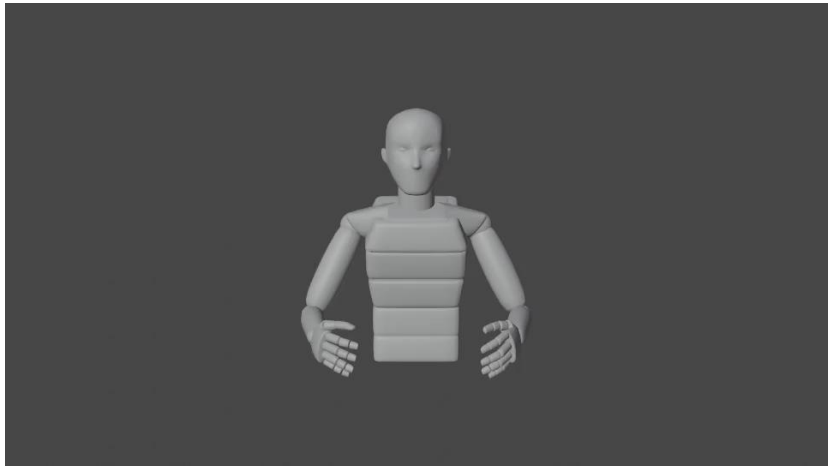


CT-Tacotron2-ISG

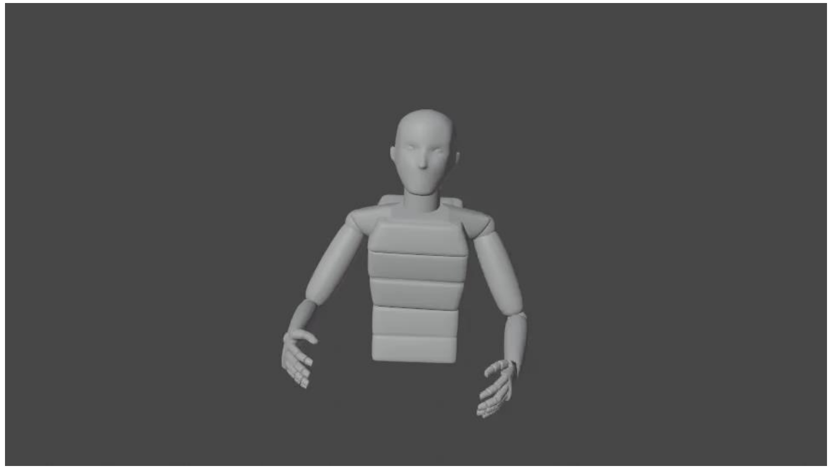


GlowTTS-ISG

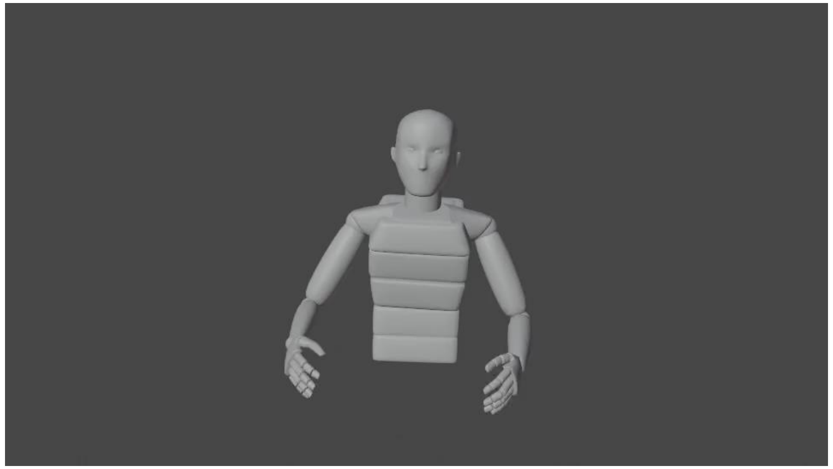
Video samples (input sentence 2)



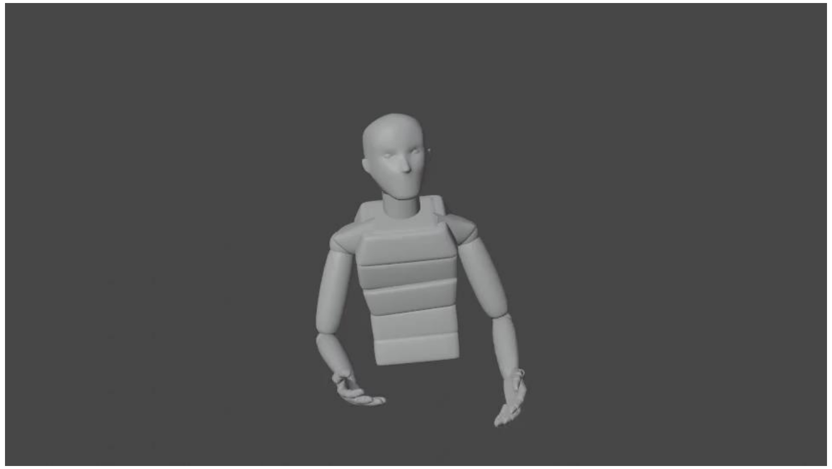
Pipeline



ST-Tacotron2-ISG



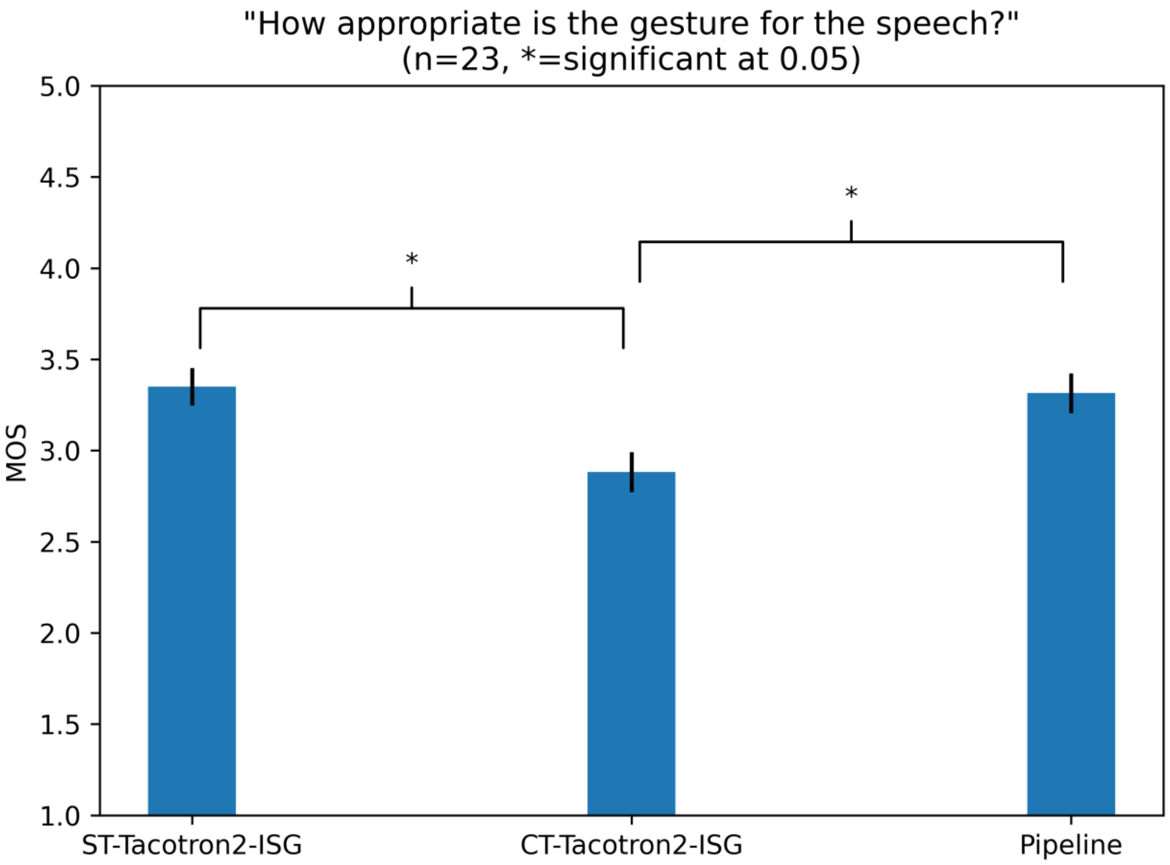
CT-Tacotron2-ISG



GlowTTS-ISG

Speech-and-gesture eval results

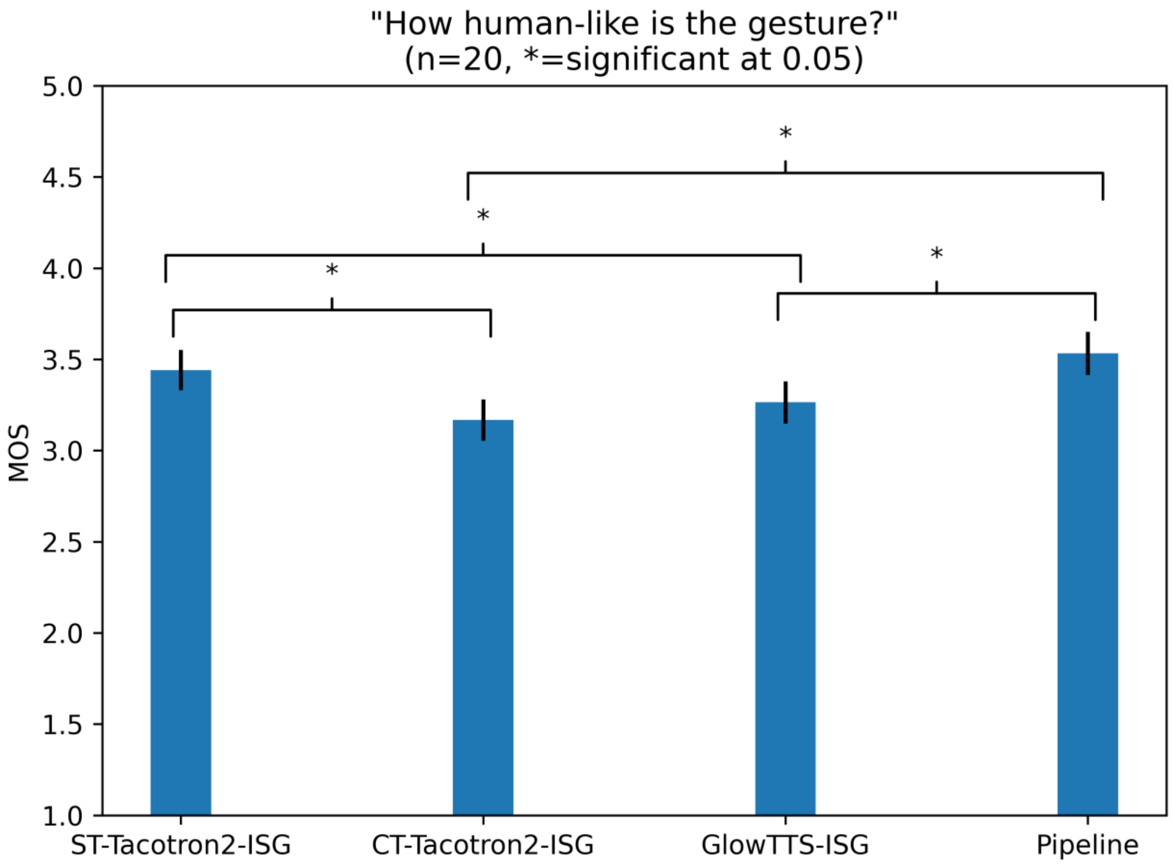
- Question is taken from GENEA Challenge 2020 (IVA 2021)
- ST-Tacotron2-ISG obtains highest MOS (not apparent from figure), but not significantly better than Pipeline system
- GlowTTS-ISG not evaluated due to poor speech quality



Reference:
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 Proc. IUI. 11–21.

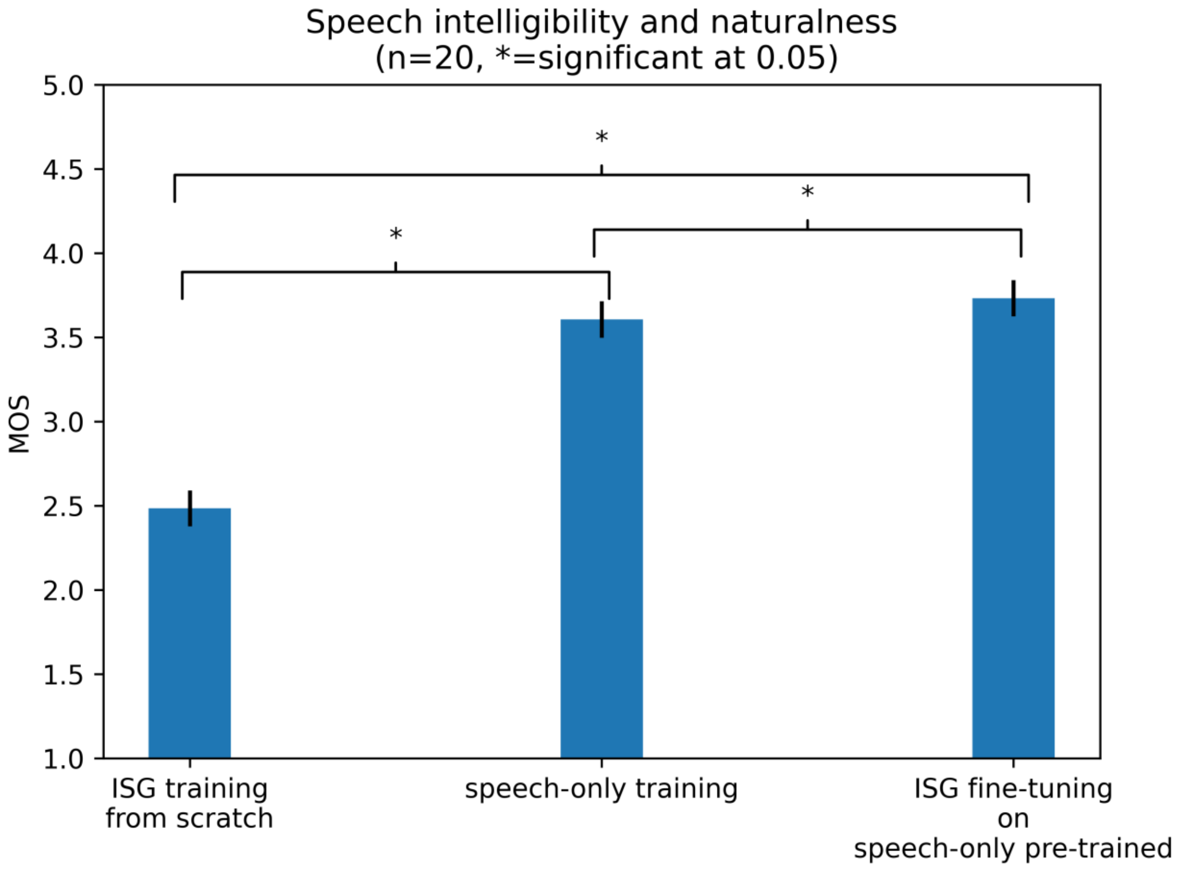
Gesture-only eval results

- Question is taken from GENEA Challenge 2020 (IVA 2021)
- Same videos from speech-and-gesture eval with audio turned off
- Pipeline obtains highest MOS (not apparent from figure), but not significantly better than second best ST-Tacotron2-ISG
- StyleGestures (gesture module in Pipeline) generates more dynamic and detailed gestures than ISG models



Speech-only eval results

- Speech audio only
- ISG fine-tuning (CT-Tacotron2-ISG) obtains highest MOS
- Shows full ISG training (both speech and gesture MSE loss) does not hurt speech quality
    - ISG training can effectively leverage uni-modal dataset along with bi-modal dataset
- ISG training from scratch is not enough to synthesize high quality speech
    - Speech-only pretraining is currently needed



ISG: same quality and faster than pipeline

- ST-Tacotron2-ISG obtains same-level of synthesis quality as Pipeline in all three evaluations (2 uni-modal, 1 bi-modal)
- Tacotron2-ISG is more parameter-efficient and faster

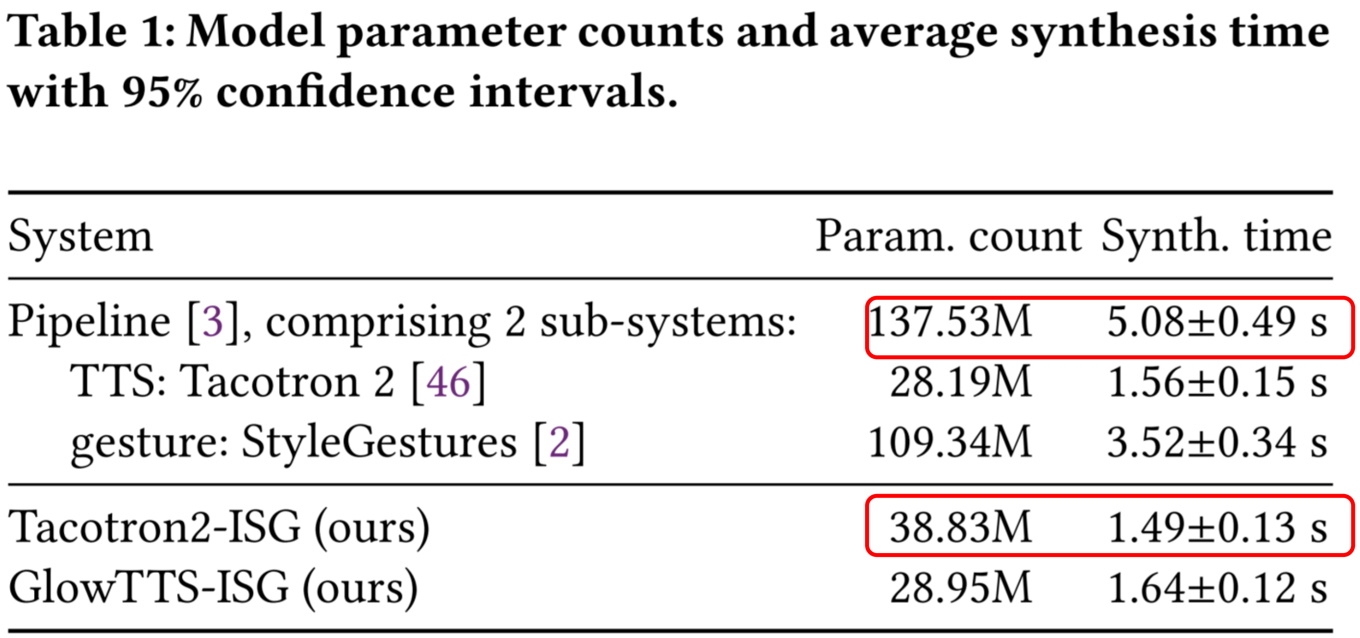


Table 1: Model parameter counts and average synthesis time with 95% confidence intervals.

System	Param. count	Synth. time
Pipeline [3], comprising 2 sub-systems:	137.53M	5.08±0.49 s
TTS: Tacotron 2 [46]	28.19M	1.56±0.15 s
gesture: StyleGestures [2]	109.34M	3.52±0.34 s
Tacotron2-ISG (ours)	38.83M	1.49±0.13 s
GlowTTS-ISG (ours)	28.95M	1.64±0.12 s

Discussion

- CT-Tacotron2-ISG not as good as ST-Tacotron2-ISG could be due to
    - Speech-gesture GAN (ST-Tacotron2-ISG)
    - Co-training speech and gesture need to rebalance the loss of the two modalities
- Tacotron 2 attention layer representation is better than mel-spec
    - Trained an ablation model with mel-spec and it synthesizes worse gesture
- GlowTTS-ISG does not give comparable results
    - Dataset too small for normalizing flow models to work well

Limitations

- Speech-and-gesture datasets are more difficult to get than uni-modal datasets
- Other TTS models might be better for ISG than the two we tried
- Evaluation for both speech and gesture synthesis remains challenging in general

website (code and video examples):
https://swatsw.github.io/isg_icmi21/

Thank you!