The DiffuseStyleGesture+ entry to the GENEA Challenge 2023

Sicheng Yang[1,*], Haiwei Xue[1,*] Zhensong Zhang[3], Minglei Li[2], Zhiyong Wu[1,4], Xiaofei Wu[3], Songcen Xu[3], Zonghong Dai[2]

[1] Tsinghua Shenzhen International Graduate School, Tsinghua University, China [2] Huawei Cloud Computing Technologies Co., Ltd, China
[3] Huawei Noah's Ark Lab, China [4] The Chinese University of Hong Kong, Hong Kong SAR, China

1. Introduction

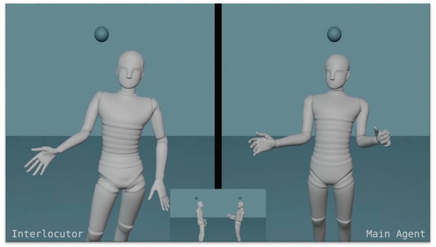
1.1 Motivation

- Goal:
  - ✓ Co-speech gesture is very important in daily communication.
- Challenge Setting:
  - Input:
    - Audio, text of the main agent
    - Audio, text and gesture of the interlocutor
  - Output:
    - Gesture of the main agent
- Problems :
  - GAN-based methods → Training difficulty
  - VAEs and Flows-based methods → Take diversity into account

1.2 Contribution

- ✓ Propose DiffuseStyleGesture+, a improved diffusion-based model for multimodal-driven co-speech gesture generation
- ✓ Our model is among the first tier at human-likeness, appropriateness for the interlocutor, and achieves competitive performance on appropriateness for speech.

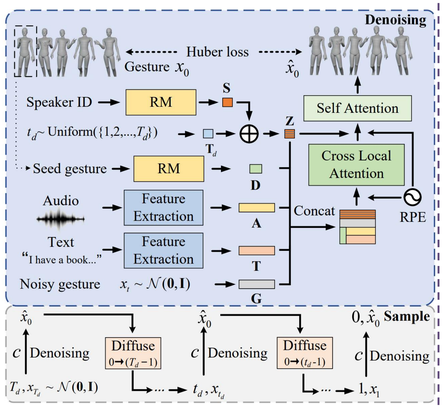
2. Visualization



3. Methodology

3.1 Feature Extraction

- Gesture
  - 62 joints including the fingers
  - Motion features: pos., vel., acc., rot., rot. vel., and rot. acc. of each joint
  - Denote natural mocap gestures clip as $X_0 \in \mathbb{R}^{(N_{seed}+N)\times[62\times(9+3)\times3]}$
- Audio
  - Audio features: MFCC, Mel Spectrum, Pitch, Energy, WavLM, and Onsets
  - Denote the features of audio clip as $\mathbf{A} \in \mathbb{R}^{N\times(40+64+2+2+1024+1)}$
- Text
  - Text features: FastText, one bit to indicate whether there is a laugh or not
  - Denote the features of text clip as $\mathbf{T} \in \mathbb{R}^{N\times302}$
- Speaker ID, one-hot vectors, 17 speakers, denote as $\mathbf{S} \in \mathbb{R}^{17}$



3.2 Gesture Denoising

- Reconstruct the original gesture x_0 from the pure noise x_t, noising step t_d and conditions c, $c = [\mathbf{S}, \mathbf{D}, \mathbf{A}, \mathbf{T}]$. *This is given by* $\hat{x}_0 = \mathrm{Denoise}(x_{t_d}, t_d, c)$.
- Loss function $\mathcal{L} = E_{x_0\sim q(x_0|c),t_d\sim[1,T_d]}[\mathrm{HuberLoss}(x_0 - \hat{x}_0)]$

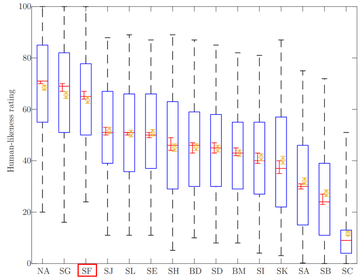
3.3 Gesture Sampling

- Initial noisy gesture x_T is sampled from the standard normal distribution; other x_{t_d}, $t_d < T_d$ is the result of the previous noising step
- Initial seed gesture is a gesture from the dataset; other clips is the last N_{seed} frames of the gesture generated in the previous clip
- For every clip, in every noising step t_d, we predict the clean gesture $\hat{x}_0$ and add Gaussian noise to the noising step x_{t_d-1}
- This process is repeated from $t_d = T_d$ until $\hat{x}_0$ is reached

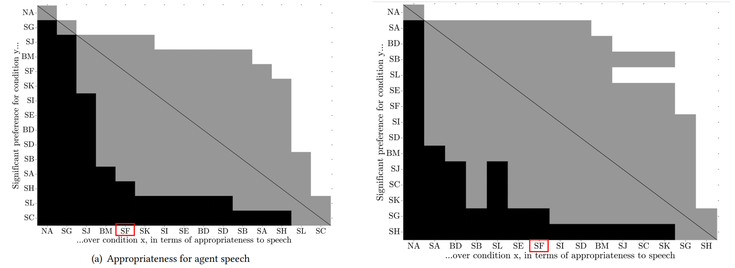
4. Experiments

4.1 Experiment Setting

- Cropped to 150 frames (5 seconds) N_{seed} = 30, N = 120
- Standard normalization
- Latent dimension 512, 8 heads, 48 attention channels, window size is 15 frames, 120000 steps, T_d = 1000



4.2 Evaluation Analysis



(a) Appropriateness for agent speech

- NA: natural mocap; BM (D): monadic (dyadic) baseline; S: submission
- Human-likeness: mean 63.6±1.3, median 65 ∈ [64, 67]
- Appropriateness for agent speech: MAS 0.20±0.06, Pref. matched 55.8%
- Appropriateness for the interlocutor: MAS 0.04±0.06, Pref. matched 51.5%

Reference

[1] The GENEA Challenge 2023: A large scale evaluation of gesture generation models in monadic and dyadic settings
[2] DiffuseStyleGesture: Stylized Audio-Driven Co-Speech Gesture Generation with Diffusion Models
[3] Talking With Hands 16.2M: A Large-Scale Dataset of Synchronized Body-Finger Motion and Audio for Conversational Motion Analysis and Synthesis
[4] The IVI Lab entry to the GENEA Challenge 2022--A Tacotron2 based method for co-speech gesture generation with locality-constraint attention mechanism

Project page