A one-to-many co-speech robot gesture synthesis

Chuang Yu and Adriana Tapus
ASR Robotics Lab, U2IS, ENSTA-Paris, IP Paris

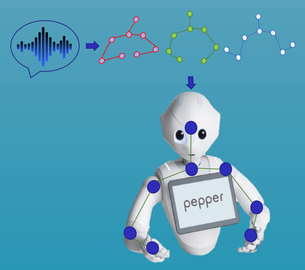
SCAN ME

Introduction

The human gestures occur spontaneously and usually they are aligned with speech, which leads to a natural and expressive interaction. Speech-driven gesture generation is important in order to enable a social robot to exhibit social cues and conduct a successful human-robot interaction. In this paper, the generation process involves mapping acoustic speech representation to the corresponding gestures for a humanoid robot. The paper proposes a new GAN (Generative Adversarial Network) architecture for speech to gesture generation. Instead of the fixed mapping from one speech to one gesture pattern, our end-to-end GAN structure can generate multiple mapped gestures patterns from one speech (with multiple noises) just like humans do. The generated gestures can be applied to social robots with arms. The evaluation result shows the effectiveness of our generative model for speech-driven robot gesture generation.

Purpose

- **Given one speech as input, the task is to generate multiple spontaneous gestures as output.**
- **Map the generated gestures to the control signal of robot joint motors for real human-robot interaction scene.**



Gesture generation model

- **Generator: Encoder + Decoder**
  - **--Encoder: Speech→Audio Presentation**
  - **--Decoder: Audio Presentation→ Gesture**
- **Discriminator**
  - **Speech+Gesture= match or not?**
- **Loss Function**

$$\mathcal{L}_{cGAN}(G,D) = \mathbb{E}_{s,g}[\log D(s,g)] + \mathbb{E}_{s,z}[\log(1 - D(s, G(s,z))]$$

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{s,z,g}\left[\|g - G(s,z)\|_1\right]$$

$$G^*_{S2GGAN} = \arg\min_G \max_D \mathcal{L}_{cGAN}(G,D) + \lambda \mathcal{L}_{L1}(G)$$

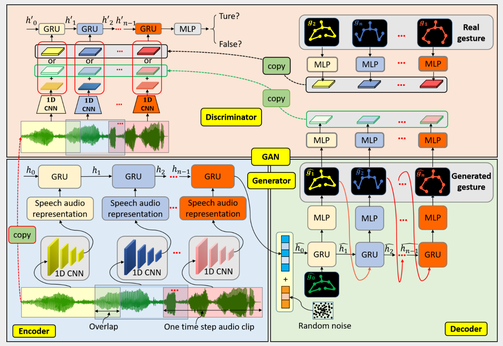


Figure 1: Gesture synthesis model with GAN

Human2Robot Gesture retargeting

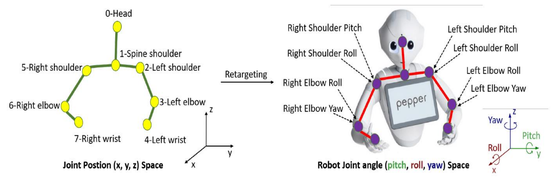
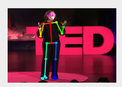
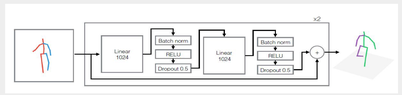


Figure 2: An overview of the gesture retargeting process.

Database building



OpenPose



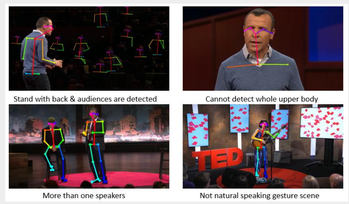
"3D-pose-baseline" model

Audio-visual Database building steps:

(1) Download 1760 public TED videos from YouTube through YouTube Data API.
(2) Extract the speech audios from videos through FFmpeg library.
(3) Extract 2D gesture data of upper body through OpenPose library.
(4) Transform the 2D gesture to the 3D gesture by "3D pose baseline" model trained by myself.
(5) Cut into clips with PySceneDetect, a Python library for detecting scene changes in videos.
(6) Select useful clips to build the audio-visual database for SRG[3].

Our gesture clip selection rule as follows:

- **The all eight upper body joints are detected in all frames.**
- **The speaker should not stand with back. [Youngwoo 2019].**
- **More than 5 seconds.**
- **Only one speaker.**
- **No audiences are detected.**
- **natural speaking gestures.**
- **no still frames where the speaker stays still without the gesture movements.**

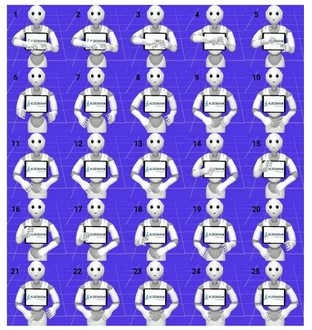


OpenPose examples with wrong results

OpenPose examples with suitable results

Results

1. Qualitative evaluation



Generated gestures on Pepper robot.
https://youtu.be/G8zoVf_Pedo

2. Quantitative validation

$$APE = \frac{1}{M \times T} \sum_{m=1}^{M} \sum_{t=T}^{T} \left| xyz_{\text{real}}(m,t) - xyz_{\text{generated}}(m,t) \right|$$

Estimate the generated pose using a **Average Position Error (APE).** Where,

- T is the time steps and is equal to 126;
- M is the number of testing samples and is equal to 960 (30 batches with batch size 32);
- $xyz_{real}(m,t)$ and $xyz_{Generated}(m,t)$ are the the ground truth and prediction of joint position x/y/z of sample m at time step t, respectively.

TABLE I: APE with noise 1 and noise 2

APE(cm)	Noise 1	Noise 2
Head	5.99	5.89
Left shoulder	3.63	3.62
Left elbow	8.97	8.91
Left wrist	21.36	21.79
Right shoulder	3.63	3.63
Right elbow	10.79	10.73
Right Wrist	22.51	22.43

- The generation with noise 1 and the generation noise 2 had **similar** results, which certify that the random noise can make the generated gestures have a random variation to a certain extent.
- The **head APE** and the **shoulder APE** are **small** while **the elbow and wrist APE** are large. Because the elbow joint and wrist joint at the end of the arm have a large movement space and other joints movements have a limited space in real contexts.

Please see our paper for details

[Oral presentation paper] Chuang Yu, and Adriana Tapus. "SRG3: Speech-driven Robot Gesture Generation with GAN." ICARCV IEEE. 2020.