Novel Realizations of Speech-driven Head Movements with Generative Adversarial Networks

Motivation

Background:
- Conversational agents (CAs) created with rules display limited variations
- Strong relation between head motion and speech
  - Goal: Speech-driven head motion for CAs
- Speech-driven frameworks tend to generate head motion with limited range of movements

Our Work:
- Learn the conditional distribution of head movements given speech features
- Increase the range of synthesized movements
- Generate multiple novel realizations of head movements for an input speech signal

Resources

Corpus:
- The IEMOCAP corpus
  - 1st female subject (1h6m)
- Head:
  - Motion capture data
  - Three head angular rotations
- Audio:
  - F0 contour, and Intensity (plus first and second derivatives) + S/L

Xface

Rendering Toolkit:

Objective Evaluation
- Parzen window density estimator
- Each frame as one sample (103.7K)
- Baselines:
  - DBN
  - BLSTM-MSE
  - BLSTM-CC

Subjective Evaluation
- Amazon mechanical turk (AMT)
  - 5 15s-segments per model (15 videos)
  - 12 workers (4 evaluations per comparison)
- Proportion preference of cGAN over baselines:
  - 0.542 (BLSTM-CC)
  - 0.737 (BLSTM-MSE)
  - 0.682 (DBN)

Conclusions
- cGAN models the intrinsic random properties of beat gestures
- cGAN generates samples that better fit the distribution of the data compared with the three baselines
- Subjective evaluations showed higher average preferences for cGAN compared with BLSTM-MSE
- We can generate as many sequences as we need

Future Work
- Considering the interlocutor may provide more predictive features for head pose generation when the CA is listening
- The model can be applied to learn facial movements during conversations

This work was funded by NSF (IIS 1718944)