Animation Generation with Speech Signal

Presented by Gusi Te and Jiajun Su

https://github.com/tegusi/animation-generation

Table of contents

Background

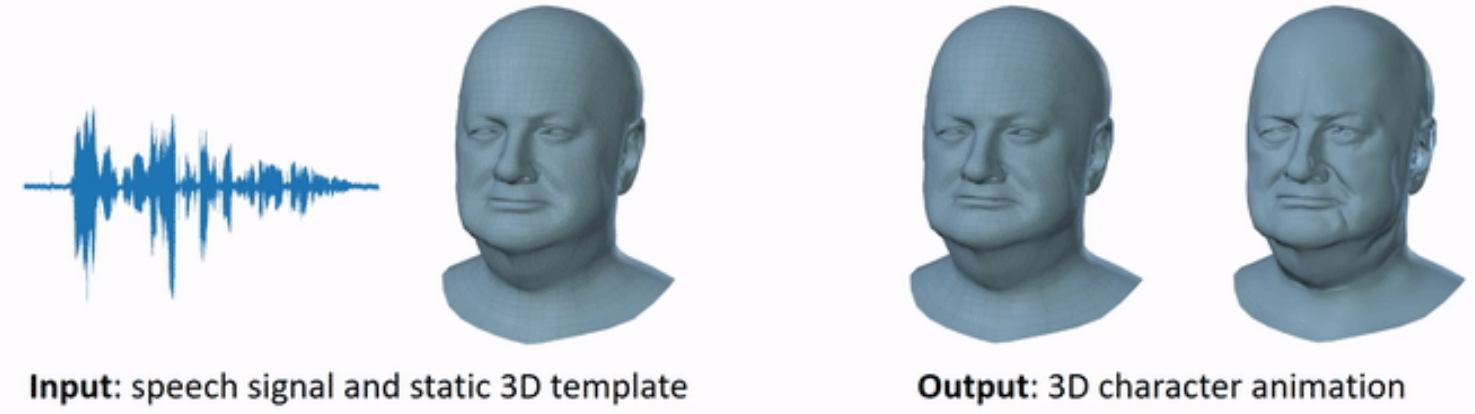
- 3D presentation provides more abundant information than 2D one
- Simply provide audio to generate vivid facial animation is challenging
- Applicable to game, virtual reality and so on

Background

- Speech and facial motion lie in different spaces
- Many-to-many mapping between phonemes and facial motion
- Limited training data relating speech to the 3D face shape
- Speaker independent representation

Contribution

- Disentangle face motion from subject identity
- Model speech signal with deep speech
- Realistically animates a wide range of face templates



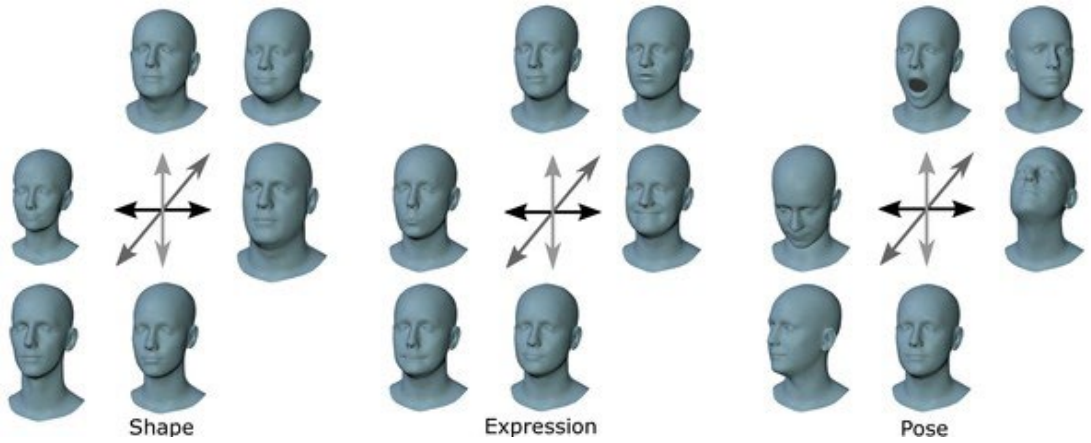
Input: speech signal and static 3D template

Output: 3D character animation

Methods

1. FLAME – 3D facial representation

- Use linear transformations to describe
  - Shape
  - Expression
  - Pose
- Linear blend skinning

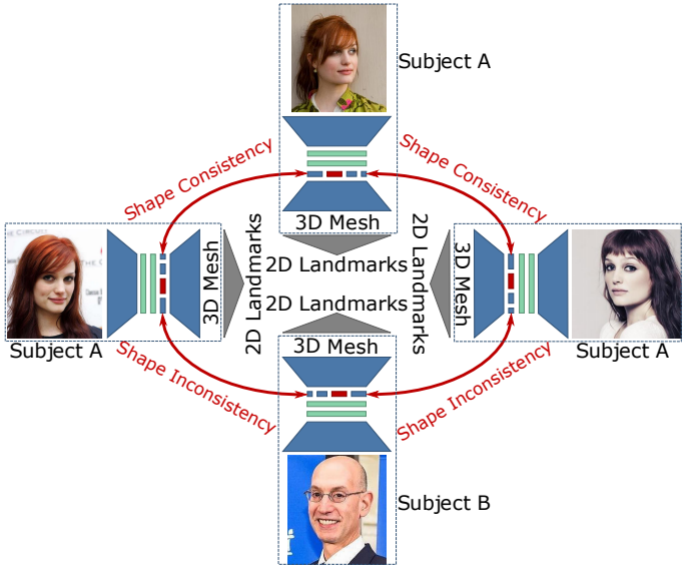


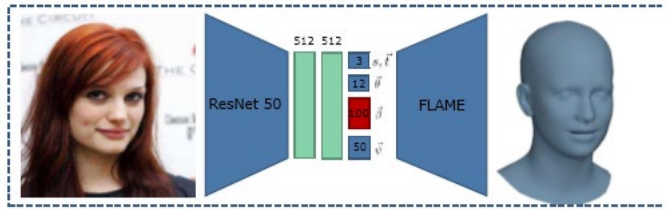
Learning a model of facial shape and expression from 4D scans (SIGGRAPH ASIA 2017)

Methods

2. RingNet – Estimate 3D mesh from single image

- Takes multiple images of the same person and single image of another person
- Propose the Ring loss to keep shape consistency





$$\sum_{i=1}^{n_b} \sum_{j,k=1}^{R-1} \max(0, \left\|\vec{\beta}_{ij} - \vec{\beta}_{ik}\right\|_2^2 - \left\|\vec{\beta}_{ij} - \vec{\beta}_{iR}\right\|_2^2 + \eta)$$

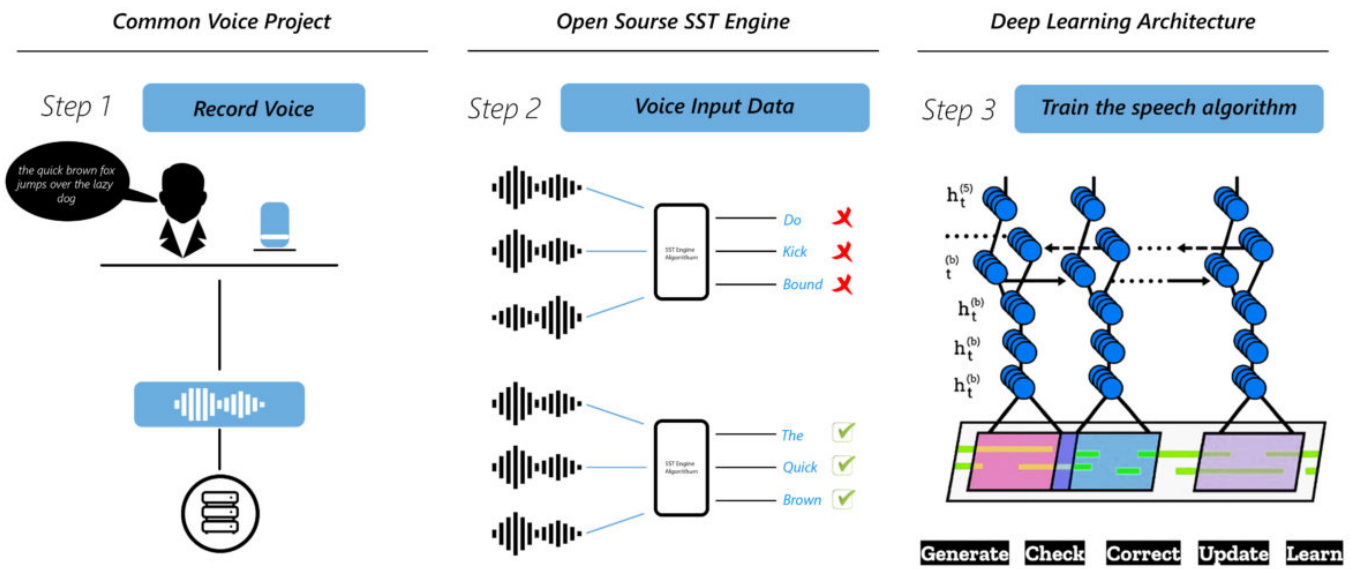
Ring loss

Learning to Regress 3D Face Shape and Expression from an Image without 3D Supervision (CVPR 2019)

Methods

3. DeepSpeech – voice model

- Provides more robust voices, regardless of noise, artifacts
- We simply adopt English alphabet to reduce complexity

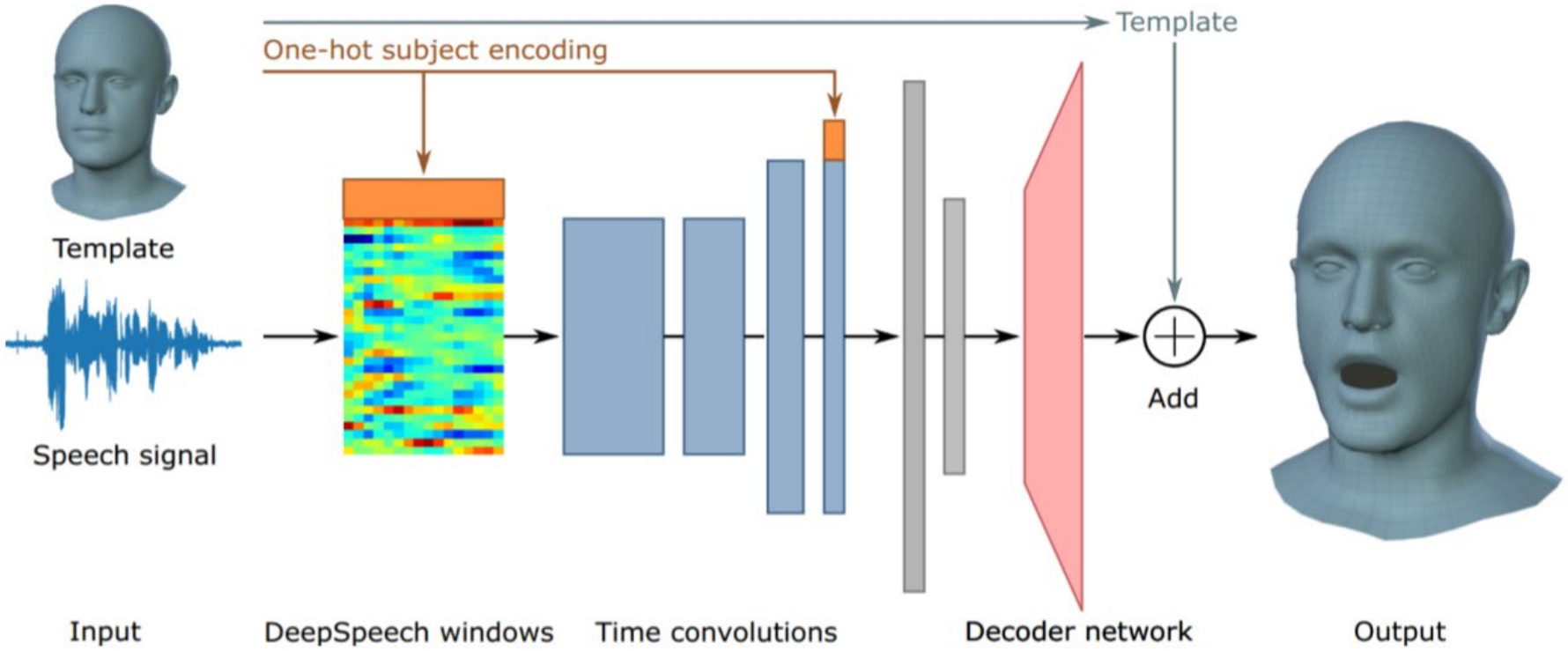


How a Speech Application Learns

Deep Speech: Scaling up end-to-end speech recognition

Methods

4. Subject-independent generation



Methods

4. Subject-independent generation

- **Speech feature extraction**
  - Window size=16
- **Encoder**
  - 4 Conv Layers + 2 FC layers
- **Decoder**
  - 3 FC Layers with PCA initialization

Methods

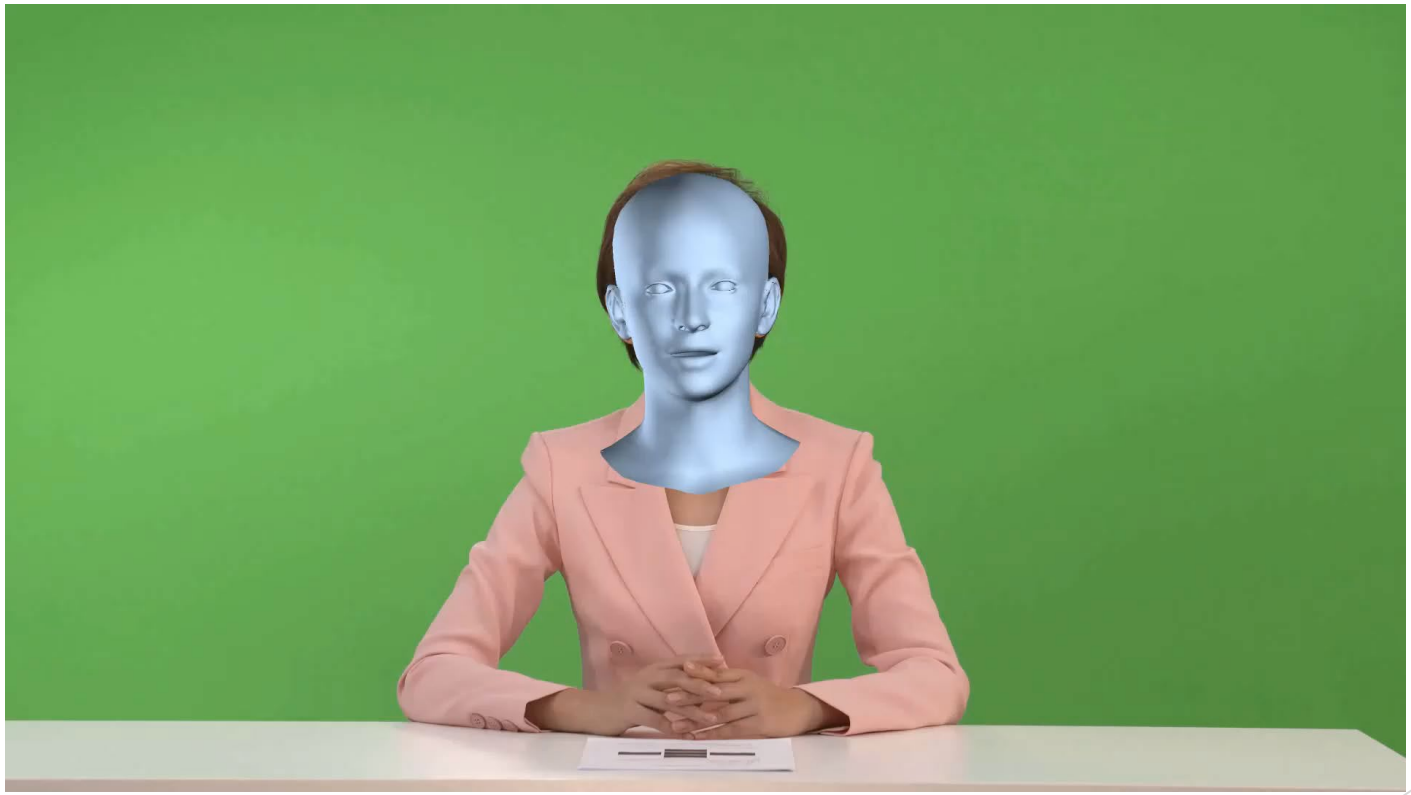
Loss function

- Position loss $E_p = \|\mathbf{y}_i - \mathbf{f}_i\|_F^2$
  - Computes the distance between the predicted outputs and the training vertices
  - Encourages the model to match the ground truth performance

- Velocity loss $E_v = \|(\mathbf{y}_i - \mathbf{y}_{i-1}) - (\mathbf{f}_i - \mathbf{f}_{i-1})\|_F^2$
  - Computes the distance between the differences of consecutive frames between predicted outputs and training vertices
  - Induces temporal stability

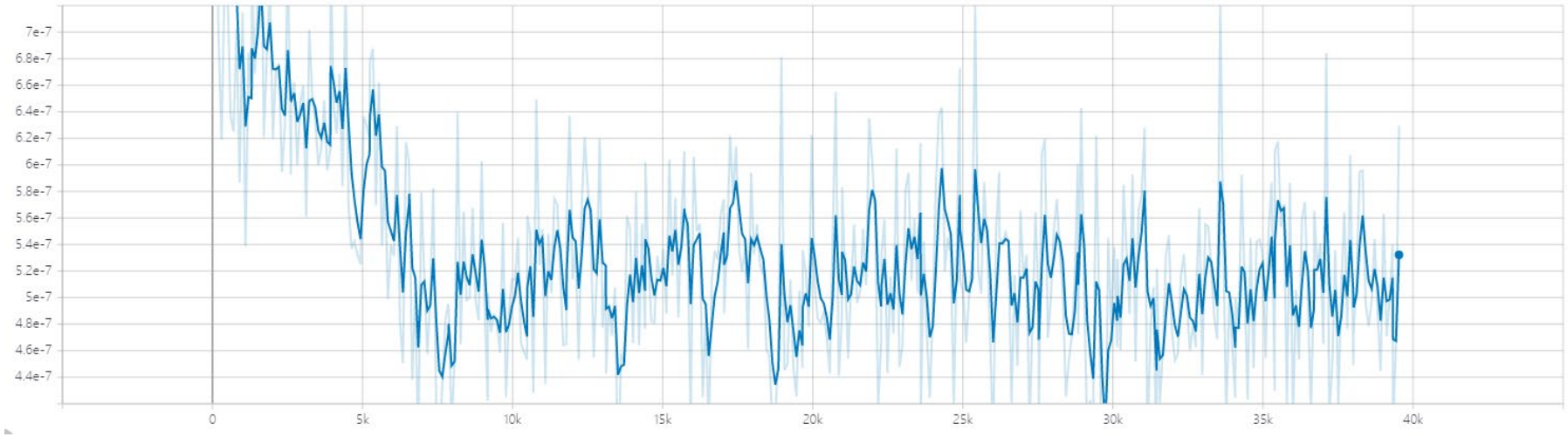
Experiments

RingNet Generation



Experiments

Validation Loss

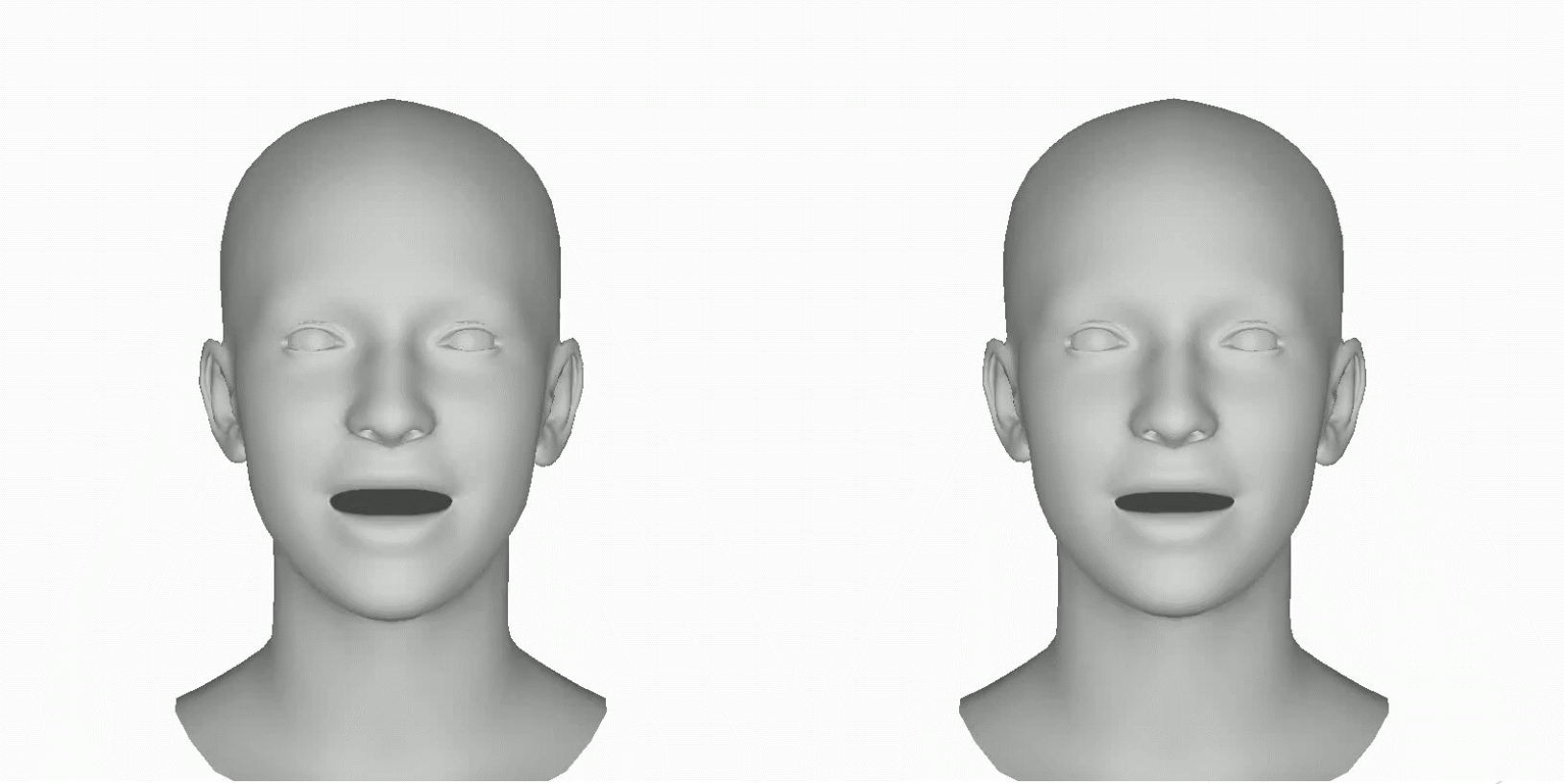


Experiments

Training Loss

Experiments

Chinese

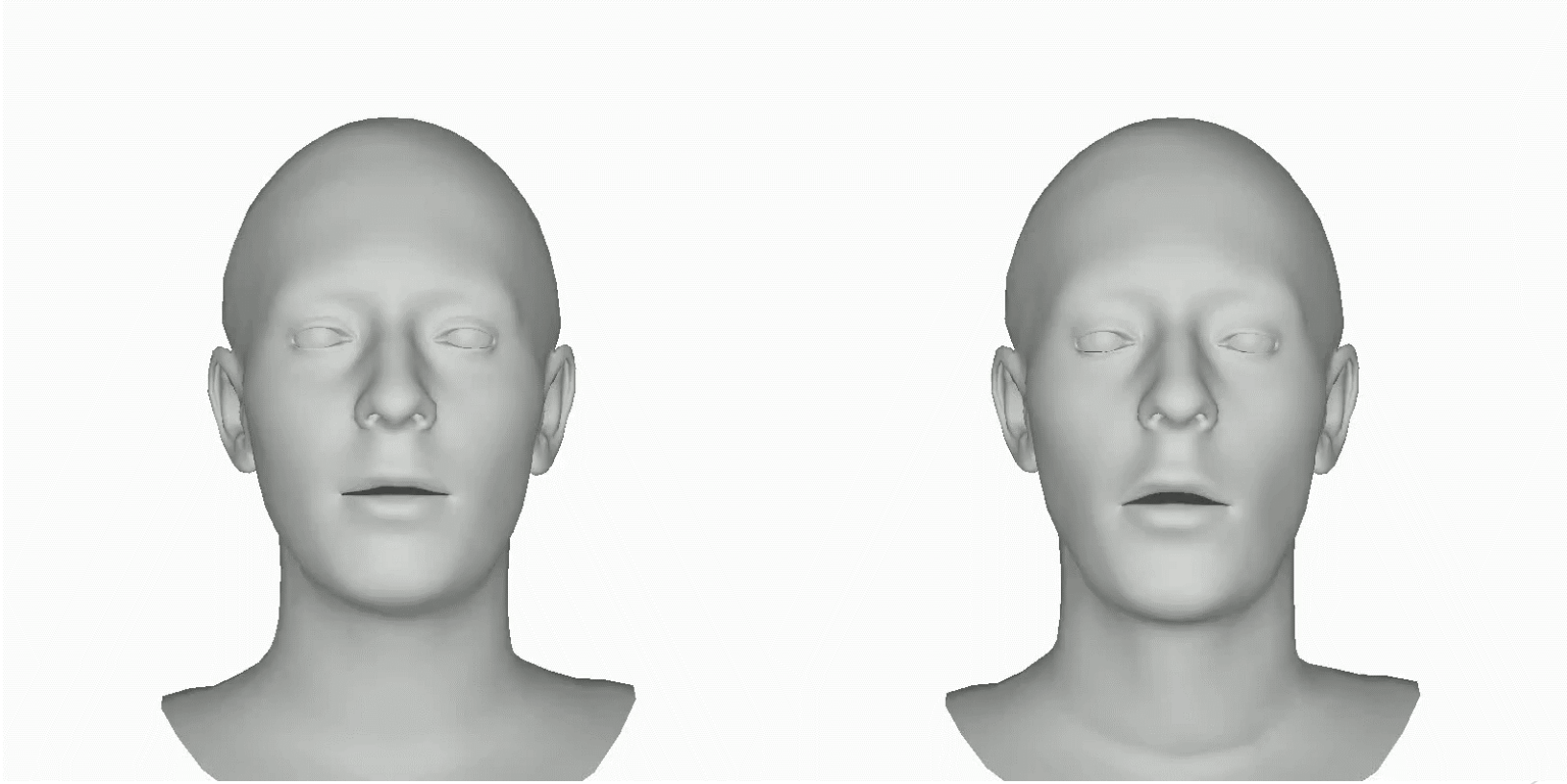


Ground Truth

Ours

Experiments

English



Ground Truth

Ours

Reference

- Cudeiro, Daniel, et al. "Capture, learning, and synthesis of 3D speaking styles." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.
- Sanyal, Soubhik, et al. "Learning to regress 3D face shape and expression from an image without 3D supervision." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.
- Li, Tianye, et al. "Learning a model of facial shape and expression from 4D scans." ACM Transactions on Graphics (ToG) 36.6 (2017): 194.

Thanks for listening