FaceFormer: Speech-Driven 3D Facial Animation with Transformers
—— Supplementary Material ——

Yingruo Fan[1] Zhaojiang Lin[2†] Jun Saito[3] Wenping Wang[1,4] Taku Komura[1*]
[1]The University of Hong Kong [2]The Hong Kong University of Science and Technology
[3]Adobe Research [4]Texas A&M University

In this supplementary material, we provide further information about FaceFormer, including detailed explanation of the FaceFormer architecture and the training details (Sec. 1), implementations of baseline methods (Sec. 2), the additional information about user study (Sec. 3) and a detailed illustration of periodic positional encoding (Sec. 4).

1. Implementation Details

Network architecture. The overall architecture of FaceFormer is illustrated in Fig. 2 of the main paper. In the encoder, the TCN is followed by a linear interpolation layer, which down/up samples the input to a given size determined by the frequency of the captured facial motion data. The interpolated outputs are then fed into 12 identical transformer encoder layers. For each transformer encoder layer, the model dimensionality is 768 and the number of attention heads is 12. Next, a linear projection layer is added on top of the transformer encoder layers, converting the 768-dimensional features to $\mathbf{d}$-dimensional speech representations ($\mathbf{d}$ = 128 for BIWI and $\mathbf{d}$ = 64 for VOCASET).

The motion encoder is a fully-connected layer with $\mathbf{d}$ outputs and the style embedding layer is an embedding layer with $\mathbf{d}$ outputs. The FaceFormer decoder has one decoder layer. The periodic positional encodings (PPE) have the same dimension as the motion encoder so that the two can be summed. For both the biased causal MH self-attention and the biased cross-modal MH attention, we employ 4 heads and set the model dimensionality to $\mathbf{d}$. The dimension of the FF layer is 2048. Similar to the encoder, the residual connections and layer normalizations are applied to the two biased attention layers and the FF layer. Finally, a fully-connected layer with $\mathbf{v}$ outputs is applied as the motion decoder ($\mathbf{v}$ = 70110 for BIWI and $\mathbf{v}$ = 15069 for VOCASET).

Training. We use the Adam optimizer with a learning rate of 1e-4. The parameters of the encoder are initialized with the pre-trained wav2vec 2.0 weights. During training, only the parameters of TCN are fixed. The models are trained for 100 epochs. The period $\mathbf{p}$ is set to 25 for BIWI and 30 for VOCASET.

2. Baseline Methods

As mentioned in the main paper, we compare FaceFormer with two state-of-the-art methods, VOCA [1] and MeshTalk [2], on both the BIWI and VOCASET datasets. For comparisons on BIWI, we use the original implementation of VOCA from their codebase[1]. Specifically, we train and test the VOCA model on BIWI. For comparisons on VOCASET, we directly use the provided trained VOCA model from their codebase and test it on VOCA-Test. Besides, we implement MeshTalk to the best of our understanding. We train and test MeshTalk on BIWI and VOCASET, respectively. One difference is the design of the UNet-style decoder: We modified the number of fully-connected layers from 7 to 3, as we found the original decoder would lead to overfitting due to the limited number of identities in BIWI and VOCASET.

3. User Study

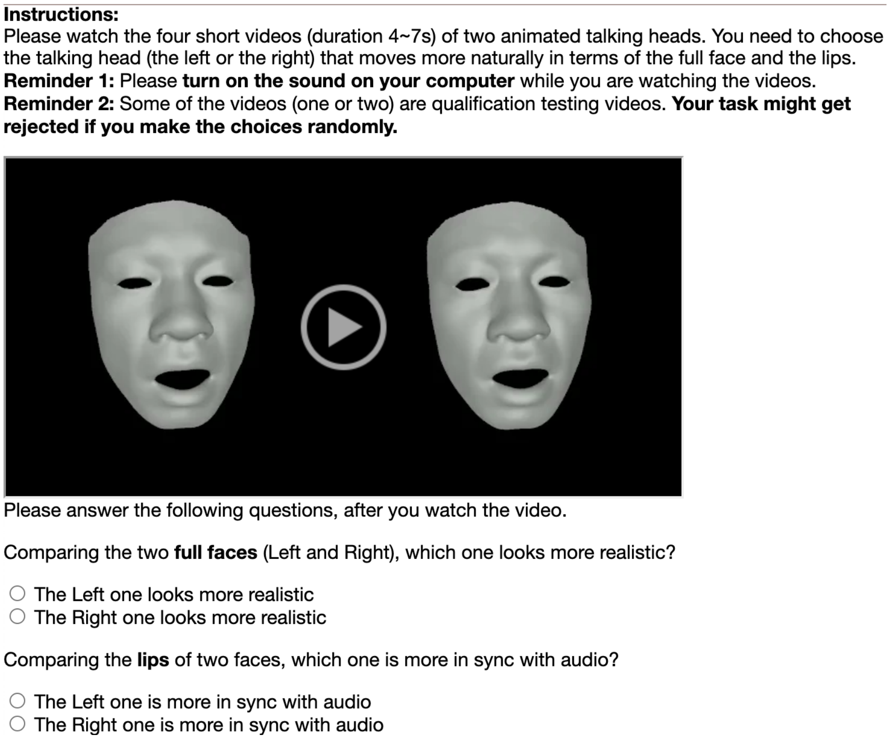
The designed user interface on Amazon Mechanical Turk (AMT) is shown in Fig. 1. To avoid Turkers' selecting an option randomly, we add one or two qualification testing videos for each HIT (human intelligence task). As shown in Fig. 2, Turkers could not submit their answers successfully if they failed to pass the hidden test. A warning message would pop up asking for checking the videos carefully before making the choices. Our recruitment requirement is that the Turkers have finished over 5000 HITs before and have an approval rate of at least 98%. In total, 576 A *vs*. B pairs (192 videos × 3 comparisons) are created for BIWI-Test-B, and 240 A *vs*. B pairs (80 videos × 3 comparisons) are created for VOCA-Test. For each HIT (human intelligence task), the AMT interface shows four video pairs in-

---

† Work done at HKUST

∗ Corresponding author

[1] https://github.com/TimoBolkart/voca



Instructions:
Please watch the four short videos (duration 4~7s) of two animated talking heads. You need to choose the talking head (the left or the right) that moves more naturally in terms of the full face and the lips.
Reminder 1: Please **turn on the sound on your computer** while you are watching the videos.
Reminder 2: Some of the videos (one or two) are qualification testing videos. **Your task might get rejected if you make the choices randomly.**

Please answer the following questions, after you watch the video.

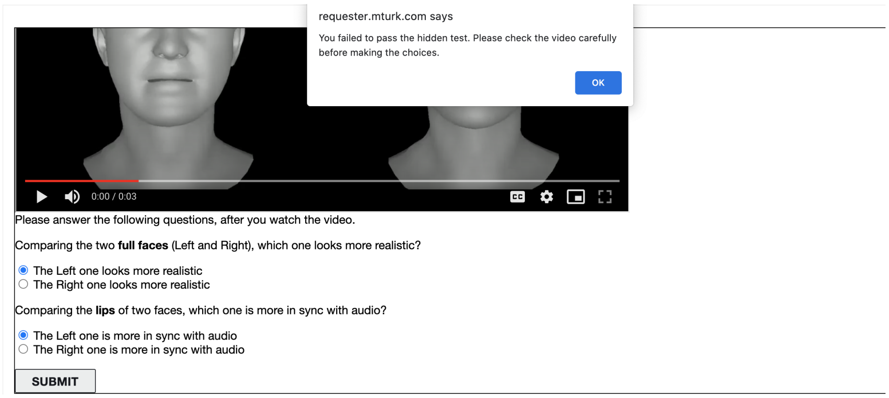
Comparing the two **full faces** (Left and Right), which one looks more realistic?

○ The Left one looks more realistic
○ The Right one looks more realistic

Comparing the **lips** of two faces, which one is more in sync with audio?

○ The Left one is more in sync with audio
○ The Right one is more in sync with audio

Figure 1. Designed user interface on AMT. Each HIT contains four video pairs and here only one video pair is shown due to the page limit.



requester.mturk.com says

You failed to pass the hidden test. Please check the video carefully before making the choices.

Please answer the following questions, after you watch the video.

Comparing the two **full faces** (Left and Right), which one looks more realistic?

● The Left one looks more realistic
○ The Right one looks more realistic

Comparing the **lips** of two faces, which one is more in sync with audio?

● The Left one is more in sync with audio
○ The Right one is more in sync with audio

Figure 2. Screen shot of the warning message.

cluding the qualification test in randomized order, and the Turker is instructed to judge the videos w.r.t two questions: “Comparing the two full faces, which one looks more realistic?” and “Comparing the lips of two faces, which one is more in sync with audio?”.

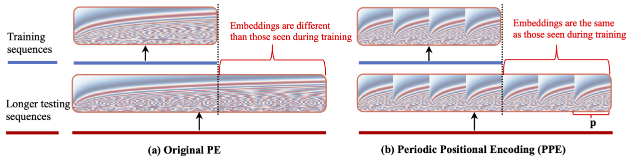

Training sequences
Embeddings are different than those seen during training
Embeddings are the same as those seen during training
Longer testing sequences
p
(a) Original PE
(b) Periodic Positional Encoding (PPE)

Figure 3. Detailed illustration of periodic positional encoding.

4. Periodic Positional Encoding

Fig. 3 provides the illustration of how PPE works. The “Original PE” has the problem in generalizing to longer sequences. Making “PE” periodic (“PPE”) can improve the generalization ability. Meanwhile, the temporal bias (“TB”) in Fig. 4 of the main paper brings the relative position information to each period and encourages the model to look at nearer previous lip motions. “ALiBi” directly removes

"PE", which improves the generalization but is less robust when encoding the temporal order information, especially for our case where adjacent frames have subtle motion variations. Thus, "TB+PPE" realizes the tradeoff between "PE" and "ALiBi".

References

[1] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10101–10111, 2019. 1

[2] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando de la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1173–1182, 2021. 1