Supplemental Materials of Talking-head Generation with Rhythmic Head Motion

Lele Chen[†], Guofeng Cui[†], Celong Liu[‡], Zhong Li[‡], Ziyi Kou[†], Yi Xu[‡], and Chenliang Xu[†]

[†] University of Rochester [‡] OPPO US Research Center
lchen63@cs.rochester.edu

We introduce the network details in Sec. A. In Sec. B, we show more results in qualitative and quantitative perspective. Note that the actual results of other comparison methods could be better, since we replicate them by ourselves. We will update those results once the code is publicly available.

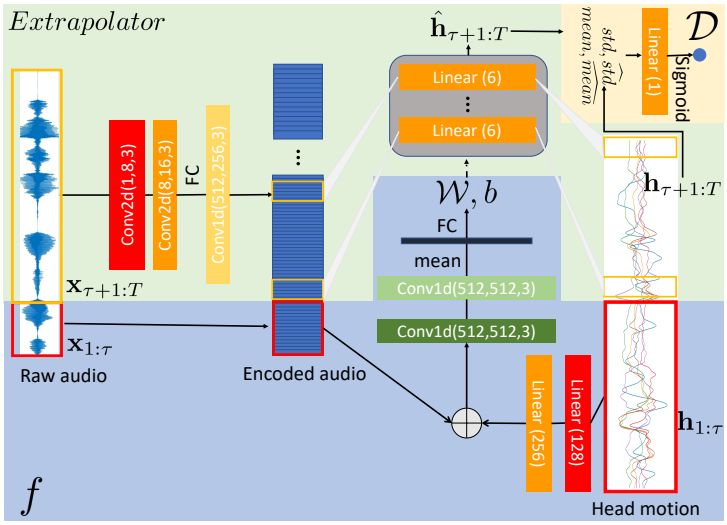

Fig. 1. The details of the head motion learner, which consists of a encoder f (blue part), an extrapolator (green part), and a discriminator (yellow part).

A Network Details

A.1 Details of The Head Motion Learner (Φ)

The head motion learner Φ consists of three sub-networks: a head motion encoding network f, a head motion extrapolation network *Extrapolator*, and a

discriminator $\mathcal{D}$. Fig. 1 shows the detailed network structure. Specifically, we first use f to encode the raw audio $\mathbf{x}_{1:\tau}$ and its paired head motion $\mathbf{h}_{1:\tau}$ to network weights $\mathbf{w}$, which contains weights and biases $\{\mathcal{W}, b\}$ for a linear layer. Since the audio sampling rate is 50000 and the image sampling rate is 25FPS, the size of the input $\mathbf{x}_{\tau+1:T}$ should be $(T-\tau) \times 0.04 \times 50000$. At time step t, we use $\mathbf{x}_{t-3:t+4}$ to represent the audio signal. So, after stacking, the size of the input to *Extrapolator* is $(7, (T-\tau) \times 0.04 \times 50000)$. In the *Extrapolator*, we apply two 2D convolutional layers on the inputs and then flatten it. Then we apply a 1D temporal convolution layer to encode it to audio feature with the size of $(T-\tau, 256)$. Then at each time step t, we forward the feature chunk (size of $1, 256$) to a linear layer, where the weights and biases $\{\mathcal{W}, b\}$ are learned from f. Once the *Extrapolator* generates all the fake head motion $\hat{\mathbf{h}}_{\tau+1:T}$, we forward it with ground truth motion $\mathbf{h}_{\tau+1:T}$ to discriminator $\mathcal{D}$. The $\mathcal{D}$ calculates the mean and standard deviation from real and fake sequences and then output a real/fake score based on the mean and standard deviation.

A.2 Details of The Facial Expression Learner (Ψ)

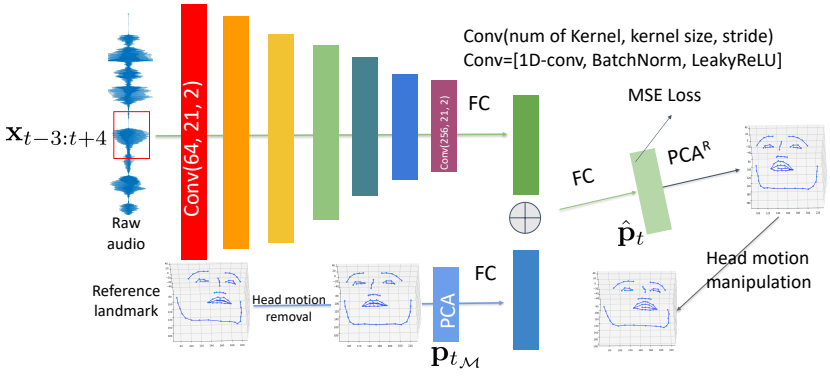

Fig. 2. The details of the facial expression learner. PCA^R denotes reversed PCA operation.

The facial expression learner Ψ is a linear regression network, which takes current audio chunk $\mathbf{x}_{t-3:t+4}$ and reference landmark PCA components $\mathbf{p}_{t_\mathcal{M}}$ as input and output current facial expression $\hat{\mathbf{p}}_t$. We use 20 PCA coefficients to represent facial expression. We list the details in Fig. 2. We directly train and optimize the model on the PCA components output. During inference, we reconstruct the 3D landmark points from the 20 PCA coefficients.

Fig. 3. The details of the unprojector. We use the method proposed in [4] as the unprojector.

A.3 Details of The 3D Unprojection Network

The unprojection network receives a RGB image $\mathbf{y}_{t_{\mathcal{M}}}$ and predict the position map image. We follow the same training strategy and network structure as [4], which employs an encoder-decoder structure to learn the transfer function. we train the unprojection network on 300W-LP [9], since it contains face images across different angles with the annotation of estimated 3DMM coefficients, from which the 3D point cloud could be easily generated. We calculate MSE loss between the predicted position map and the ground truth position map. For training details, please refer to [4].

Fig. 4. The testing results on President Barack Obama's weekly address footage dataset [5].

B More Results

B.1 Controllable Videos

We show one testing results on VoxCeleb2 dataset to demonstrate our ability of generating controllable head motion and facial expression. From Fig. 5, we can find that our model can generate controllable videos with desired head motion and facial expressions.

B.2 Test on President Barack Obama Footage Dataset

To further improve the image quality, we fine-tuned our model with $K = 8$ on five videos from the President Barack Obama's weekly address footage dataset [5] and leave the rest videos as testing set. Fig. 4 shows two example testing results.

Table 1. Ablation studies on VoxCeleb2 dataset. Our model mentioned in this table are trained from scratch.

Method	CSIM↑	SSIM↑	FID↓
Baseline	0.19	0.67	112
Full Model	0.44	0.71	40.8
w/o 3D-Aware	0.21	0.61	109
w/o Hybrid-Attention	0.37	0. 73	57.8
w/o Non-Linear Comp.	0.40	0.69	64.5
w/o warping	0.34	0.67	78.2

B.3 Ablation Studies

We conduct ablation experiments to study the contribution of four components: 3D-Aware, Hybrid-Attention, Non-Linear Composition, and warping. The Baseline model is a straightforward model without any features (e.g. 3D-Aware, Hybrid-Attention). Table. 1 shows the quantitative results of ablation studies.

B.4 Settings of User Studies

Human subjects evaluation is conducted to investigate the visual qualities of our generated results compared with Zakharov et al. [8], Wang et al. [6] and Wiles [7]. The ground truth videos are selected from different sources: we randomly select samples from the testing set of LRW [3], VoxCeleb2 [2] and LRS3 [1]. Three methods are evaluated w.r.t. two different criteria: whether participants could regard the generated videos as realistic and whether the generated talking-head videos temporally sync with the corresponding audio. We shuffle all the sample videos and the participants are not aware of the mapping between videos to methods. They are asked to score the videos on a scale of 0 (worst) to 10 (best). There are overall 20 participants involved (at least 50% of them are native English speaker), and the results are averaged over persons and videos.

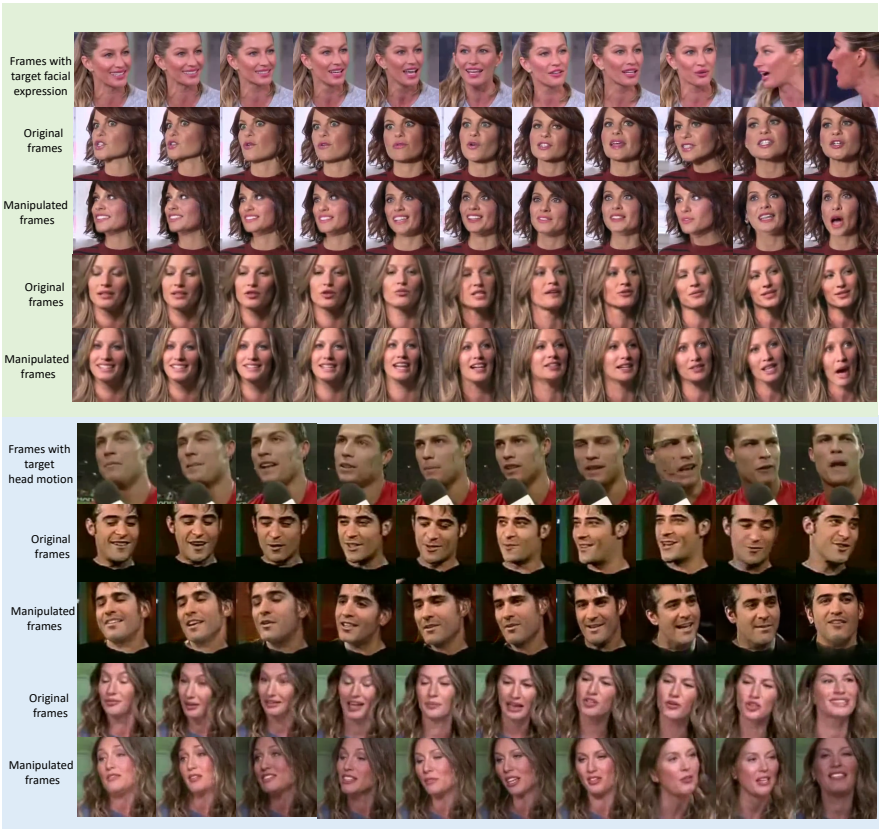

Fig. 5. The controllable results. The videos in upper part are manipulated with target facial expressions while keep the head motion unchanged. We show the target facial expression in the first row. The videos in lower part are manipulated with target head motion while keep the facial expression unchanged.

References

1. Afouras, T., Chung, J.S., Zisserman, A.: Lrs3-ted: a large-scale dataset for visual speech recognition. In: arXiv preprint arXiv:1809.00496 (2018)
2. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. In: INTERSPEECH (2018)
3. Chung, J.S., Zisserman, A.: Lip reading in the wild. In: Asian Conference on Computer Vision (2016)
4. Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X.: Joint 3d face reconstruction and dense alignment with position map regression network. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 534–551 (2018)
5. Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing obama: learning lip sync from audio. ACM Transactions on Graphics (TOG) **36**(4), 95 (2017)
6. Wang, T.C., Liu, M.Y., Tao, A., Liu, G., Kautz, J., Catanzaro, B.: Few-shot video-to-video synthesis. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
7. Wiles, O., Sophia Koepke, A., Zisserman, A.: X2face: A network for controlling face generation using images, audio, and pose codes. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 670–686 (2018)
8. Zakharov, E., Shysheya, A., Burkov, E., Lempitsky, V.: Few-shot adversarial learning of realistic neural talking head models. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
9. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: A 3d solution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 146–155 (2016)