Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation (Supplemental Materials)

Hang Zhou[1], Yasheng Sun[2,3], Wayne Wu[2,4], Chen Change Loy[4], Xiaogang Wang[1], Ziwei Liu[4] ✉
[1]CUHK - SenseTime Joint Lab, The Chinese University of Hong Kong [2]SenseTime Research
[3]Tokyo Institute of Technology [4]S-Lab, Nanyang Technological University
{zhouhang@link,xgwang@ee}.cuhk.edu.hk, wuwenyan@sensetime.com, {ccloy,ziwei.liu}@ntu.edu.sg

1. Descriptions to the Supplemental Video

The video [7] consists of 4 parts, including the introductory sentences, comparisons with previous methods, ablation study, and frontalization results.

Experimental Settings. All of our videos are generated based on one *identity reference* image and driven by a clip of audio with a *pose source* video. The mouth motion of our generated results should be synced with the audio and the head poses are controlled by the pose source. The pose source videos are retrieved from 50 randomly sampled videos in the test set based on the distances between the pose feature $f_{p(1)}$ encoded from the first frame of the videos and $f_{p(ref)}$ encoded from the reference image.

Introductory Sentences. At first, we show results of several different identities speaking the same sentence, and one fixed identity speaking another sentence with different poses. All identity references are places above the generated results, and the driving pose sources are placed underneath. During inference, the mouths of the pose source videos are not shaded. The effect in the videos is an illustration that the mouth shapes of pose source videos have no effect on our results.

Comparisons with Previous Methods. We compare our method with ATVG [2], Wav2Lip [6], MakeitTalk [8] and Rhythmic Head [1], which are the representatives of 2D landmark-based, reconstruction-based, 3D landmark-based and 3D model-based methods respectively. As free pose control is non-applicable in most of the method, we keep the rest of the settings the same as ours when re-implementing their results. We use the same pose source to drive Rhythmic Head [1] as ours. The results of Rhythmic Head are produced by their authors, thus only two samples are given. Please note that under large poses, the integrated landmark detector of ATVG [2] would fail. Also, the mouth movements of MakeitTalk [8] will be less accurate under such circumstances, showing the non-robustness of structural information. While the lip sync of Wav2Lip [6] is basically accurate, they change only the mouth shapes. Realistic videos cannot be created with their method when only one image is given as reference.

Table 1: **Additional quantitative results on VoxCeleb2**. For SyncNet offset Sync_{off}, the closer to the ground truth the better.

Method	PSNR ↑	Sync_{off}	CSIM ↑
ATVG [2]	29.41	0.2	NA
Wav2Lip [6]	29.54	0.5	**0.987**
MakeitTalk [8]	29.51	1.9	0.935
Rhythmic Head† [1]	29.55	0.7	0.921
Ground Truth	NA	0.8	0.954
PC-AVS (Ours)	**29.68**	0.9	0.950

Ablation Study and Additional Results. We also show the results of our ablation studies. As can be seen in the video, the head poses of ours without $\mathcal{L}_c$ is non-changeable. We show the result of a model whose pose-dim is set as 36, which is shakier than ours. Certain artifacts would appear when changing the generator in an AdaIN style. Finally, we show the results of talking face frontalization by setting all values in the pose feature to 0s. The results of two different languages are illustrated, showing that our model is robust to large poses and different languages.

2. More Quantitative Evaluations

We conduct more quantitative evaluations on the VoxCeleb2 dataset. These metrics are not discriminative enough but are also widely used in previous studies. They are **PSNR** that is used for image quality; the cosine similarity (**CSIM**) between embedding vectors of a pretrained face recognition model [5] to verify the identity preserving; and the **offset**

(Sync_{off}) is used to measure how much video lags the audio in SyncNet [4]. Note that it is not the main problem in talking face generation, thus this value is only shown as a reference.

The results are shown in Table 1. As Wav2lip keeps the head pose unchanged, it is natural that their facial identity preserving is better than other methods. We have also conducted an experiment to test the lip reading accuracy on LRW [3] dataset. However, as there is no accurate lip reading model available, the results are not informative. Wav2lip [6] achieves the best on the lip reading evaluation.

References

[1] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. *European Conference on Computer Vision (ECCV)*, 2020. 1

[2] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[3] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *ACCV*, 2016. 2

[4] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *ACCV*, 2016. 2

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[6] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia (ACMMM)*, 2020. 1, 2

[7] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[8] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makeittalk: Speaker-aware talking head animation. *SIGGRAPH ASIA*, 2020. 1