Meta Talk: Learning to Data-Efficiently Generate Audio-Driven Lip-Synchronized Talking Face with High Definition

Yuhan Zhang[1], Weihua He[1] *, Minglei Li[2], Kun Tian[1], Ziyang Zhang[1#], Jie Cheng[1], Yaoyuan Wang[1#], Jianxing Liao[1]*
1 Advanced Computing and Storage Lab, Huawei Technologies Co. Ltd., China
2 Language & Speech Innovation Lab, Huawei Technologies Co. Ltd., China

Introduction

Audio-driven talking face, driving talking face by audio, has received considerable attention in multi-modal learning due to its widespread use in virtual reality. However, long-time recording of target high-quality video is needed by most existing audio-driven talking face studies, which significantly increases customization costs. The method based on 3D morphable model (AudioDVP) [1] reduces the burden of target video acquisition, but lip shape is not synchronized well with an arbitrary new piece of audio in the generated video as its audio-driven performance strongly depends on the audio identity. The approach based on GAN (Wav2Lip) [2] uses a pre-trained discriminator to accurately detect lip-sync errors and force the generator to accurately morph the lip movements in sync with a new audio in the wild instead of the target's audio. Although it produces a decent lip-syncing video of the talking face and achieves disentanglement of audio identity and model and target identity, the definition of the lip area is always poor for visual experience and cannot meet application requirements.

Method	Training data	Definition of generated face	Lip-sync of generated video
Our method	3 min high-definition video	High-definition (1024x1024)	✓
AudioDVP	3 min high-definition video	Limited-definition (256x256)	limited
AudioDVP	5~6h high-definition video	Limited-definition (256x256)	✓
Wav2Lip	Over 30h low-definition video	Low-definition (96x96)	✓

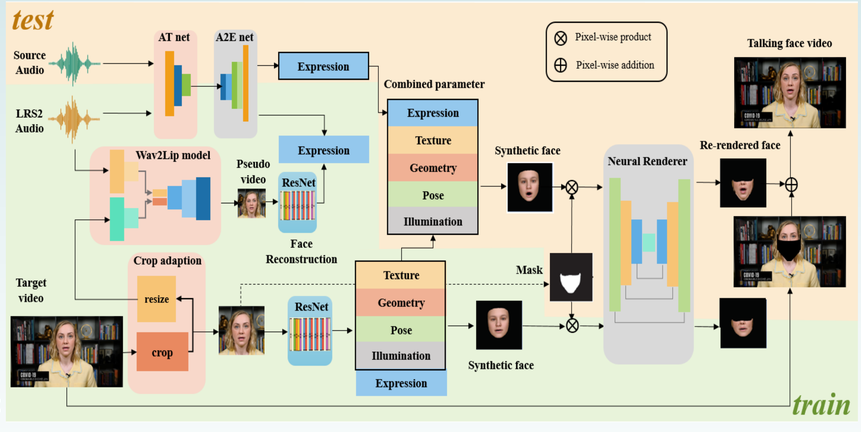
Contribution

In this paper, we propose a novel talking face generation framework and strive to transfer the powerful audio driven lip-syncing abilities from a pretrained model [2] to ours using only short training target video. The key contribution of this paper are summarized as follows:

- the low-definition pseudo video predicted by Wav2Lip with the target video and LRS2 audio is introduced to enhance the audio- driven identity-disentangled ability of talking face generation.
- We train a modified audio-to-expression (A2E) network to guarantee the accurate lip motion driven by arbitrary audio, which makes our method possess an powerful audio driven performance comparable to Wav2Lip [2].
- A modified crop module is introduced for automatically adapting the size of the 3DMM synthetic face to the original face area, then enabled our framework to meet the requirements of 4K-definition photo-realistic talking face video.

Method

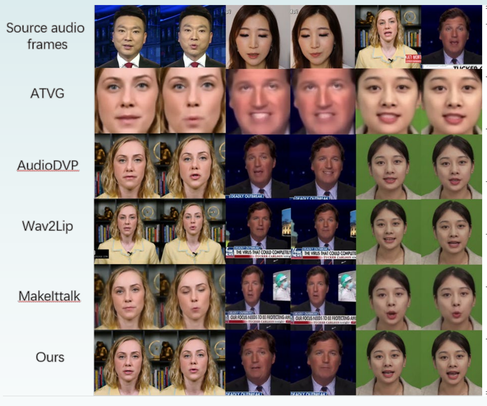
During the training phase, we first crop the original target video into the target face video, which is then resized to be low-resolution to generate a low-definition talking face video with LRS2 audio using the pre-trained model Wav2Lip [2]. The generated video-audio pair is the pseudo label which possesses abundant phonemes and corresponding talking face video with excellent lip synchronization. 3D face reconstruction is performed on both the pseudo video and the target video, and the facial 3D morphable model (3DMM) parameters including expression, geometry, texture, pose, illumination coefficients are extracted from each frame of them. To obtain a powerful model mapping audio to expression parameters, a new audio-to-expression transformation network is trained with audio-expression pairs of the pseudo video. Then, the 3DMM parameters are used to re-render the synthetic facial images in the target video. Finally, we train a neural rendering network with the lower half of synthetic and real target faces to generate a high-definition photo-realistic talking face video.



During the testing phase, arbitrary audio can be input and fed into the trained audio-to-expression transformation network to predict audio-driven expression parameters. Then, the predicted expression parameters replace the original ones of 3DMM parameters obtained by 3D reconstruction. We re-render the face to audio-driven synthetic face using the combined 3DMM parameters. Then the lower half of the generated synthetic face is translated to a realistic lower half face. Finally, the generated photo-realistic lower half face is sewn into the background of original target video to generate a high-definition lip-synchronized video.

Experiments

We tested our method on the videos of seven characters collected from the previous work [1]. Only 3min of them are used to train the model. We first aligned all the speaking faces by detecting their landmarks, and then cropped the video to a 512x512 or 768x768 frame size centered around the lower half of the face. Then, the center image frame are used as the paired image data to finally generate a 28 × 80 MFCC feature for each 10ms audio block. We compared our method with ATVG [3], Wav2lip [2], AudioDVP [1] and Makelttalk [4] by testing their driven performance on audio from multi persons. The comparison results among these methods are shown in the figure. Our method generates more synchronized lip movements compared with the other four methods. The generated video can show more texture details of the face and even freckles on the F's face more clearly. Then, metrics LSE-D and LSE-C from [2] are adopted for quantitative evaluation of lip-syncing performance in the wild, and FID [2] and SSIM for image quality (see the table). The lip-sync performance of our method is comparable to Wav2Lip, and our method produces videos with the best image quality among these methods.



	Methods	ATVG	Wav2Lip	AudioDVP	MakeIttalk	Ours
A	LSE-D↓	9.114	7.756	10.195	9.977	8.636
	LSE-C↑	5.653	7.555	4.138	4.716	6.060
	FID↓	21.572	11.847	9.437	23.158	6.734
	SSIM↑	0.5298	0.6072	0.9490	0.5526	0.9832
B	LSE-D↓	10.400	7.540	14.978	11.911	9.878
	LSE-C↑	5.234	6.066	0.238	2.135	4.963
	FID↓	19.983	13.120	10.234	19.315	7.065
	SSIM↑	0.6721	0.6049	0.9645	0.6238	0.9896
C	LSE-D↓	10.581	6.637	11.712	16.170	9.530
	LSE-C↑	5.122	8.951	3.322	0.06	6.141
	FID↓	19.311	11.154	9.677	19.46	6.498
	SSIM↑	0.350	0.5647	0.9316	0.4781	0.9744
D	LSE-D↓	10.005	6.546	11.804	11.444	9.091
	LSE-C↑	5.808	9.023	2.884	3.725	6.155
	FID↓	17.969	12.485	9.076	19.93	6.881
	SSIM↑	0.5682	0.5766	0.9439	0.4705	0.9875
E	LSE-D↓	12.506	6.571	11.713	14.339	9.831
	LSE-C↑	2.734	8.989	3.063	1.018	5.493
	FID↓	18.697	11.185	9.523	19.502	7.131
	SSIM↑	0.5716	0.6482	0.9237	0.4872	0.9867
F	LSE-D↓	9.567	6.343	9.953	11.167	8.817
	LSE-C↑	5.803	9.314	4.794	4.277	5.841
	FID↓	20.839	13.457	11.348	17.746	6.775
	SSIM↑	0.5321	0.7015	0.9577	0.4764	0.9894
G	LSE-D↓	9.687	6.013	13.332	8.831	8.880
	LSE-C↑	7.261	10.237	2.074	6.795	7.539
	FID↓	21.894	14.94	10.795	23.880	6.579
	SSIM↑	0.5794	0.6102	0.9102	0.7736	0.9866

References

[1] XinWen, MiaoWang, Christian Richardt, Ze-Yin Chen, and Shi-Min Hu, "Photorealistic audio-driven video portraits," IEEE Transactions on Visualization and Computer Graphics, vol. 26, no. 12, pp. 3457–3466, 2020.
[2] 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, New York, NY, USA, 2020, MM '20, p. 484–492, Association for Computing Machinery.
[3] Lele Chen, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7824–7833.
[4] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li, "Makelttalk," ACM Transactions on Graphics, vol. 39, no. 6, pp. 1–15, Nov 2020.