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 target identity, the definition of the lip area is always poor for visual experience and cannot meet application requirements.

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.

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 is used as the paired image data to finally generate a 28x38 MFCC feature for each 10ms audio block. We compared our method with ATVG [3], Wave2lip [2], AudioDVP [1] and MakeIttalk [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.

References