Brand-new Speech Animation Technology based on First Order Motion Model and MelGAN-VC

Shaomin Chen, Xinyi Gao, Jiangtao Wang*, Yu Xiao, Yueling Zhang, Gang Xu
National Trusted Embedded Software Engineering Technology Research Center, East China Normal University, 3663, North Zhongshan Road, China
*Email: jtwang@sei.ecnu.edu.cn

Abstract. Speech animation has huge application potential in instant messaging and entertainment media fields such as videophones, virtual meetings, audio and video chats. The traditional voice-driven speech animation has the problem of a single adaptation language, and the performance-driven speech animation has the problem of high cost of capture equipment and difficult mass production. Based on the above existing problems, we propose a new method of speech animation generation, that is, given a static portrait of a person and a face-driven video, finally generate a face animation video of the character in the given portrait. The conversion system consists of two parts: face conversion and voice conversion. We noticed that the final generated face animation video has problems such as low definition, not smooth playback, and metallic sound. On this basis, this article proposes to increase the animation enhancement experiment and replace the encoder measures for improvement. Through comparative experiments, the above measures are proved to be effective.

1. Introduction
Speech animation is designed to drive the vertices of the head graphic model to synthesize the movement of the articulators corresponding to the given audio. Highly natural voice animation can effectively improve user experience. Recently, the rapid development of virtual reality technology has further highlighted the urgent need for human-machine natural communication in immersive environments.

Although researchers at home and abroad have made certain achievements in the research of speech animation [1], there are still following problems: 1) Although the speech animation method based on performance-driven can obtain high-quality face animation, the cost of motion capture equipment is high, and it is not suitable for batch synthesis speech tasks. 2) The language recognized by the voice-driven speech animation technology is single, and the reusability is not strong. 3) The synthesized speech animation does not perform well in facial details.

To solve above problems, there is a method be proposed for generating facial animation video based on a static portrait picture and a source video, which mainly includes two parts: facial animation generation and voice conversion. The main function of the facial animation module is to synthesize a facial animation video of a character in a given picture. This paper uses a motion-based First Order Model [2] as a facial animation conversion model. In order to improve the frame rate and video resolution, VFIASC network was added to carry out video frame insertion experiments, and the video was super-resolution processed through ESRGAN network to obtain more realistic videos. The voice conversion module in this article uses
MelGAN-VC [3], a one-to-one voice conversion model, which is composed of a generator, a discriminator and a siamese network. In order to reduce the metallic feeling of synthesized audio, on this basis, the WaveGlow is used to replace the original Griffin-Lim vocoder, making the synthesized sound closer to the real human voice. The speech animation generation method be proposed by this paper is very low-cost and easy to perform batch voice synthesis tasks. The synthesized speech animation video is kept in sync with the auditory and visual aspects, and looks close to a real person.

2. Background
Researchers at home and abroad have done a lot of research work on speech animation generation technology. How to generate a high-quality, high-natural, and low-cost speech animation has always been a valuable research focus.

Facial Animation: Recently, deep learning has become an effective technology for face animation and video relocation [4]. In particular, GAN [5] and VAEs [6] have been used to transfer video between facial expressions or motion patterns. However, these methods mostly rely on pretrained models to extract object-specific representations, such as key point locations. Compared with these works, the technology does not rely on tags, prior information about animated objects, or the specific training process of each object instance. In addition, it is applied to different objects of the same type.

Speech Conversion using Generative Adversarial Networks: Generative Adversarial Networks [7] are widely used for image generation and image-to-image translation [8]. A lot of work has shown that it is possible to apply the GAN designed for image conversion to other types of data (such as voice data). [9] shows that it is feasible to use GAN to generate audio data, that is, use convolutional architecture on waveform and spectrogram data. [10] suggests using different GAN architectures to convert different voice features instead of spectrograms. In this way, the network does not rely on pixel-wise differences, and is proven to be more flexible when converting between domains with basically different low-level pixel features.

3. Proposed Approach
Given a static portrait of character A and a speech video of character B as input, our goal is to generate a speech video of character A while retaining the facial expression features and semantic information of character B. The method be proposed mainly includes two parts, face animation module and voice conversion module.

3.1. Facial Animation Model
Video is composed of video frames, so the training process of generating facial animation can be regarded as an image reconstruction process. This paper draws on the first-order motion based facial animation model proposed by [2] (see Figure 1). Firstly, an unsupervised face key point detection module is designed to estimate the feature key points in the input image and video frame respectively. Secondly, a motion estimation module is designed to transform the sparse key points into dense motion optical flow, and the occlusion mask is obtained. After that, a picture generation module is used to feed the motion optical flow and occlusion mask into the module to get the final output, that is, the original face animation. The originally generated face animation video still has the problems of low playback smoothness, blurred pictures and unpleasant artifacts. On this basis, the VFIASC is proposed to be added to network to perform frame insertion experiments on the original generated video. Then input the video after the frame insertion experiment to the ESRGAN network for super-resolution experiment. The edge and texture details of the image reconstructed by the ESRGAN network will be more delicate than the original image, and it is widely used in image restoration.
3.2. Voice Conversion Model
MelGAN-VC model is composed of a generator, a discriminator and a siamese network (see Figure 2). In the training process, the generator learns the mapping function from the A domain to the B domain, and the spectrogram $x$ generates the data $b$ mapped to the B domain through the generator $G$, in an attempt to deceive the discriminator. The discriminator strives to distinguish the real spectrogram from the generated spectrogram, and encourages the generator to generate a more realistic spectrogram. The siamese network translates the spectrogram of the source domain to the target domain through cooperation with the generator. The siamese network mainly learns the high-level semantic features in the spectrogram, so that the translated spectrogram is similar to the source domain. Because the actual effect of the voice converted using the MelGAN-VC model is not ideal, it sounds metallic, and some semantic information is lost, instead of the traditional vocoder Griffin-Lin, WaveGlow is used in the original method. Using the new vocoder can speed up the synthesis of the spectrogram to the speech signal and improve the conversion efficiency.

3.3. Loss Function
Perceptual loss: Since the training process is actually a process of image reconstruction, the perceptual loss proposed by Johnson [11] is used for reference. Pass the driving image $D$ and the generated image through VGG19 respectively, and the reconstruction loss is as follows, among them, $N_i(\cdot)$ represents the feature value of the i-th layer:

$$L_{rec}(\widehat{D}, D) = \sum_{i=1}^{l} |N_i(\widehat{D}) - N_i(D)|$$

TraVeL loss: Originally introduced in [12], the TraVeL loss (Transformation Vector Learning loss) aims at keeping transformation vectors between encodings of pairs of samples. A transformation vector is defined between $(x_i, x_j) \in X$ as
\( t(x_i, x_j) = x_j - x_i \) \hspace{1cm} (2)

A siamese network \( S \) is used to encode samples in a semantic latent space and formulate a loss to preserve vector arithmetic in the space such as
\[
t(S(\frac{a}{2}, i), S(\frac{a}{2}, j)) = t(S(\frac{b}{2}, i), S(\frac{b}{2}, j)) \forall i, j \quad (3)
\]

**Identity loss:** In the case of using only conversion loss, although the generated voice samples sound real, these samples lose the timbre of the target. Therefore, the identity loss function is used to ensure that it is consistent with the identity of the target object:
\[
L_{G, id} = E_{\hat{b} \sim B} \left[ \left\| G(\hat{b}^{id}) - b^{id} \right\|_2^2 \right] \quad (4)
\]

4. Implementation

4.1. Implementing Environment
Due to laboratory conditions, and the training of spectrum images involves many matrix calculations, multiple GPUs are required to speed up the parallel calculation speed and training speed. All experiments in this article are based on the cloud server provided by Google. The GPU provided by Google Colab is Tesla P100, the video memory is 32G, the memory is 128G, and the deep learning framework is pytorch-1.0.0.

4.2. Face Animation Experiment
In order to save memory and reduce calculations, both the unsupervised key point prediction module and the dense motion estimation module are based on pictures with a resolution of 64*64. The Adam optimizer [14] is used to with a batch size of 20 and a learning rate of 0.0002 to train the network, and train this model iteratively for a total of 40,000 times. After 20,000 iterations, the learning rate begins to decay linearly. The datasets VoxCeleb and UVA-Nemo that be downloaded from YouToBe were used as training face animation models.

The experiment used Leo's speech video as the driving video, and used three Chinese celebrity photos and a girl statue picture as test samples for testing, and intercepted three different expressions of the characters at different moments as the test results. Among the 10 key points detected by unsupervised learning, most of the key points have clear directions (see Figure 3), which lays the foundation for the next expression transfer algorithm. By comparing the expressions of all the characters at the same moment, the facial expression features of the characters in the original video are perfectly mapped to the faces of the target characters. Within the allowable error range, the facial expressions and head rotation angles in the driving video and the target video appear almost the same at the same time, the facial edge deformation is small, and the overall effect is small.

![Figure 3. Experimental results of face key point detection.](image-url)
The enlarged picture has unpleasant artifacts and the picture definition is not high. In order to further improve the image quality, video super-resolution experiments are conducted on the video after the frame interpolation process. ESRGAN network and EDVR network are used for experiment. ESRGAN (Enhanced Super Resolution ResNet) is an advanced version of SSSResNet (Super Resolution ResNet). The edge and texture details of the image reconstructed by the ESRGAN network will be more delicate than the original image, and it is widely used in image restoration. EDVR handles two tasks: video super-division and deblurring. Here it is used for video super-resolution experiments. What is different from ESRGAN is that EVDR does not regard the video super-division task as a simple extension of the image super-division, it also uses the redundant information between video frames. Three experiments have been designed in total, the first one uses only the ESRGAN network, the second only uses the EVDR network, and the third is the superposition of the ESRGAN network and the EVDR network. The effects of the three experiments are shown in Figure 4. It can be seen that the performance is best when only the ESRGAN network is used for the experiment, and the edges and texture details of the video image are more delicate than the original image. The video images processed by the remaining two methods still have artifacts. Through comparative experiments, the first strategy has been chosen to apply to actual model.

![Figure 4. Super-resolution experiment results.](image)

4.3. Voice Conversion Experiment

A full convolution model architecture had been used in the design of generators, discriminators, and twin networks. The generator adopts the U-Net architecture, uses a convolution kernel with a step size of 2 for down-sampling, and up-sampling by sub-pixel convolution. This method can eliminate the artifacts caused by deconvolution. In addition, each convolutional layer in the generator and discriminator uses spectrum normalization, which makes the training process more stable. The experiments use two data sets to train the speech model, the author’s voice and the audio downloaded from the Internet by Degang Guo. Each piece of audio data lasts from 3 to 10 seconds, a total of 680 pieces, a total of 1500 cycles of iterative training, after 200 epochs, the generator and discriminator have stabilized.

It can be seen from Figure 5 that the contours of the source speech spectrum and the synthesized speech spectrum are very similar. Compared with the MelGAN-VC model, our proposed method can retain more original sounds details of the information. Because the actual effect of the voice converted using the MelGAN-VC model is not ideal, it sounds metallic, and some semantic information is lost, instead of the vocoder Griffin-Lin, the WaveGlow is used in the original method. It can be seen from Figure 6 that the MOS score of the speech generated by the improved method has been significantly improved.
Figure 5. Voice pectrum diagram.

The picture on the left is a spectrogram obtained by training with MelGAN-VC. From left to right are the target voice, source voice and generated voice. The figure on the right is the spectrogram generated by the method proposed in this article. From left to right are the source voice, target voice, and generated voice.

Figure 6. Comparison of MOS scores using vocoder Griffin-Lin and WaveGlow

5. Conclusion
This article studies two aspects: facial animation and voice conversion. In order to synthesize face animation, this article uses face animation based on first-order motion as the basic model, and the voice conversion module draws on the MelGAN-VC model. On this basis, an animation enhancement experiment has been proposed on the generated video to improve the smoothness and clarity of the video, and reduce the metallic feeling of the synthesized sound by replacing the encoder, so that the generated facial animation looks more natural, real and lifelike as a whole. Compared with the traditional performance-driven method, the equipment cost of this method is negligible, and it is easy to mass produce. It has good practicability and application prospects. Compared with the traditional voice-driven voice animation generation method, our method only converts voice features in the voice conversion part, and can adapt to multiple languages. Although applications such as voice animation do not seem to have dangerous uses, such technologies can easily be misused to create fake video data for political or personal reasons. Crucially, resources must also be invested in developing methods to identify fake audio data.

References
[1] Aleksic P S and Katsaggelos A K 2004 Speech-to-video synthesis using MPEG-4 compliant visual features IEEE Transactions on Circuits and Systems for Video Technology, 14(5) 682-692
[2] Siarohin A, Lathuilière S, Tulyakov S, Ricci E, Sebe N 2019 First order motion model for image animation. In: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019
[3] Pasini M 2019 MelGAN-VC: Voice Conversion and Audio Style Transfer on arbitrarily long samples using Spectrograms arXiv preprint arXiv:1910.03713
[4] Ravikumar S., Davidson C., Kit D., Campbell N., Benedetti L., and Cosker D. 2016 Reading between the dots: combining 3D markers and FACS classification for high-quality blendshape facial animation. Proc. Graphics Interface 2016 143–151
[5] Wen S P, Liu W W, Yang Y, Zeng Z G, and Huang T W 2019 Generating realistic videos from keyframes with concatenated GANs IEEE Transactions on Circuits and Systems for Video Technology 29(8) 2337-48
[6] Bagautdinov T, Wu C, Saragih J, Fua P, Sheik Y 2018 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3877-3886
[7] Goodfellow J I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A C, and Bengio Y 2014 Generative adversarial networks CoRR, abs/1406 2661

[8] Brock A, Donahue J, and Simonyan K 2019 Large scale GAN training for high fidelity natural image synthesis In 7th International Conference on Learning Representations, ICLR 2019 (New Orleans, LA, USA, May 6-9, 2019)

[9] Donahue C, McAuley J, and Puckette M 2018 Adversarial audio synthesis arXiv preprint arXiv:1802.04208

[10] Kameoka H, Kaneko T, Tanaka K, and Hojo N 2018 Stargan-vc: non-parallel many-to-many voice conversion using star generative adversarial networks In Proc IEEE Spoken Language Technology Workshop (SLT) pp 266–273

[11] Johnson J., Alahi A., Fei-Fei L. 2016 Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9906. Springer, Cham. https://doi.org/10.1007/978-3-319-46475-6_43

[12] Amodio A and Krishnaswamy S 2019 Travelgan: Image-to-image translation by transformation vector learning CoRR, abs/1902.09631

[13] Taigman Y, Polyak A, and Wolf L 2017 Unsupervised cross-domain image generation In 5th International Conference on Learning Representations (ICLR 2017) (Toulon, France, April 24-26, 2017)

[14] Wen S P, Xiao S X, Yang Y, Yan Z, Zeng Z G, and Huang T W 2019 Adjusting the learning rate of memristor-based multilayer neural networks via fuzzy method IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems 38(6) 1084-94