INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Speech Driven Talking Head Generation via Attentional Landmarks Based Representation

Wentao Wang[1], *Yan Wang*[1*], *Jianqing Sun*[2], *Qingsong Liu*[2], *Jiaen Liang*[2], *Teng Li*[1]

[1]School of Electrical Engineering and Automation, Anhui University, Hefei, China
[2]Unisound Intelligent Technology Co., Ltd, China

[*]ywanglt@gmail.com

Abstract

Previous talking head generation methods mostly focus on frontal face synthesis while neglecting natural person head motion. In this paper, a generative adversarial network (GAN) based method is proposed to generate talking head video with not only high quality facial appearance, accurate lip movement, but also natural head motion. To this aim, the facial landmarks are detected and used to represent lip motion and head pose, and the conversions from speech to these middle level representations are learned separately through Convolutional Neural Networks (CNN) with wingloss. The Gated Recurrent Unit (GRU) is adopted to regularize the sequential transition. The representations for different factors of talking head are jointly feeded to a Generative Adversarial Network (GAN) based model with an attentional mechanism to synthesize the talking video.Extensive experiments on the benchmark dataset as well as our own collected dataset validate that the propose method can yield talking videos with natural head motions, and the performance is superior to state-of-the-art talking face generation methods.

Index Terms: talking head generation, speech driven, head motion

1. Introduction

Talking head generation is to synthesize human face video from any given input speech, so that the generated video is smooth and natural, and human lip movement is synchronized with the speech. Usually a human face image is given to guide the generated figure of video. It has wide applications in real life, such as virtual anchors, movie animation, teleconference, and enhancing speech comprehension while preserving privacy.

In recent years, talking face generation has been a hot research topic, and impressive results have been achieved for a specific person or arbitrary identities [1][2]. Previous works mostly focused on precise synchronization of facial movements with the input speech [3][4], or high quality videos generation [5]. Despite the great progress on feature representation for talking video generation, most works neglected natural person head motion during talking. Head motion can be an important cue for human communication, and its absence can not lead to acceptable realistic talking video, since human visual system has low tolerance to any unnaturalness in facial videos. Furthermore, learning only from calibrated frontal faces could cause facial jitter effect in the results.

However, to synthesize realistic talking video with head motion that exactly match certain speech is challenging. First, various types of visual features of talking head video such as lip movement and head motion have different characteristics in mapping to the audio feature. For example human head may move before or after the corresponding utterance. Second, audio-video synchronization is strictly required when head moving, and generating plausible facial expression with head motion is especially difficult because of the geometric and kinematic complexity of faces.

In this paper, we propose a novel and robust method to generate talking head video with natural head motion from any given speech. Motivated by the remarkable results in facial landmark tracking, we leverage landmark points as the middle representation to bridge audio and video. Features of head motion and lip movement are computed separately. Cross modality alignment and mapping from speech to visual representations of multiple factors are learned jointly with CNN and GRU.

Based on the above mentioned visual representations driven by speech, as well as a given target image to guide the output background and person identity, a GAN based network is adopted to synthesize the talking head video. An attentional mechanism is proposed in the learning process to naturally integrate different cues to produce high quality video frames.

We conduct extensive experimental evaluations on a benchmark dataset, as well as our own collected videos containing full head motions of human talking. The results of quantitative metrics, generated video quality and audio visual synchronization show the superiority of our proposed algorithm against existing state-of-the-arts.

The rest of this paper is organized as follows. Section 2 provides a description of the related works. In Section 3, we present our proposed method and Section 4 gives our experimental results as well as the discussion. Section 5 concludes the paper.

2. Related Works

Previous related works mostly concerned the synthesis of face animation. In this section, we briefly review the related works of talking face generation.

Talking face generation is originally closely related to computer graphics [2][6]. Earlier works mainly targeted at modeling the mapping from audio to mouth action. Karras et al. [2] proposed to transfer the input speech to 3D facial space, and deep learning is also introduced in [3] to convert speech directly to the JALI model representation for facial actions [7]. Speech2Vid [1] treated audio-visual pair as temporal-independent image generation problem, but the generated videos look unnatural since only mouths move with audio.

Recently with the growth in CNN and GAN, end-to-end synthesis from speech to face video becomes more popular [8]. Suwajanakorn et al. [9] used a time delayed Long Short Term Memory (LSTM) to generate facial key points synced to the audio and use another network to generate the video frames conditioned on the key points. [4] utilized facial landmarks to assist generation, and exploited a dynamically adjustable pixel-wise loss along with an attention mechanism. [10] proposed an ad-

http://dx.doi.org/10.21437/Interspeech.2020-2304

versarial learning method to disentangle the different information for one image during generation.

As mentioned above, previous works mainly focused on frontal facial movements generation, and barely considered the head motion. The recent work of [5] tried to extract head poses via 3D face model reconstruction from 2D face images, and integrate this information into video synthesis. However, the 3D reconstruction model is difficult to learn accurately, and its coefficients for pose representation are not rich enough. In order to generate head motion-tailored talking video, we proposed to disentangle the head action and lip movement information explicitly, and integrate them naturally in our model.

3. The Proposed Method

This section introduces the details of our proposed method. Figure 1 gives the overall framework, which can be divided into two main parts: STL net for speech to landmark points conversion and LTV net for landmarks to video conversion. The whole procedure can be formulated by:

$$R : t = \Phi(\mathrm{A} : t) \tag{1}$$

$$V : t = \Psi(R : t, P) \tag{2}$$

$$V : t = \Psi(\Phi(A : t), P) \tag{3}$$

where $A : t$ represents the speech sequence, P is the given reference face image, $R : t$ is the generated landmark points sequence based xy coordinates, $V : t$ is the generated image sequence.

Through two parallel networks, the STL net converts the speech signal $A : t$ to the moving sequence of the corresponding facial contour points and the mouth contour points, which are then input to LTV net together with the reference image P to generate the resulted talking head video. In the following we introduce the two components in detail.

3.1. Speech to Landmark Representations

The original landmark points are obtained with DLIB [11] which can yield reliable 68 landmark points for talking head with pose variations in a certain range. To normalize the scale and location, we use a template face image, and extract its landmark points as the reference. The 20 points of the mouth contour are extracted and calibrated according to the positions of three template landmark points: the nose tip and the two mouth corners. It is then used to represent the human lip action which needs to be exactly synchronized with speech content. The head motion is represented by the original face contour points without calibration, so that we can keep natural temporal transition.

The computing process of STL net can be formulated by:

$$R_{\text{mouth}} : t = \Phi\left(GRU\left(f_{\text{a1}}(A : t), R_{\text{template\_mouth}}\right)\right) \tag{4}$$

$$R_{\text{head}} : t = \Psi\left(GRU\left(f_{\text{a2}}(A : t)\right), R_{\text{template\_head}}\right) \tag{5}$$

$$R_{\text{full}} : t = R_{\text{mouth}} : t \oplus R_{\text{head}} : t \tag{6}$$

where $A : t$ denotes the speech sequence, $R_{\text{template\_mouth}}$ is the template mouth landmark points, $R_{\text{template\_head}}$ is the template head landmark points, $R_{\text{mouth}} : t$ is the mouth sequential representation driven by speech. $R_{\text{head}} : t$ is the landmarks based representation for head pose, and $R_{\text{full}} : t$ is the full representation containing both head motion and lip movement information.

We design two parallel CNN based sub-networks, i.e. lip branch and head branch, to convert input speech to representations of lip animation and head motion of talking person. We basically use CNN networks to encode the speech representation and convert it to the landmark based representations with the fully connected networks. Through the network the one-dimensional speech is converted to two-dimensional features, which are then input to the GRU network together with the template representation, to predict the corresponding landmark points sequence.

In the lip branch network, the speech features are encoded to 512-dimension vectors by the network of convolution, Batch-Normalization and ReLU followed by full connection and the template landmark points are encoded to 256-dimension features by full connection. Then, we concatenate the two features as the input and encode it into a 42-dimension feature by GRU followed by full connection. The head branch network structure is similar, and it takes 100-dimension template landmark features as input and obtains 100-dimension output.

To train the model effectively we adopt the wingloss function, defined in Equation (7). Since for the key points based representation, conventional Mean Square Error (MSE) loss tended to reach saturation in a few epoches, which hindered the fully optimization the model. We experimentally validated this, and the wingloss could guarantee a stable procedure in the network training.

$$\text{wingloss}(x) = \begin{cases} w \ln(1 + |x|/\epsilon) & \text{if } |x| < w \\ |x| - C & \text{otherwise} \end{cases} \tag{7}$$

where w is empirically set to 10.0 and ϵ is 2.0, c is based on w and ϵ, x represents the difference of the predict landmarks based representation and its corresponding groud truth.

3.2. GAN Based Video Synthesis with Attention

The LTV net can be formulated by the following Equation (8) :

$$V : t = G\left((R : t) \otimes f_{\text{img}}, f_{\text{img}}\right) \tag{8}$$

where $R : t$ represents the sequential landmarks representation yielded by the previous mentioned STL net. $V : t$ is the frame sequence to be generated, and f_{img} denotes the reference face image to guide the synthesized person figure, as well as the background. The operation of $\otimes$ means the landmark points are jointly computed with the reference image to produce the head part of expecting results. Firstly the generated two sets of landmark points from previously mentioned STL net are integrated to one representation by aligning the three points, i.e., nose lip and two mouth corners. The whole landmarks based representation are merged with the reference image to get the face region information for video synthesis.

In the proposed GAN based video generation, several factors need to be considered simultaneously: head and lip motions, facial appearance and the background. These factors have different properties. For example, background information is independent of the representation for motions, but it directly determines the resulted picture quality of the video. The generator of GAN is based on UNet network. To integrate multiple cues naturally while preserving important parts well, we design an attention mechanism to consider the head information with motion separately from the background information.

As shown in Figure 2, first, we encode the landmark points and the template image. Then, the Sigmoid function is used to operate on the landmark points, which are then multiplied by the corresponding reference image, to obtain the face information, as formulated in Equation (9). Later the background part of the input reference image is introduced through another branch to

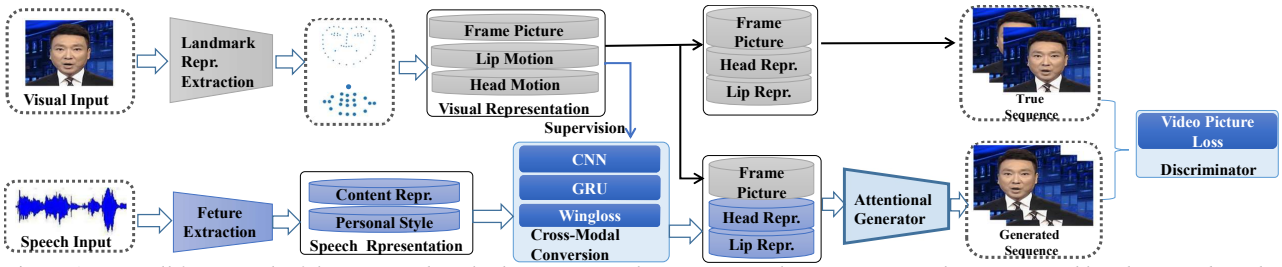

Figure 1: *Overall framework of the proposed method. We separately extract visual representations lip motion and head motion based on landmark points. The speech sequence is converted to these representations with CNN and GRU models. The multiple representations are jointly input to the GAN base network with attentional mechanism to generate the talking head video, which should satisfy both video naturalness and audio-visual synchronization.*

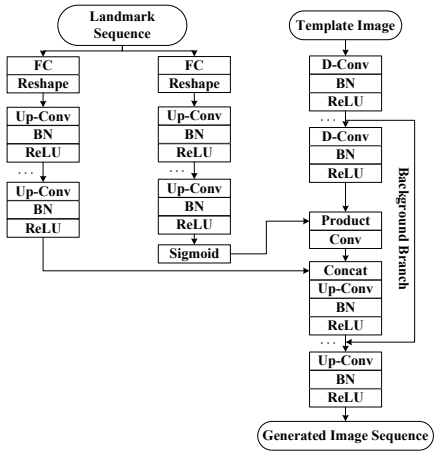

Figure 2: *Illustration of the multiple cues integration with attentional mechanism.*

integrated together with the facial part to produce high quality video frames. The attentional mechanism can be formulated by:

$$f_{\text{face}} = f_{\text{img}} \otimes \text{Sigmoid}(f(L:t)) \tag{9}$$

Where f_{img} denotes the reference template face image, $L:t$ denotes the landmark sequence.

To design the discriminator for the GAN network training, we jointly consider the factors of lip synchronization with speech, head motion naturalness, and video quality to design the loss function. Modules in our framework are jointly trained by optimizing the following composite loss:

$$L_{\text{GAN}} = L_d + L_g + L_{mse} \tag{10}$$

where L_d is the Binary Cross Entropy (BCE) loss of discriminator and L_g is generator of GAN respectively, and L_{mse} represents the image MSE loss.

4. Experiments

4.1. Dataset and Settings

The experiments are conducted on two datasets. One is composed of our own collected videos recording realistic talking contents from three persons, and the other is the widely used benchmark GRID dataset [13]. Since available public datasets for talking face generation only contain frontal facial information without head motions, we collected our own dataset with the upper part of talking person body, containing natural facial movements and head motions. The data is divided to training and test sets at the ratio of $9:1$. For speech, we extract the Mel Frequency Cepstrum Coefficient (MFCC) features at the sampling rate of 16000Hz with the windows size of 10ms. The video frames are cropped to the size of 300×300 around the talking head in preprocessing.

Our network is implemented with the Pytorch1.0 library. During training, we used Adam optimizer and a fixed learning rate of 10^{-4}. All network layers were initialized in a random way. All the training procedure was taken with a TITAN V100 GPU. We use common reconstruction metrics such as Peak Signal to Noise Ratio (PSNR) and Structure Similarity Index Measure (SSIM) [14] to evaluate the quality of the synthesized talking faces. Furthermore, we use Landmark Distance (LMD) [12] to evaluate the accuracy of the generated lip by calculating the landmark distance between the generated video and the ground truth. The lower LMD, the better of the generation.

Table 1: *The comparison of SSIM values of generated videos from different methods on our own dataset.*

Method	SSIM
Chen [12]	0.74
ATVG [4]	0.80
The proposed (w/o attention)	0.83
The proposed (w attention)	**0.88**

4.2. Evaluation on Our Own Dataset

In order to quantify the effect of each component of our method, we conduct ablation studied on our own datasets. Figure 3 shows some sample results of the generated head landmark points from the input speech and a guide image for scale calibration. Compared with landmark points extracted from the corresponding ground-truth videos, we can see that the mouth shape is almost the same while the head pose has variations. This coincides with our expectation that lip movement should be precisely aligned with speech but there could be some randomness for the head motion.

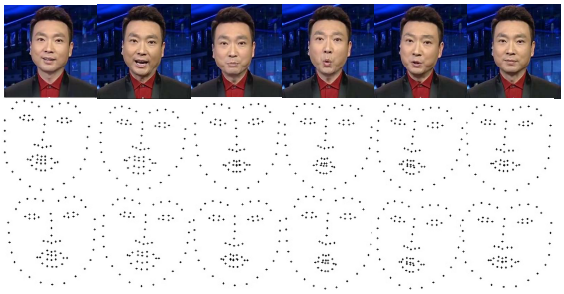

Figure 3: *Sampling results of the STL net of the proposed method. The first line shows the ground truth video corresponding to the input speech, the second line shows the landmark points from the ground truth video, and the third line shows the predicted landmark points driven by speech.*

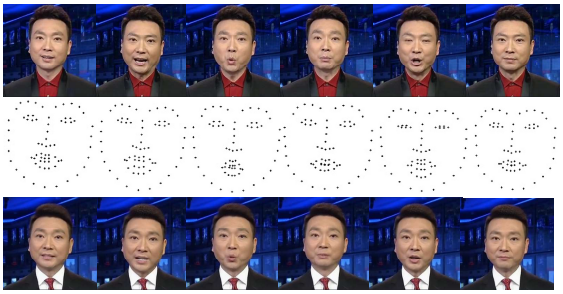

Figure 4: *The output of LTV network. The first line is the original image, the second line is the key point corresponding to the original image, and the third line are the image synthesized by the key points.*

Figure 4 gives some results of the landmark points to video step. The generated talking head frames using our proposed method are compared with the ground-truth frame corresponding to the input landmark points. It can be seen that the head pose and facial expression in generated frames are very close to the ground-truth, while the background has variations including the human hair style, clothes, tie.

Table 1 gives an overall evaluation of the proposed method on the self-collected dataset. We also compared the previous talking face generation methods of Chen [12] and ATVG [4] with our proposed method using the SSIM metric on the synthetic video. It can be seen that our attention network achieves the best performance. Here we also evaluate the impact of the proposed attentional mechanism in the method by comparison. It can be seen that the synthesized video quality can be much improved by introducing attention to integrate the different cues in the generative network. Figure 5 shows examples of the output synthesized video by our proposed method. We can observe in detail that the generated frames are natural with high quality.

4.3. Evaluation on GRID dataset

On the benchmark dataset our propsed model is compared with recent state-of-the-art talking face generation methods, including Chen [12], ATVG [4], and Chung [1]. To evaluate the impact of different components in the proposed approach, we also conduct ablation experiments with such designations: STL network without considering head motion and lip movement separately and generative network without attention mechanism. Table 2 shows the quantitative results of the proposed method with high PSNR, SSIM and low LMD, indicating the good quality of the generated face video frames. The baseline is a straightforward model without separating multiple cues in STL net, nor it adopted the attention mechanism in video generation. The proposed (w/o Splitnet) represents without separating multiple cues in STL net only, and the proposed (w/o Attention) represents without the attention mechanism. Though our PSNR is not the highest, the proposed method yielded the best effect in terms of LMD and SSIM. It validates effectiveness of the components of the proposed method for speech driven visual representation and talking video generation.

Figure 5: *The output of the proposed method. The 1st and 3rd lines show the ground truth videos, and the 2nd and 4th lines show the synthesized videos each driven by the speech of its upside video.*

Table 2: *Quantitative evaluation on GRID dataset, for fair comparison, we generate the full face and do not apply any post process.*

Method	LMD	SSIM	PSNR
Chen[12]	1.59	0.76	29.33
Wiles[15]	1.48	0.80	29.39
Chung[1]	1.44	0.79	29.87
Baseline	1.82	0.77	28.78
ATVGnet[4]	1.29	0.83	**32.15**
The proposed(w/o Splitnet)	1.07	0.84	29.23
The proposed(w/o Attention)	1.33	0.89	29.68
The proposed	**1.01**	**0.91**	29.73

5. Conclusion

In this paper, we propose a GAN based network based on the attentional multiple representations to synthesize talking head video from given speech. Head motion and lip movement of talking person are separately represented through landmark points representations, and the attention mechanism is used to integrate the different part features into the generative network to synthesize natural results. Experimental results showed that the proposed method can generate not only high quality facial appearance, accurate lip movement, but also natural head motion. For future work, we could consider upper body movements including gesture information in the talking person video generation.

6. References

[1] J. S. Chung, A. Jamaludin, and A. Zisserman, "You said that?" *arXiv preprint arXiv:1705.02966*, 2017.

[2] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–12, 2017.

[3] Y. Zhou, Z. Xu, C. Landreth, E. Kalogerakis, S. Maji, and K. Singh, "Visemenet: Audio-driven animator-centric speech animation," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–10, 2018.

[4] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7832–7841.

[5] R. Yi, Z. Ye, J. Zhang, H. Bao, and Y.-J. Liu, "Audio-driven talking face video generation with natural head pose," *arXiv preprint arXiv:2002.10137*, 2020.

[6] S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. Hodgins, and I. Matthews, "A deep learning approach for generalized speech animation," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–11, 2017.

[7] P. Edwards, C. Landreth, E. Fiume, and K. Singh, "Jali: an animator-centric viseme model for expressive lip synchronization," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 1–11, 2016.

[8] K. Vougioukas, S. Petridis, and M. Pantic, "End-to-end speech-driven realistic facial animation with temporal gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 37–40.

[9] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–13, 2017.

[10] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, "Talking face generation by adversarially disentangled audio-visual representation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 9299–9306.

[11] D. King, "Dlib c++ library," *Access on: http://dlib. net*, 2012.

[12] L. Chen, Z. Li, R. K Maddox, Z. Duan, and C. Xu, "Lip movements generation at a glance," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 520–535.

[13] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[14] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[15] O. Wiles, A. Sophia Koepke, and A. Zisserman, "X2face: A network for controlling face generation using images, audio, and pose codes," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 670–686.