Hindawi
Complexity
Volume 2020, Article ID 6629634, 8 pages
https://doi.org/10.1155/2020/6629634

Research Article

Realistic Speech-Driven Talking Video Generation with Personalized Pose

Xu Zhang and Liguo Weng

Jiangsu Key Laboratory of Big Data Analysis Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China

Correspondence should be addressed to Liguo Weng; liguoweng@hotmail.com

Received 30 October 2020; Revised 18 November 2020; Accepted 8 December 2020; Published 29 December 2020

Academic Editor: Zhijie Wang

In this work, we propose a method to transform a speaker's speech information into a target character's talking video; the method could make the mouth shape synchronization, expression, and body posture more realistic in the synthesized speaker video. This is a challenging task because changes of mouth shape and posture are coupled with audio semantic information. The model training is difficult to converge, and the model effect is unstable in complex scenes. Existing speech-driven speaker methods cannot solve this problem well. The method proposed in this paper first generates the sequence of key points of the speaker's face and body postures from the audio signal in real time and then visualizes these key points as a series of two-dimensional skeleton images. Subsequently, we generate the final real speaker video through the video generation network. We take a random sampling of audio clips, encode audio contents and temporal correlations using a more effective network structure, and optimize and iterate network outputs using differential loss and attitude perception loss, so as to obtain a smoother pose key-point sequence and better performance. In addition, by inserting a specified action frame into the synthesized human pose sequence window, action poses of the synthesized speaker are enriched, making the synthesis effect more realistic and natural. Then, the final speaker video is generated by the obtained gesture key points through the video generation network. In order to generate realistic and high-resolution pose detail videos, we insert a local attention mechanism into the key point network of the generated pose sequence and give higher attention to the local details of the characters through spatial weight masks. In order to verify the effectiveness of the proposed method, we used the objective evaluation index NME and user subjective evaluation methods, respectively. Experiment results showed that our method could vividly use audio contentsto generate corresponding speaker videos, and its lip-matching accuracy and expression postures are better than those of previous work. Compared with existing methods in the NME index and user subjective evaluation, our method showed better results.

1. Introduction

The task of a speech-driven speaker video refers to a technology that automatically generates a video of a corresponding character's speech through a computer-based audio information. The content of the talking must be consistent with the character's pose in the video. Traditional speech-driven talking video requires professional equipments and operators to perform character modeling, which is usually very expensive for custom use. In recent years, with the successful application of deep neural networks, data-driven speech and video synthesis methods have been proposed. These methods often require the use of a large amount of high-quality audio and video data, and the production process is complex, but the synthesized speaker's mouth posture matching effect is poor.

The current mainstream methods mainly focus on facial speaker synthesis and do less work on body postures and facial expressions. Specifically, the existing methods [1, 2] input the speaker's voice information into the recurrent neural network to obtain 3D face model parameters, then map the fitted 3D face model to 2D key points as inputs of the video synthesis module, and then output corresponding speaker pictures through the video synthesis model. Due to

the weak representation ability of the 3D face model parameter network, the key point error obtained from the 3D face model conversion is larger, the 3D face model needs to be used as an intermediate state for conversion, resulting in a complicated overall process. Eskimez et al. [3] converted the facial key points into the average face space in the dataset to remove ID features and simplified the task. Although the key point indicators obtained from the network output are relatively low, the posture expressionsare very monotonous and rigid, and hence , the synthesized speaker video is not realistic enough.

As mentioned above, the matching effect of existing speech-driven speaker methods is not ideal, and the synthesized speaker video has a jitter phenomenon. In order to solve the above problems, this paper proposes a method to convert the speaker's voice information into the target person's talking video. We use the Dilated Depthwise Separable Residual (DDSR) unit to encode the audio features [4, 5], and then use the GRU network layer [6] to learn the temporal features and constrain the network outputs using content loss functions. Through this network structure, the audio content and temporal correlation information are effectively encoded simultaneously, the facial key point index of the model output is lowered, and the mouth shapes and postures of the synthesized speaker video are matched with audio contents better, plus, the synthesized speaker video is more natural and realistic. In the process of training and testing, we insert the specified pose sequence frame into the pose sequence, which makes the audio conversion to the speaker's mouth shape and posture more natural and vivid. In order to enrich the speaker's detailed texture, we introduce a local attention mechanism in the key point network and add spatial weights to the face, fingers, and other parts of the character to get higher attentions.

Finally, in order to better evaluate our system, we used high-resolution and frame rate (FPS) cameras to create a dataset containing audio and video for multiple targets reading selected articles. Compared with the existing methods, our method produces better visual perception. In Figure 1, we show some images of our synthesized speaker video.

In summary, the contributions of our work are

(1) We use a novel Dilated Depthwise Separable Residual (DDSR) unit. This network structure can effectively represent the audio content and temporal correlation, and the facial key point index of the model output is lower. At the same time, the network model is used to model the key points of the face and human posture, respectively. After preprocessing , it uses the loss function to optimize iteratively. The results show that the face details and human postures are better.

(2) We use the first-order differential loss function and the pose perception loss function [7, 8] to optimize the model. Among them, the first-order differential loss function can smooth the pose of the front and rear frames, and the pose perception loss function uses the spatiotemporal graph to form a hierarchical representation of the pose sequence, so as to constrain the temporal-spatial information output from the network.

(3) We establish a pose keypoint map to add richer poses and expressions to the generated human poses. In addition, we also provide a method to convert the pose in the existing sequence window into the corresponding keyframe pose sequence.

2. Related Work

Given a speaker's audio information, the generation of the corresponding person speaking video has attracted many researchers' interests. Earlier works mainly used the Hidden Markov model (HMM) to generate corresponding relationships between speech and facial motions [9–14]. Among them, Brand [15] proposed voice puppetry as an HMM-based method for generating conversation faces driven only by voice signals. In another study, Cosker et al. [10, 11] proposed a hierarchical model that can animate the subregions of the face independently of speech and merge them into a complete face video.

In recent years, with successful applications of deep neural networks, the related work of speech-driven speaker based on deep learning method has been proposed. Among them, Suwajanakorn et al. [16] designed an LSTM network to directly generate the target identity talking face video from the audio. However, this method needs to record a large number of facial videos with specific target identities, it limits its application in many scenarios. Linsen et al. converted audio information into the 3D face model parameter space and then the fitted 3D face model to 2D facial key points. Their network uses several layers of recurrent neural networks as encoding, and the network feature learning ability is relatively weak. The facial key points obtained by the conversion of the 3D face model have a large error, and the 3D face model needs to be used as an intermediate state for conversion. This leads to the complexity of the overall process.

In addition, including the single-stage method of direct conversion of audio to speaker video space, many researchers divide the task of speech generation into two stages. Usually, the key point information only responds to the voice content information. Pham et al. [17] first used the LSTM network to map voice features to 3D deformable shapes and rotation parameters and finally generated 3D animated faces in real time based on the predicted parameters. In literature [18], they further improved this method, replacing speech features with original waveforms as inputs and the LSTM network with a convolutional structure. However, compared with the speech-generated gesture keypoint network in our method, their method is less intuitive in shapes and rotation parameters, and the mapping from these parameters to specific gestures or facial expressions is not clear. In another related work, the key points of the face that they generated are for a standardized average face, rather than for a specific target identity. Although this helps to eliminate factors that are not directly related to voice, the predicted sequence of key points for the

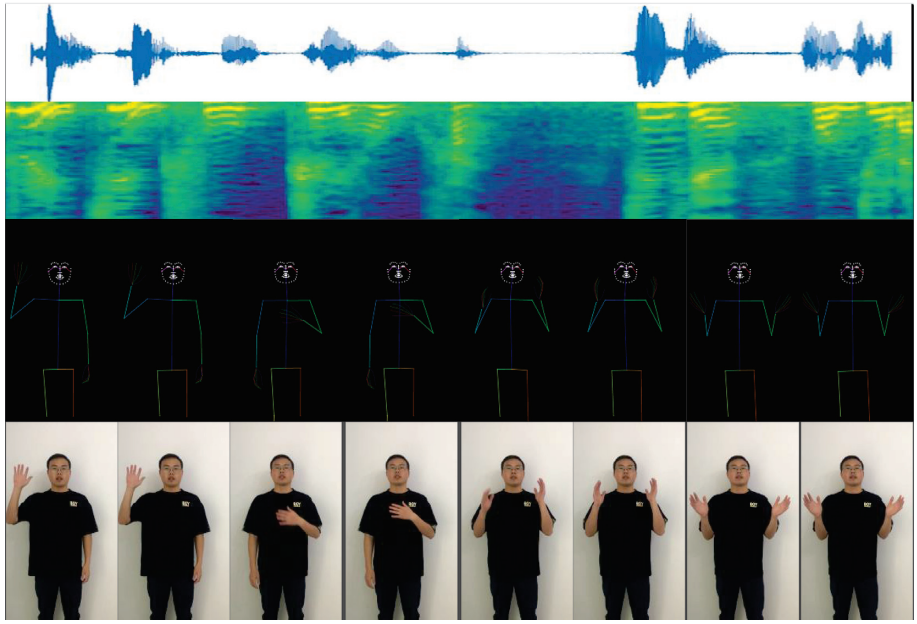

Figure 1: Speech-driven talking video: a given piece of audio/text can be used to drive the video of the specified speaker.

posture is unnatural. [19] An extended complex human motion synthesis method based on autotuning recurrent network is proposed. They can simulate more complex movements, including dances or martial arts. In the second stage of work, most methods use vid2vid [20] to enhance the time consistency between adjacent frames. Shysheya et al. [21] proposed a method to generate realistic videos from skeleton sequences without establishing a 3D model. Our method also uses the vid2vid network to synthesize the final speaker video from the posture skeleton picture and obtains better results. For the detailed texture information of the face and hands, we use separate discriminators to optimize these parts in vid2vid.

Our method expands the data of random audio samplings and uses a more effective network structure to learn audio contents and timing correlations. The loss function uses the first-order differential loss and poses perception loss to optimize output pose timing stability and matching accuracy. At the same time, the keyword wake-up technology is used to convert the generated sequence poses into specified action poses. A large number of experimental results show that our method generates a natural and realistic speaker video for talking audio, and its lip matching and expression posture are more expressive than those of the previous work.

3. Methods

In this section, we mainly introduce different modules of the network. The overall network structure is shown in Figure 2. In our approach, the input information can be either audio or text. When the audio information is used as the speaker synthesis network input, we convert the audio data into log-mel features; the aud2kps network is used to get the human body postures and facial key points. Using the Dictionary Building and Key Pose Insertion method to insert a specified action frame into the generated key point sequence, the synthesis effect is more natural and realistic, and then the output key points of facial and human posture are visualized as a series of 2D skeleton images, and these 2D skeleton images are further fed into the Vid2vid generation network to generate the final talking images. When the input is text information, it is necessary to use the acoustic model to convert the text information to obtain a unified log-mel feature as the input of the Aud2Kps network. The following steps are the same as the audio signal input process. The text-to-speech method (TTS) is currently very mature and commercialized, and we use the open source tactron2 [22] to complete the text conversion results which we want. In the following sections, we describe each module of our architecture.

3.1. Pose Keypoints. In the process of audio-video conversion, we use the key points of human body posture as the intermediate state representation so that the span of the two spatial features will not be too large. Compared with using the 3D human body model as the intermediate state representation, it is more convenient and universal in the process of training and reasoning. We use the open source method OpenPose [23, 24] to obtain the key points of the human body posture. These key points include a total of 137 position coordinate information of the body, feet, hands, and faces. Firstly, we construct these 2D key points and audio information into a content sequence and then train the Aud2Kps network to generate 2D coordinates corresponding to the posture key points from the audio speech information.

3.2. Audio to Keypoints (Aud2Kps). As shown in Figure 2, our Aud2Kps network takes log-mel spectrogram as the input. $[x_0, x_1, \ldots, x_n,]$ is the input vector of audio/text

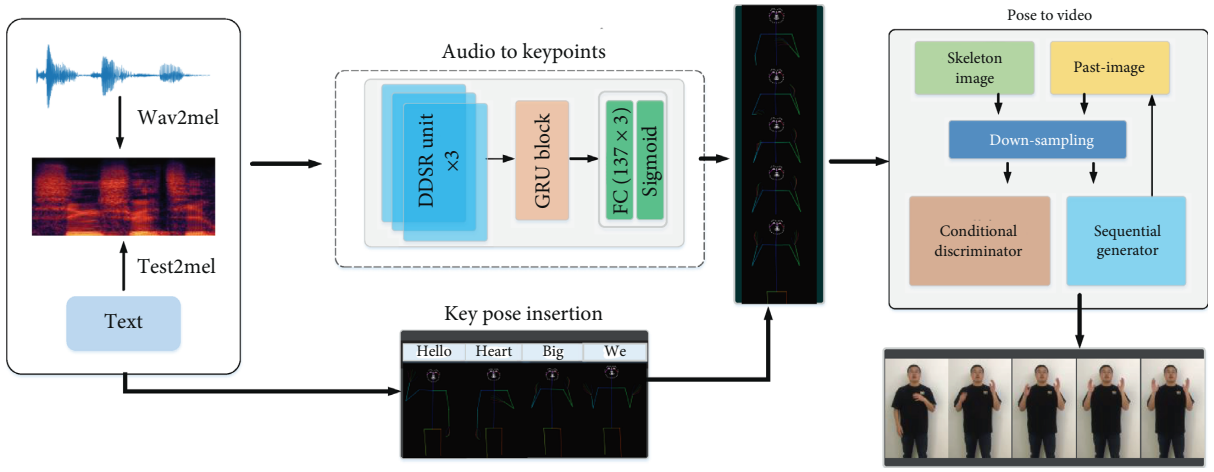

FIGURE 2: Pipeline of our method: the input information can be audio or text. When the audio information is used as the speaker synthesis network input, we convert the audio data into log-mel features and then input the Aud2Kps model to get the pose key points. When the input is text information, it is necessary to use the acoustic model to convert the text information to the log-mel feature as the input of the Aud2Kps network. The following steps are the same as the audio signal input process.

encoding and $[y_0, y_1, \ldots, y_n,]$ is the output open-pose key point vector. The log-mel spectrum feature extracted from audio [25] is a set of 80-dimensional vectors. We designed a DDSR unit to encode the semantic content of features, then input the GRU model to learn the timing features, and finally input the full connection layer and sigmoid activation function to obtain the key point information of the face and human body posture. Our network structure effectively characterizes the audio content information and the correlations between the front and rear time series so that the NME index of the facial key points output by the model is lower. When Aud2Kps maps the audio sequence to the pose sequence, since different parts of the human body have different scales, we need to give them different weights. Therefore, for the body, hands, facial contours, and mouth positions, we set the attention weightsas 1, 10, 50, and 100, respectively. We also use the first-order differential loss between two consecutive poses to ensure that the output pose key points are more smooth and natural.

The MSE loss function L_{MSE} is given by

$$L_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}\left\|y_i - \hat{y}_i\right\|_{l_2}. \tag{1}$$

The first-order temporal differential loss L is given by

$$L_{\text{First-order}} = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{y}_i - \hat{y}_{i-1}\right\|_{l_1}. \tag{2}$$

At the same time, we use a pose-perception loss function to calculate the content loss between the real and generated pose key points. In most content loss, the VGG network is used as the feature extractor [26, 27], the pose perception loss function uses ST-GCN as the feature extractor of the perception loss function, and the hierarchical representation of the skeleton sequence is formed by using the space-time graph and can be obtained from automatically learn spatial and temporal patterns in the data. We use a dilated residual block in each DDSR unit [28] so that each subsequent layer has a long time span, and the receptive field of the convolutional layer after expansion increases exponentially with the number of layers. This method can effectively increase the sensing receptive field of each output time step and obtain a better long-range correlation. The implementation details of the DDSR unit are shown in Figure 3.

Given a pretrained GCNnetwork φ, we define a collection of layers φ as φ_l. For a training pair (P, M), where P is the ground truth skeleton sequence and M is the corresponding piece of audio, our perceptual loss is

$$L_{\text{Perceptual}} = \sum_{l=1}^{N}\beta_l\left\|\varphi_l(P) - \varphi_l(G(M))\right\|_{l_1}. \tag{3}$$

Here, G is the first-stage Aud2Kps network in our framework. The hyperparameters β_i balance the contribution of each layer l to the loss.

Since the text input will not affect the model efficiency even there is difference in voice characteristics between people, the text input will make the network model more general. Similar to the process of using audio-training Aud2Kps, we convert the text segmentation into phonemes and then use the acoustic model through feature encoding to generate log-mel features as the input of the subsequent speaker synthesis model. We use the open source tacotron2 model to convert the text into a log-mel feature. The following process is the same as the process of audio-to-keypoint.

3.3. Key Pose Insertion. During the model training process, we found that although the Aud2Kps model can synchronize the audio and video content of the speaker very well, the generated character action sequence is too monotonous. This is mainly because the character action sequence is the same at most times in the training set, and the action sequence with posture change is very sparse in the whole

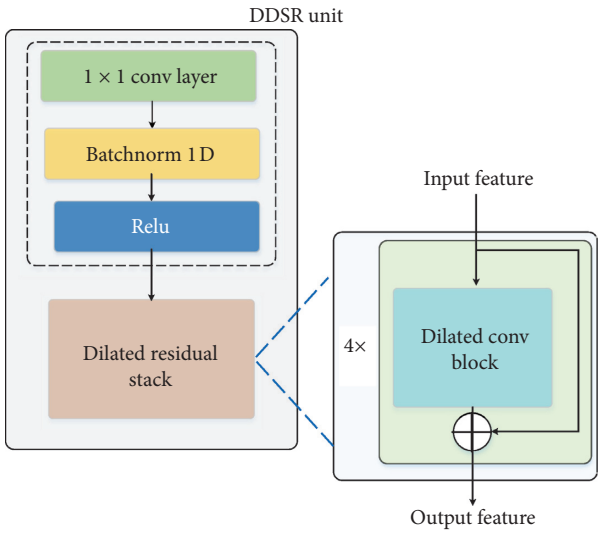

Figure 3: Dilated Depthwise Separable Residual (DDSR) unit network.

training set [29]. In order to make the gesture actions in the synthesized speaker video more expressive and diverse, we designed a gesture sequence dictionary. When the specified keywords appear in the audio content, the corresponding window of the gesture sequence output by Aud2Kps is converted into the specified action, and the posture transformation here uses the posture transformation matrix stored in the posture sequence dictionary.

We select some posture action sequences from the recorded videos and then construct these posture sequences and the corresponding wake-up words into a posture sequence transformation dictionary (composed of transformation matrix). Once the input audio content appears in the dictionary, we will transform the existing pose sequence with a certain probability. The probability between different words may be different. In order to maintain a smooth transition to this pose, we smooth the adjacent frames.

3.4. Pose to Video. We use the vid2vid generator network to convert our generated skeleton images into corresponding speaker videos. After the key points of the human body posture are obtained from the Aud2Kps network, they are visualized as a series of 2D skeleton images, and these 2D images are further fed into the Vid2vid generator network [20] to synthesize the final speaker video. In our network structure, different positions of the human body pay attention to different degrees of importance and people tend to pay more attention to the part of the face and hands. In order to make the vid2vid network pay more attention to the detail texture synthesis of face and hands, we use a separate discriminator network to train the models of face and hand regions to ensure that the discriminator pays more attention to the generated facial and hand details.

4. Experiments

4.1. TalkingPose Dataset. Our audio and video data can be from related speeches or broadcast videos on websites. However, most of the video resources on websites are shot at different times with change of character decorations and clothing styles, which increases uncontrollable factors of samples and increases the difficulty of training. Therefore, we specify speakersto perform audio and video recording. Our speakers read different themes and scripts, and the entire recording time of audio and video is about 2 hours. The video resolution is 1920×1080, and the speed is 30 frames per second.

After recording the video data, the audio data can be directly separated from the corresponding video data. We sample audio data with a sampling rate of 16 kHZ and convert them into log-mel features as the network input. Since audio may have different volume levels, we first normalize its volume through RMS-based normalization [29]. Then, through sparse fast Fourier transform (sfft), the audio is converted from time-domain representation to frequency-domain representation. The value on each frequency represents the energy of the frame of speech signal at the current frequency, and a set of multiple triangular filters are used. The linear spectrum after sfft is processed to obtain 80-dimensional low-dimensional features to simulate the suppression of high-frequency signals by human ears. This method is widely used in speech feature extraction. We use random sampling strategies to expand the dataset for the audio features in the same segment, and the log-mel feature and the posture key point sequence are 1 : 4 as the model input. Figure 2 is a partial example of our dataset.

4.2. Implementation Details. All the models are trained on 8 Nvidia GeForce GTX 1080 Ti GPUs. For the first stage of the Aud2Kps model in our framework, the model is implemented in PyTorch [24] and takes approximately one day to train for 500 epochs. For the hyperparameters, the dimensions of the output channels of the three DDSR units are set to [128, 256, 512], the number of hidden nodes in the GRU timing network is set to 256, and the number of nodes in the final fully connected layer of the network is set to the number of OpenPose parameters 137×3. For the pretraining process of ST-GCN, ST-GCN achieves 49% precision on our TalkingPose dataset. By using the Adam optimizer [30] to minimize the L_2 norm loss of key points in Pytorch, we ensure that the audio features are effectively converted to the corresponding pose key points. The network training batch size is 64, and the learning rate is 0.001. For the second stage that transfers pose to video, the Vid2vid model takes approximately seven days to train for 20 epochs, and the hyperparameters of it adopts the same as [20]. During model training, the data preprocessing part will automatically crop the original video resolution to 1024×1024. Therefore, our results are all 1024×1024 resolution.

4.3. Evaluation Metrics. The task of evaluating speech-driven talking videos is not simple because (1) there is no benchmark dataset to evaluate speech-to-human pose video; (2) the effect of people's speech-driven talking video performance is very subjective, so it is difficult to define model performance. We choose to compare our results with SoTA approaches using the user study. We compare Learning-Gesture [31], neural-voice-puppetry [32], EverybodyDance [33], and Personalized-bodyPose [29] in our user study. In the evaluation metrics of the user study, we refer to the Mean Opinion Score (MOS) [30] of the evaluation index in the text-to-speech (TTS) method [34] to measure the effectiveness of different models. Table 1 shows the MOS of user study for all methods. We get the best overall quality score over the other 4 SOTA methods.

The quantitative model predicts the effect of speaking posture. Even if the people speak the same sentence, he will not perform the same gesture at different moments. It is difficult to judge whether the speech content is correctly converted to the human body posture. However, the facial and mouth shapes of the same sentence are almost the same. Therefore, we evaluate the performance of the model through facial key points. We use the NME indicator [35] to measure the deviation degree that the audio information is converted into corresponding real facial key points. NME is widely used in facial landmark detection to evaluate the quality of models. It is calculated by the average Euclidean distance between predicted and ground truth landmarks, and then it is normalized to eliminate the impact caused by the image size inconsistency. NME for each pose is defined as

$$\mathrm{NME} = \frac{1}{L}\sum_{k=1}^{L}\frac{\left\|p_k - \widehat{p}_k\right\|_2}{d}, \tag{4}$$

where L refers to the number of landmarks, p_k and $\widehat{p}_k$ refer to the predicted and ground truth coordinates of the k_{th} landmark, respectively, and d is the normalization factor, such as the distance of eye centers (interpupil normalization, IPN) or the distance of eye corners (interocular normalization, ION).

To evaluate the effect of pose to video,we use a subjective evaluation method, a user study. In order to evaluate the final output video, we invited 100 participants on the Internet to conduct a subjective test. We showed a total of three videos to participants. Two of them are our synthetic videos, of which, one is a speaker video generated from real human audio, and the other one is a speaker video generated from TTS synthetic audio, and the remaining one is the original real speaker video. These 3 videos are randomly scrambled, and we did not tell the participants the tags behind the videos. Participants need to subjectively rate the quality of these videos, from 1 (strongly disagree) to 5 (strongly agree). The evaluation options include (1) the integrity of the human body; (2) the face of the speaker in the video is clear; (3) the posture of the person in the video looks natural and smooth; (4) the overall visual experience of the video is realistic.

As shown in Table 2, the overall score of our synthetic video four items is 3.795, and the real video is 4.365, which means that the overall effect of our proposed synthetic talking video reaches 86.94% of the real video. It is closer to the real speaker effect in terms of facial details and human body posture integrity. The video score generated by TTS is worse than the voice generation effect, and the reasons are the same as those in Table 3. The main reason is that the synthesized audio has information loss, and hence it is different from the original audio. This loss brings errors into the generated human body postures so that the visual score of the synthesized speaker video is low.

Table 1: Mean Opinion Score (MOS) of 100 participants on 4 questions. *Q*1: completeness of body. *Q*2: the face is clear. *Q*3: the body movement is correlated with audio. *Q*4:overall quality.

	*Q*1	*Q*2	*Q*3	*Q*4
Learning gesture [31]	3.414	3.659	3.914	3.308
Neural-voice-puppetry[32]	3.202	3.840	3.180	3.542
EverybodyDance [33]	3.944	3.662	3.680	3.681
Personalized-bodyPose[29]	3.894	4.011	3.383	3.762
Our method	3.901	4.083	3.526	3.778

Table 2: Mean Opinion Score (MOS) of 100 participants on 4 questions. *Q*1: completenessof body. *Q*2: the face is clear. *Q*3: the body movement is correlated with audio. *Q*4: overall quality.

	*Q*1	*Q*2	*Q*3	*Q*4
Synth.	4.14	4.37	2.92	3.75
TTS	4.10	3.80	2.58	3.39
Real	4.31	4.42	4.33	4.40

4.4. Ablation Study. We use the NME index to evaluate facial key points on the test set. As shown in Table 3, we use different time-length datasets (0.5 h, 1.0 h, 1.5 h, and 2.0 h, respectively) to train the model and observe the impact on the accuracy of pose prediction. In addition, we evaluate the audio data of text synthesis to observe the impact of sound changes on the results, use text to train and test the network, and compare the results with the audio results. Finally, we compare the training using only the GRU network with that using our network structure.

From Table 3, we can notice the following. (1) After the audio training set is increased to 1.5 h, the model benefit will not be great by increasing the dataset, but the model effect can also be improved by further increasing the amount of data on the text training set. (2) From the model indicators obtained from audio and text data, it can be seen that the effect of audio is worse than that of text , indicating that the audio conversion to the key points of the face is more accurate. (3) The audio data synthesized by text is tested on the model. The effect is not as good as the original audio mainly because the synthesized audio has information loss, and hence it is different from the original audio. (4) Using the DDSR unit network model is better than only using the GRU network structure as feature extractor. Although only using the GRU network can capture the correlation between the front and rear frames, the feature representation ability is

Table 3: Evaluation metrics used NME (%) on facial landmarks (lower is better).

	Orig.	Only-GRU	TTS-mel	Text
0.5	4.925	5.673	5.871	5.693
1.0	4.921	5.640	5.885	5.690
1.5	4.853	5.644	5.877	5.614
2.0	4.907	5.647	5.829	5.607

weak. The combination of the DDSR unit and the GRU can make up for this shortcoming.

To prove the effectiveness of our key pose insertion method, we conducted another user study. In this study, we simply presented a pair of composite videos with and without inserting key poses. Participants only need to evaluate which of the two videos is more natural and realistic. From the final user rating, it is shown that the synthesized video with gesture actionsbeing inserted into its existing posture sequence scored 81.3% and the synthesis video without the key frame poses only received 18.7% of votes. This illustrates the effectiveness of inserting pose key points to enrich speech-driven talking video synthesis.

5. Conclusion and Future Work

In this work, we propose a new method to generate realistic talking video from audio information. We sample the audio data randomly and use a more effective network structure to learn the audio content and timing correlation.We use first-order differential loss and pose perception loss to optimize the network output so that the face and pose key points obtained by audio conversion are smoother and the index performance is better. At the same time, by inserting a specified action frame into the synthesized human pose sequence window, the synthesized speaker's action posture is more natural and realistic. Our objective and subjective evaluation comparison results are very competitive over the existing methods. Our current method has good results in single-speaker scenarios. In multispeaker audio-video conversion tasks, we use TTS technology to convert speech to text to eliminate the inconvenience caused by voice ID information. In the future, we will further explore the work related to multispeaker to multitarget character video synthesis.

Data Availability

The data used to support the findings of the study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of PR China (42075130).

References

[1] 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*, vol. 36, no. 4, pp. 1–12, 2017.

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

[3] S. E. Eskimez, R. K. Maddox, C. Xu, and Z. Duan, "Generating talking face landmarks from speech," in *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation*, pp. 372–381, Guildford, UK, June 2018.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[5] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, Salt Lake City, UT, USA, June 2018.

[6] K. ChoB. Van Merriënboer et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Doha, Qatar, October 2014.

[7] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the Association for the Advancement of Artificial Intelligence*, pp. 1–9, New Orleans, LA, USA, February 2018.

[8] J. B. Estrach, P. Sprechmann, and Y. LeCun, "Super-resolution with deep convolutional sufficient statistics," in *Proceedings of the 4th International Conference on Learning Representations, ICLR*, San Juan, Puerto Rico, May 2016.

[9] K. Choi, Y. Luo, and J.-N. Hwang, "Hidden markov model inversion for audio-to-visual conversion in an mpeg-4 facial animation system," *The Journal of VLSI Signal Processing*, vol. 29, no. 1/2, pp. 51–61, 2001.

[10] D. Cosker, D. Marshall, P.L. Rosin, and Y. Hicks, "Speech driven facial animation using a hidden markov coarticulation model," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*, pp. 128–131, Cambridge, UK, August 2004.

[11] D. Cosker, D. Marshall, P. Rosin, and Y. Hicks, "Video realistic talking heads using hierarchical non-linear speech-appearance models," in *Proceedings of the Mirage 2003*, Le Chesnay-Rocquencourt, France, March 2003.

[12] L. D. Terissi and J. C. Gómez, "Audio-to-visual conversion via HMM inversion for speech-driven facial animation," in *Proceedings of the Brazilian Symposium on Artificial Intelligence*, pp. 33–42, Salvador, Brazil, October 2008.

[13] L. Xie and Z.-Q. Liu, "A coupled HMM approach to video-realistic speech animation," *Pattern Recognition*, vol. 40, no. 8, pp. 2325–2340, 2007.

[14] X. Zhang, L. Wang, G. Li, F. Seide, and F.K. Soong, "A new language independent, photo-realistic talking head driven by voice only," in *Interspeech*, pp. 2743–2747, Springer, Berlin, Germany, 2013.

[15] M. Brand, "Voice puppetry," in *Proceedings of the 26th Annual Conference on Com- Puter Graphics and Interactive Techniques*, pp. 21–28, Los Angeles, CA, USA, August 1999.

[16] 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, p. 95, 2017.

[17] H.X. Pham, S. Cheung, and V. Pavlovic, "Speech-driven 3D facial animation with implicit emotional awareness: a deep learning approach," in *Proceedings of the 1st DALCOM Workshop, CVPR*, Guildford, UK, 2017.

[18] H. X. Pham, Y. Wang, and V. Pavlovic, "End-to-end learning for 3D facial animation from speech," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pp. 361–365, Boulder, CO, USA, October 2018.

[19] Y. Zhou, Z. Li, and S. Xiao, "Auto-conditioned recurrent networks for extended complex human motion synthesis," in *Proceedings of theInternational Conference on Learning Representations*, Vancouver, Canada, May 2018.

[20] T. C. Wang, M. Y. Liu, and J. Y. Zhu, "Video-to-video synthesis," in *Proceedings of theConference on Neural Information Processing Systems (NeurIPS)*, Montreal, Canada, December 2018.

[21] A. ShysheyaE. Zakharov et al., "Textured neural avatars," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2387–2397, San Juan, PR, USA, June 2019.

[22] J. Shen, R. Pang, R. J. Weiss et al., "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proceedings of the2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783, Calgary, Canada, April 2018.

[23] Z. Cao, G. Hidalgo Martinez, T. Simon, S.-E. Wei, and Y. A. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, p. 1, 2019.

[24] S. E. Wei, V. Ramakrishna, and T. Kanade, "Convolutional pose machines," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4724–4732, Las Vegas, NV, USA, June 2016.

[25] K. Kumar, R. Kumar, and T. De Boissiere, "Melgan: generative adversarial networks for conditional waveform synthesis," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 14910–14921, Vancouver, Canada, 2019.

[26] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1511–1520, Venice, Italy, October 2017.

[27] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of the European Conference on Computer Vision*, pp. 694–711, Amsterdam, Netherlands, October 2016.

[28] S. Mehta, M. Rastegari, and A. Caspi, "Espnet: efficient spatial pyramid of dilated convolutions for semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 552–568, Munich, Germany, September 2018.

[29] M. Liao, S. Zhang, and P. Wang, "Speech2video synthesis with 3D skeleton regularization and expressive body poses," in *Proceedings of the Asian Conference on Computer Vision*, Kyoto, Japan, December 2020.

[30] R. Skerry-Ryan and E. Battenberg, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *Proceedings of the 35th International Conference on Machine Learning*, pp. 4693–4702, Stockholm Sweden, July 2018.

[31] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik, "Learning individual styles of conversational gesture," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3497–3506, Long Beach, CA, USA, June 2019.

[32] J. Thies, M. Elgharib, and A. Tewari, "Neural voice puppetry: audio-driven facial reenactment," in *Proceedings of the European Conference on Computer Vision*, pp. 716–731, Glasgow, UK, August 2020.

[33] C. Chan, S. Ginosar, T. Zhou, and A.A. Efros, "Everybody dance now," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5933–5942, Seoul, Republic of Korea, October 2019.

[34] J. M. Valin and J. Skoglund, "LPCNet: improving neural speech synthesis through linear prediction," in *Proceedings of theICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5891–5895, Brighton, UK, May 2019.

[35] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D& 3D face alignment problem?(and a dataset of 230,000 3D facial landmarks)," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1021–1030, Venice, Italy, October 2017.