Shallow Diffusion Motion Model for Talking Face Generation from Speech

Xulong Zhang[1], Jianzong Wang[1(✉)], Ning Cheng[1], Edward Xiao[2], and Jing Xiao[1]

[1] Ping An Technology (Shenzhen) Co., Ltd., Shenzhen, China
jzwang@188.com
[2] Aquinas International Academy, La Palma, USA

Abstract. Talking face generation is synthesizing a lip synchronized talking face video by inputting an arbitrary face image and audio clips. People naturally conduct spontaneous head motions to enhance their speeches while giving talks. Head motion generation from the speech is inherently difficult due to the nondeterministic mapping from speech to head motions. Most existing works map speech to motion in a deterministic way by conditioning certain styles, leading to sub-optimal results. In this paper, we decompose the speech motion into two complementary parts: pose modes and rhythmic dynamics. Accordingly, we introduce a shallow diffusion motion model (SDM) by equipping a two-stream architecture, *i.e.*, a pose mode branch for primary posture generation, and a rhythmic motion branch for rhythmic dynamics synthesis. On one hand, diverse pose modes are generated by conditional sampling in a latent space, guided by speech semantics. On the other hand, rhythmic dynamics are synced with the speech prosody. Extensive experiments demonstrate the superior performance against several baselines, in terms of fidelity, similarity, and syncing with speech.

Keywords: Talking face · Shallow diffusion · Head motion generation · Speech

1 Introduction

Head motion generation from the speech is to synthesize spontaneous head motions synchronized with input speech audio. Professional speakers are experts in utilizing such motions to effectively deliver information. This task is essential for applications such as digital avatars and social robots [11]. Notably, with this technique, amateur speakers can also generate their own "professional" talking videos, by mimicking moves from professional speakers.

With the development of deep neural networks for generation-related tasks [29,31,38,43,44], talking face can be driven by audio speech. While generating lip motions has been extensively studied in talking face generation [23], synthesizing plausible speech head motions remains an open issue. Specifically,

B. Li et al. (Eds.): APWeb-WAIM 2022, LNCS 13422, pp. 144–157, 2023.
https://doi.org/10.1007/978-3-031-25198-6_11

lip motions can be well matched with the input audio using a deterministic mapping, *i.e.*, one to one mapping from phonemes to lip shapes. However, such models can not be trivially extended to the head, due to the highly stochastic nature of head motions during a talk speech. Practically, the speech head motion is highly freedom. Even if the same person gives the same speech twice in a row, there is no guarantee that the speaker would exhibit the same head motions. Moreover, a person usually switches poses from time to time during a long talking speech. The same speech audio does not necessarily lead to a fixed form of motions, and different speeches may go well with the same motion sequence.

Most existing works treat head and lip motion generation in a similar way [14, 41, 46], *i.e.*, the head landmarks are directly inferred from the input audio via a deep network. To simplify the non-deterministic mapping, some methods [15, 18] rely on a set of pre-defined postures, or condition on person-specific styles and templates. These solutions can mimic motions of certain speakers/styles to some degree, but they are limited in terms of motion diversity and fidelity, especially for long talk speeches. Therefore, it is critical to developing algorithms that model the non-deterministic mapping between speech and head motions.

Based on studies in linguistics and psychology [39], speech motion helps the organization and presentation during speech delivery and contributes to both semantics and intonation. Semantically, head motions contribute to the utterance content. For example, some motions are conventionalized and attached to certain linguistic properties (e.g., "nod"). These motions are widely used to facilitate communication. In terms of intonation, the rhythmic movement that matches the prosody of audio could attract the attention of the audience, with the stressed syllable during speech. Moreover, proper rhythmic motions also reflect the progress of the speech and deliver a vivid listening experience. Such speech motion usually has no specific linguistic meaning and manifests as simple and fast hand dynamics related to prosody.

Motivated by these studies, we consider the structure of speech motions from a novel perspective. We introduce the concept of pose mode as the mode of the pose distribution that speakers have for fragments of speech. Considering the speaker's posture in a speech video as a random vector, it follows a multi-modal distribution in the high dimensional space. Modes in such distribution (values with local maximal density) correspond to the habitual postures of speakers. Our work focuses on motions in talk videos, where speakers organize a long speech around a certain topic. Under this setting, the pose modes are mostly habitual postures with no specific global meaning. Consequently, the structure of speech motions can be considered as the sequential transitions of pose modes with rhythmic dynamics under each pose mode. Therefore, the non-deterministic mapping from speech to head motion is decomposed into two parts: a stochastic mapping from speech semantics to pose modes, and the mapping from speech prosody to rhythmic motion dynamics. Our contributions are summarized as follows:

1. To address the non-deterministic mapping from speech to head motions, we propose to decompose the motion into pose modes and rhythmic motions.

The former is stochastically generated with a shallow diffusion model, and the latter is effectively inferred by speech prosody.
2. Extensive experiments demonstrate that our model generates plausible freedom motions well synced with the speech, outperforming other baselines in terms of the fidelity, similarity, and syncing with speech.

2 Related Work

Talking face generation is a cross-modal image synthesis task, Brand *et al.* [2] proposed *Voice Puppetry* for the generation of full facial animation from speech. With audio-driven facial animation, it can assist animation generation and film production. In the following paragraphs, we will overview the prior works about the audio-driven facial animation methods, which consist of facial landmarks, lip-sync animation, speaker-related animation, and image generation.

Facial Landmarks. A deep neural network-based facial landmarks generation is proposed by Eskimez *et al.* [10]. It was used in the talking face generation and improved speech intelligibility robust to noisy conditions. Chen *et al.* [5] proposed a cascade GAN-based method to generate a talking face, instead of learning a direct mapping between audio and image, a high-level structure of facial landmarks is used as a middle representation. First, transfer audio to landmarks and then generate the image conditioned on the landmarks. Greenwood *et al.* [13] jointly learn full-face animation and head pose, the landmarks were used as the image representation. In the image, each person had 62 landmarks distributed about the face, the landmarks along with lip edges and eyes. and translation combined.

Lip-Sync Animation. Given an arbitrary audio speech and one image of an arbitrary speaker, generating lip movement sync with the speech content is the lip-sync animation task. With the increased power of GPU computation, end-to-end learning [24,25,27,30,35] from audio to video frames have huge progress. Chen *et al.* [4] proposed to train an end-to-end model with a novel correlation loss to synchronize lip changes and speech changes, which is robust to view angels, lip shapes, and facial texture. Song *et al.* [26] propose a conditional recurrent generation network to build a temporal model for accurate lip synchronization, it considers the temporal dependency across video frames. To boost the accuracy of lip synchronization, a lip-reading discriminator is added. Vougioukas *et al.* [34] proposed an end-to-end method, using a static image of a speaker and an audio speech, without relying on handcrafted intermediate features. The model is based on a temporal GAN, that uses discriminators for the audio-visual synchronization, it generates lip movements sync with the speech. The speech styles like shouting or mumbling are related to the motion of face motion, Zhou *et al.* [47] proposed a three-stage LSTM network architecture to produce animator-centric speech motion curves, it is a real-time lip-sync from audio.

Speaker Related Animation. Given audio of a specific person, to synthesize a high-quality video of him speaking, replicate the sound and cadence of a person's voice. The speaker-related animation needs to model not only the speech content, but also requires to model the target style how it speaks, and how it expresses itself. Suwajanakorn *et al.* [28] used a recurrent neural network to learn the mapping between audio to mouth shapes conditioned on the same person of Obama. With the speaker-related model, it learns the texture of the lip. Cudeiro *et al.* [8] proposed a model that factors identity from facial motion, conditioning on speaker labels during training allows the model to learn different speaking styles. Thies *et al.* [32] proposed method with a latent variable to model the face of the target speaker, it learns temporal stability while rendering to generate video frames.

Image Generation. Fišer *et al.* [12] introduced a method of wrapping-based portrait video generation, with a controllable amount of landmarks to perform non-parametric texture synthesis. For the face image, image to image translation is popularly used to talking face synthesis. Thies *et al.* [33] proposed Face2Face to animate the facial expressions of the target speaker and re-render the output video in a photo-realistic fashion. It shows the robust appearance of face transfer between talking face videos. GAN-based method was proposed by Kim *et al.* [19], a recurrent GAN captures the Spatio-temporal features of talking face and could copy facial expressions from source to target speaker. A cycle-consistency loss [42] is added to the model for the facial expression styles transfer. Zakharov *et al.* [40] proposed few-shot talking face generation method, it performs meta-learning on a large dataset. The model embeds the face landmarks into embedding vectors, and the generator network maps the face landmarks into the output frames.

3 Method

We proposed a method called the shallow diffusion motion model (SDM) to generate a talking face sequence according to a given speech. To this end, a mapping from speech to face motion is required. We decomposed the talking face into pose motion and rhythmic motion. Additionally, to address the over-smoothing of generation of the talking face, a shallow diffusion mechanism was proposed for the generation of the motion sequence. Correspondingly, there are three modules for the proposed method, and the framework is shown in Fig. 1. For the input image and the video frames, we use a pre-trained face landmark detector [3] to do a preprocess and use the movement of the landmarks as the motion of the talking face.

Given a speech audio S and the corresponding video frame sequence contain the taking face F. The content encoder extracted feature on the input speech S, the content encoder is built up by four convolutional layers. The model of SDM is to learn the mapping between $S_{(i)}$ and $F_{(i)}$ of the i^{th} frame. In this work, we used mel-spectrum as the feature representation of speech $S_{(i)}$ and keypoint landmarks of the human face as the representation of the visual face frame $F_{(i)}$.

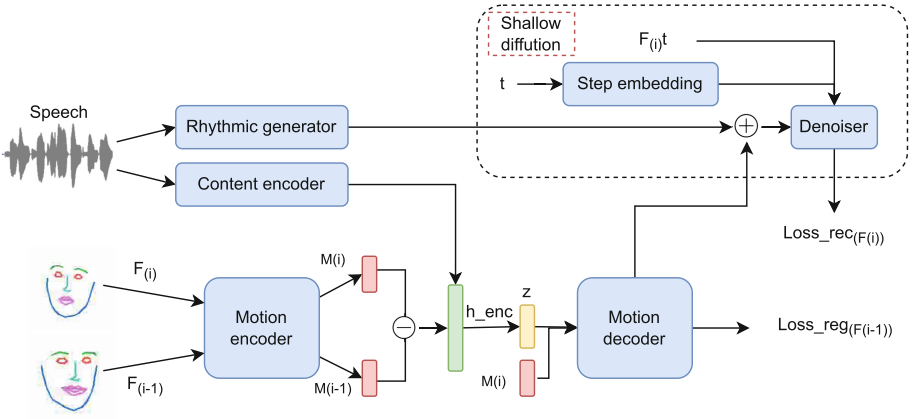

Fig. 1. The framework of talking face motion model with shallow diffusion model. The whole framework consists of two branches, a speech branch for the mapping of speech content and rhythm motions, a visual branch to learn the pose motion between frames.

To learn this mapping, we decompose the motion of talking face $F_{(i)}$ into two parts:

$$F_{(i)} = \overline{F}_{(i)} + \widetilde{F}_{(i)} \tag{1}$$

where $\overline{F}_{(i)}$ is content-related motion, and $\widetilde{F}_{(i)}$ is the rhythmic related motion of the talking face. The content-related motion can be regarded as the main motion of the head and the rhythmic-related motion could be the dynamics of the talking head. Finally, we use a shallow diffusion mechanism model for image generation.

3.1 Pose Motion Generation

The pose motion contains a motion encoder and a motion decoder. The content related pose motion conditioning on the content of the audio speech S_{cont} and the pose motion of the previous frame $F_{(i-1)}$.

$$\overline{F}^{*}_{(i)} = G_c(F_{(i-1)}, S_{cont}) \tag{2}$$

where the G_c represent the content related pose motion generator, and we use the superscript $*$ for the representation of the result of the generator.

The motion encoder encodes the Frame $F_{(i)}$ and $F_{(i-1)}$ into the latent vectors $M_{(i)}$ and $M_{(i-1)}$ separately. Conducting a subtraction between the neighbor frames could get the change the motion. With the condition on speech content to sync the motion change with the speech. The motion decoder does a reconstruction of the previous frame $F_{(i-1)}$ with the motion change variable and the latent vector $M_{(i)}$. The content-related motion can be formulated as a condition motion predictor. During the training phase, for each frame $F_{(i)}$ has fixed

the previous frame of $F_{(i-1)}$, the module of the motion decoder conducts the reconstruction of the motion of frame $F_{(i-1)}$ as:

$$M^*_{(i-1)} = f_{dec}(z, M_{(i)}) = f_{dec}(M_{(i-1)}) \tag{3}$$

where f_{dec} is the motion decoder, and z is the latent variable of the layer h_enc. We use the the motion reconstruct to regularize the embedding space of the motion encoder and decoder:

$$\mathcal{L}_{reg} = ||M_{(i)} - f_{dec}(M_{(i)})|| + ||M_{(i-1)} - f_{dec}(M_{(i-1)})|| \tag{4}$$

This forces the motion decoder to use the information of the latent variable z.

3.2 Rhythmic Motion Generation

The rhythmic motion is changed according to the temporal domain, it is important for the talking face to control the motion with dynamics. We generate the rhythmic motion through the rhythmic dynamics of the prosodic information in speech. It can keep the sync of prosody between the visual and audio.

In the control of rhythmic motion generation, we use a rhythmic generator for the dynamics motion embedding. The rhythmic generator is mainly built up with a convolutional network. The rhythmic motion is independent of the motion learned from the contend related pose motion, and the loss is defined as:

$$\mathcal{L}_{ind} = ||\widetilde{M^*}_{(i)} - \overline{M^*}_{(i)}|| \tag{5}$$

The $\mathcal{L}_{ind}$ ensures the generated rhythmic motion pose $\widetilde{M^*}_{(i)}$ independent to the content related motion pose $\overline{M^*}_{(i)}$. It helps the dynamics of motion are not affected by the content of speech.

3.3 Shallow Diffusion Mechanism

The shallow diffusion mechanism is applied to the image animation generation. The main module of the shallow diffusion mechanism comes from the diffusion model [20–22]. The diffusion model contains two processes, a diffusion process to convert the image data into a Gaussian distribution step by step, and a reverse process to reconstruct the image data from Gaussian white noise. The pipeline of a diffusion model is shown in Fig. 2.

Diffusion Process. Let the distribution of data $F^0_{(i)}$ as $p(F^0_{(i)})$, the diffusion process converts the $F^0_{(i)}$ into $F^T_{(i)}$ step by step with a Markov chain with fixed parameters. The T steps conversion can be formulated as:

$$q(F^{1:T}_{(i)}|F^0_{(i)}) = \prod_{t=1}^{T} q(F^T_{(i)}|F^{t-1}_{(i)}) \tag{6}$$

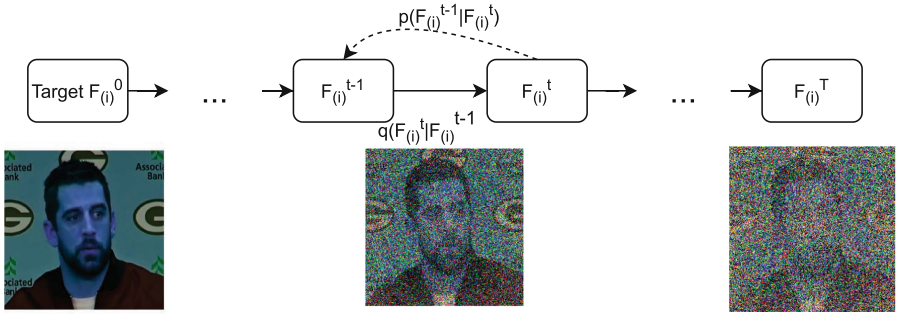

Fig. 2. The two processes of diffusion model. The diffusion process is from $F_{(i)}^0$ to $F_{(i)}^T$, the reverse process is from $F_{(i)}^T$ to $F_{(i)}^0$.

At each step $t \in (1, T)$, a Gaussian noise multiply with a variance of $\alpha \in [\alpha_1, \cdots, \alpha_T]$ is added to the $F_{(i)}^{t-1}$ to obtain $F_{(i)}^t$.

$$q(F_{(i)}^t | F_{(i)}^{t-1}) = \mathcal{N}(F_{(i)}^t; \sqrt{1-\alpha_t} F_{(i)}^{t-1}, \alpha_t \mathbf{I}) \tag{7}$$

If the parameters of α are well designed, and the step T is larger enough, the final $q(F_{(i)}^T)$ is equally an isotropic Gaussian distribution.

Reverse Process. The reverse process is from $F_{(i)}^T$ to $F_{(i)}^0$, which is follow the Markov chain with learnable parameters θ. The reverse process can be approximate it with the neural networks with the parameters θ. It can be formulated as:

$$p_\theta(F_{(i)}^{0:T}) = p(F_{(i)}^T) \prod_{t=1}^{T} p_\theta(F_{(i)}^{t-1} | F_{(i)}^t) \tag{8}$$

To learn the parameters θ, we optimizing the loss with stochastic gradient descent on:

$$\mathcal{L}_{diff} = D_{KL}(q(F_{(i)}^{t-1} | F_{(i)}^t, F_{(i)}^0) || p_\theta(F_{(i)}^{t-1} | F_{(i)}^t)) \tag{9}$$

where $D_{KL}()$ is the Kullback-Leibler divergence. Finally, with the trained network, we can sample from $p(F_{(i)}^T) \sim \mathcal{N}(0, I)$ to generate the target data with the reverse process.

When the step of T is big enough, the trajectory from $F_{(i)}^0$ to Gaussian $F_{(i)}^T$ and the trajectory from $F_{(i)}^T$ to $F_{(i)}^0$ will meet in a step t. Inspired by this point, we can use an auxiliary predictor to predict the step of t. With the step of t to do a shallow diffusion process. And the reverse process could also start at the predicted step of t.

3.4 Training Losses

The reconstruct loss is applied to final reconstution of the image F_i^*:

$$\mathcal{L}_{rec} = ||F_i^* - F_{(i)}|| \tag{10}$$

Overall the total loss is defined as:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{reg} + \lambda_2 \mathcal{L}_{ind} + \lambda_3 \mathcal{L}_{rec} \tag{11}$$

where the $\lambda_1 \sim \lambda_3$ is hyperparameters for balancing the different losses.

3.5 Testing Stage

The pipeline of talking face inference phase is followed as the function:

$$\begin{aligned} F_{(i)}^* &= G(F_{(i-1)}^*, S(i)) \\ &= f_{denoi}(f_{dec}(f_{con}(S(i)), f_{enc}(F_{(i-1)})), f_{rhy}(S(i)), t, F_{(i)^t}) \end{aligned} \tag{12}$$

where f_{denoi} is the reverse processes of diffusion, f_{rhy} is the rhythmic generator. The final results of talking face video are a stack of the frames $F_{(1)}^*, \cdots, F_{(n)}^*$ with the tool of ffmpeg.

4 Experiments and Results

4.1 Experimental Setup

Datasets. We used three datasets for the experimental evaluation, it contains VoxCeleb2 [6], LRW [7] and LRS3-TED [1]. The VoxCeleb2 contains more than 6000 celebrities and covers 1 million utterances in speech. The LRW is a large dataset containing 1000 speakers, and each speaker spoke 500 different words. The LRS3-TED includes face track over 400 h of videos from TED and TEDx, it has more challenges with head movements than others. We follow the raw split as the ratio of the dataset.

Training Details. We use the optimizer of ADAM with the learning rate of 2×10^{-4}, and the β_1 of 0.4, β_2 of 0.999. In the traning phase, we set the loss weight in Eq. 11 as λ_1 of 5, λ_2 of 2, and λ_3 of 1. The experiment was conducted on a single GPU of NVIDIA Tesla V100 with 16 GB memory.

Metrics of Evaluation. For the quantitative evaluation, we adopted several criteria, it includes Frchet Inception Distance (FID) [16], which was used to quantify the fidelity of the synthesized image, and structured similarity (SSIM) [36], it was used to compare the similarity of the synthesized image and real images. We use cosine similarity (CSIM) [40] to identify the speaker identity preserving ability, which computed the cosine distance between the embedding vectors of a face recognition network [9]. To check if the synthesized video contains sync movement of the lip to speech content, we use Landmarks Distance (LMD) [4] for evaluation.

4.2 Results and Analysis

Comparision with Talking Face Methods. We first compare the proposed method with the related works of talking face methods, we select the audio-driven method. With given a single images and an audio to generate the video of talking face, which has been studied in Zhou *et al.* [45], Song *et al.* [26], Chung *et al.* [17], Vougioukas *et al.* [34], Chen *et al.* [5], and Wiles *et al.* [37]. For a fair comparison, all the methods were input with the same image and speech from the test dataset. And we do a preprocess on the input image with the same cropping area. The quantitative evaluation results are shown in Table 1.

Table 1. Comparisions with different audio to video methods on the three public dataset of VoxCeleb2, LRW, and LRS3-TED. The score of FID and LMD smaller is better, while for SSIM and CSIM bigger is better. We bold each leading score.

Method	Datasets											
	VoxCeleb2				LRW				LRS3-TED			
	FID	SSIM	CSIM	LMD	FID	SSIM	CSIM	LMD	FID	SSIM	CSIM	LMD
Zhou *et al.* [45]	137	0.84	0.32	4.8	149	0.85	0.39	3.7	221	0.72	0.27	6.2
Song *et al.* [26]	163	0.78	0.27	5.6	134	**0.91**	0.45	**3.1**	204	0.62	0.28	6.5
Chung *et al.* [17]	159	0.79	0.29	5.4	132	**0.91**	0.44	**3.1**	212	0.58	0.32	6.7
Vougioukas *et al.* [34]	127	**0.85**	0.33	6.3	116	0.88	0.35	3.6	196	0.63	0.26	6.4
Chen *et al.* [5]	142	0.82	0.31	4.9	151	0.84	0.38	3.3	294	0.66	0.31	4.8
Wiles *et al.* [37]	117	0.65	0.31	4.8	107	0.69	0.31	3.2	172	0.57	0.28	5.6
Ours	**97**	0.74	**0.42**	**3.4**	**102**	0.76	**0.49**	**3.1**	**122**	**0.79**	**0.44**	**3.2**

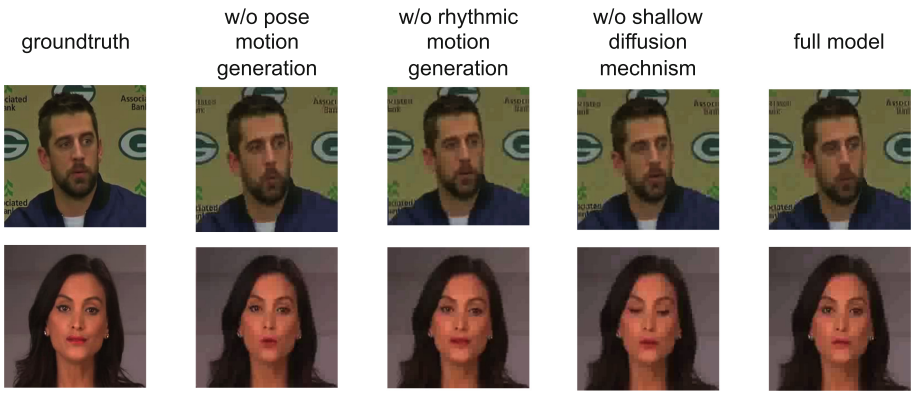

Fig. 3. The ablation studies with visualization, three main modules of pose motion generation, rhythmic motion generation, and shallow diffusion mechnism are compaired with the full model.

Note that in the preprocess of our method, we did not include the affine transformation, which leads to a lower score in terms of SSIM. From the results shown in Table 1, we can see except for the SSIM score, our method could achieve

the best performance than other related audio to video methods in most evaluation metrics. As shown, the proposed method outperforms other baselines, suggesting better generation ability in relating audio and motion. Our model shows strong performances on the fidelity of the synthesized image by the low FID scores, while other baselines fail to generate high fidelity images on some speakers. Our model is more robust to lip motion syncing, leading to lower averaged LMD scores. Our model is more accurate keep the speaker identity in the synthesized image, which leads to a high score of CSIM.

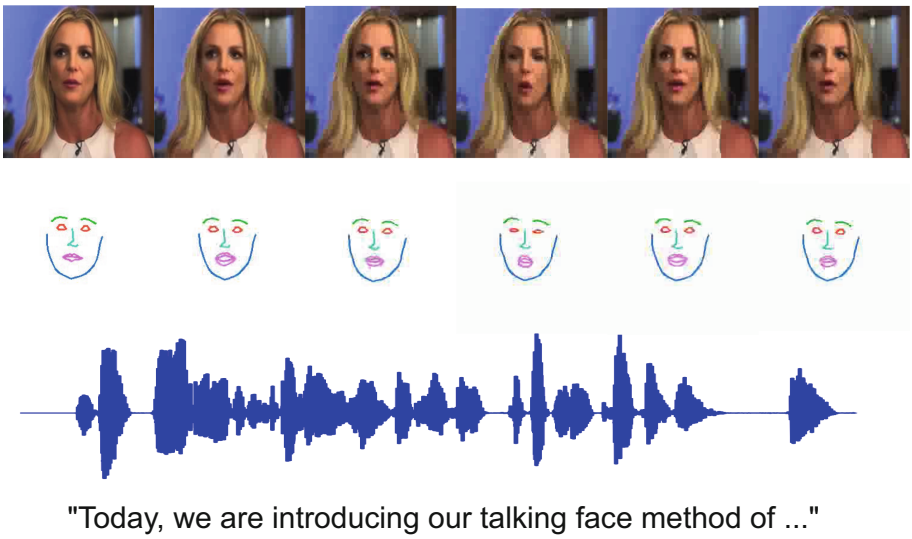

Fig. 4. End to end video generation results. From bottom to top row are speech text, speech audio, generation keypoint, and the final video sequence.

Ablation Studies. We compare the contributions of different modules in the ablation studies, the primary modules described in Sect. 3. We conduct the experiments on the dataset of VoxCeleb2. As shown in Fig. 3, we visualize the result of each module compared with the full model.

From the results shown in Fig. 3, we can see the synthesized frames without a shallow diffusion mechanism in the fourth column, the motion of the face has a bigger distance from the groundtruth. We attribute this to the shallow diffusion model, the diffusion-based module could synthesize the target image more robustly, which could stabilize the generation and could lead to a faster convergence during training. Another case we found in the ablation studies is the pose motion generation module affects the lip part of the face in the second column. Without the pose motion generation module, the synthesized image could not control the mouth for the speech content.

Video Results. Further, we show the results based on our generated motion to the video frames in Fig. 4. The video can be generated end to end by inputting a speech and an image. We can apply the method to the arbitrary input image in the wild, it can generate any identity. It can be used for the recording of video presentations.

5 Conclusion

In this work, we propose an approach based on a shallow diffusion mechanism that synchronizes faces with speech content through rhythmic movements of the head. We solve the non-deterministic mapping problem by decomposing the difficult task into complementary parts. Given input speech audio, pose motion generation generates different pose patterns sequentially through conditional sampling, while rhythmic motion generation simultaneously enriches each pose pattern dynamically with audio-conditioned rhythms to achieve spontaneous movements. Our model generates highly diverse and visually plausible face images in a shallow diffusion mode, from the prediction time step to conducting the reverse process of diffusion.

Acknowledgement. This paper is supported by the Key Research and Development Program of Guangdong Province under grant No.2021B0101400003. Corresponding author is Jianzong Wang from Ping An Technology (Shenzhen) Co., Ltd (jzwang@188.com).

References

1. Afouras, T., Chung, J.S., Zisserman, A.: Lrs3-ted: a large-scale dataset for visual speech recognition. arXiv preprint arXiv:1809.00496 (2018)
2. Brand, M.: Voice puppetry. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, pp. 21–28 (1999)
3. Bulat, A., Tzimiropoulos, G.: 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 (2017)
4. Chen, L., Li, Z., Maddox, R.K., Duan, Z., Xu, C.: Lip movements generation at a glance. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 520–535 (2018)
5. Chen, L., Maddox, R.K., Duan, Z., Xu, C.: Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7832–7841 (2019)
6. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: deep speaker recognition. arXiv preprint arXiv:1806.05622 (2018)
7. Chung, J.S., Zisserman, A.: Lip reading in the wild. In: Lai, S.-H., Lepetit, V., Nishino, K., Sato, Y. (eds.) ACCV 2016. LNCS, vol. 10112, pp. 87–103. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54184-6_6
8. Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., Black, M.J.: Capture, learning, and synthesis of 3d speaking styles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10101–10111 (2019)

9. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4690–4699 (2019)
10. Eskimez, S.E., Maddox, R.K., Xu, C., Duan, Z.: Generating talking face landmarks from speech. In: Deville, Y., Gannot, S., Mason, R., Plumbley, M.D., Ward, D. (eds.) LVA/ICA 2018. LNCS, vol. 10891, pp. 372–381. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93764-9_35
11. Eskimez, S.E., Zhang, Y., Duan, Z.: Speech driven talking face generation from a single image and an emotion condition. IEEE Trans. Multimedia **24**, 3480–3490 (2021)
12. Fišer, J., et al.: Example-based synthesis of stylized facial animations. ACM Trans. Graph. (TOG) **36**(4), 1–11 (2017)
13. Greenwood, D., Matthews, I., Laycock, S.: Joint learning of facial expression and head pose from speech. In: Interspeech (2018)
14. Guo, Y., Chen, K., Liang, S., Liu, Y.J., Bao, H., Zhang, J.: Ad-nerf: audio driven neural radiance fields for talking head synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5784–5794 (2021)
15. Gupta, A., Khan, F.F., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: Intelligent video editing: incorporating modern talking face generation algorithms in a video editor. In: Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing, pp. 1–9 (2021)
16. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Adv. Neural Inf. Processi. Syst. **30**, 1–12 (2017)
17. Jamaludin, A., Chung, J.S., Zisserman, A.: You said that?: synthesising talking faces from audio. Int. J. Comput. Vision **127**(11), 1767–1779 (2019)
18. Ji, X., Zhou, H., Wang, K., Wu, W., Loy, C.C., Cao, X., Xu, F.: Audio-driven emotional video portraits. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14080–14089 (2021)
19. Kim, H., et al.: Neural style-preserving visual dubbing. ACM Trans. Graph. (TOG) **38**(6), 1–13 (2019)
20. Kingma, D., Salimans, T., Poole, B., Ho, J.: Variational diffusion models. Adv. Neural Inf. Process. Syst. **34**, 21696–21707 (2021)
21. Kong, Z., Ping, W., Huang, J., Zhao, K., Catanzaro, B.: Diffwave: a versatile diffusion model for audio synthesis. In: 9th International Conference on Learning Representations, ICLR 2021 (2021)
22. Lam, M.W., Wang, J., Su, D., Yu, D.: Bddm: bilateral denoising diffusion models for fast and high-quality speech synthesis. In: International Conference on Learning Representations (2021)
23. Meshry, M., Suri, S., Davis, L.S., Shrivastava, A.: Learned spatial representations for few-shot talking-head synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13829–13838 (2021)
24. Qu, X., Wang, J., Xiao, J.: Enhancing data-free adversarial distillation with activation regularization and virtual interpolation. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3340–3344. IEEE (2021)
25. Si, S., Wang, J., Peng, J., Xiao, J.: Towards speaker age estimation with label distribution learning. In: ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4618–4622 (2022). https://doi.org/10.1109/ICASSP43922.2022.9746378

26. Song, Y., Zhu, J., Li, D., Wang, A., Qi, H.: Talking face generation by conditional recurrent adversarial network. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, pp. 919–925 (2019)
27. Sun, A., et al.: Reconstructing dual learning for neural voice conversion using relatively few samples. In: IEEE Automatic Speech Recognition and Understanding Workshop, pp. 946–953. IEEE (2021)
28. Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing obama: learning lip sync from audio. ACM Trans. Graph. (ToG) **36**(4), 1–13 (2017)
29. Tang, H., Zhang, X., Wang, J., Cheng, N., Xiao, J.: Avqvc: one-shot voice conversion by vector quantization with applying contrastive learning. In: 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2022), pp. 1–5. IEEE (2022)
30. Tang, H., et al.: TGAVC: Improving autoencoder voice conversion with text-guided and adversarial training. In: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU2021), pp. 938–945. IEEE (2021)
31. Tang, J., Wu, Y., Li, M., Wang, Z.: Talking face generation based on information bottleneck and complementary representations. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp. 3443–3447 (2021)
32. Thies, J., Elgharib, M., Tewari, A., Theobalt, C., Nießner, M.: Neural voice puppetry: audio-driven facial reenactment. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12361, pp. 716–731. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58517-4_42
33. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: real-time face capture and reenactment of rgb videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2387–2395 (2016)
34. Vougioukas, K., Petridis, S., Pantic, M.: Realistic speech-driven facial animation with gans. Int. J. Comput. Vision **128**(5), 1398–1413 (2020)
35. Wang, Q., Zhang, X., Wang, J., Cheng, N., Xiao, J.: Drvc: a framework of any-to-any voice conversion with self-supervised learning. In: 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2022), pp. 3184–3188. IEEE (2022)
36. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
37. Wiles, O., Koepke, A., Zisserman, A.: X2face: a network for controlling face generation using images, audio, and pose codes. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 670–686 (2018)
38. Xu, L., Zhou, X.: A crowd-powered task generation method for study of struggling search. Data Sci. Eng. **6**(4), 472–484 (2021)
39. Yao, X., Fried, O., Fatahalian, K., Agrawala, M.: Iterative text-based editing of talking-heads using neural retargeting. ACM Trans. Graph. (TOG) **40**(3), 1–14 (2021)
40. Zakharov, E., Shysheya, A., Burkov, E., Lempitsky, V.: Few-shot adversarial learning of realistic neural talking head models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9459–9468 (2019)
41. Zhang, C., Ni, S., Fan, Z., Li, H., Zeng, M., Budagavi, M., Guo, X.: 3d talking face with personalized pose dynamics. IEEE Trans. Visualization Comput. Graph. **29**, 1438–1449 (2021)

42. Zhang, X., Wang, J., Cheng, N., Xiao, E., Xiao, J.: CycleGEAN: cycle generative enhanced adversarial network for voice conversion. In: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU2021), pp. 1–6. IEEE (2021)
43. Zhang, X., Wang, J., Cheng, N., Xiao, J.: Susing: su-net for singing voice synthesis. In: International Joint Conference on Neural Networks, IJCNN 2022. IEEE (2022)
44. Zhang, X., Wang, J., Cheng, N., Xiao, J.: Tdass: target domain adaptation speech synthesis framework for multi-speaker low-resource tts. In: International Joint Conference on Neural Networks, IJCNN 2022. IEEE (2022)
45. Zhou, H., Liu, Y., Liu, Z., Luo, P., Wang, X.: Talking face generation by adversarially disentangled audio-visual representation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 9299–9306 (2019)
46. Zhou, H., Sun, Y., Wu, W., Loy, C.C., Wang, X., Liu, Z.: Pose-controllable talking face generation by implicitly modularized audio-visual representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4176–4186 (2021)
47. Zhou, Y., Xu, Z., Landreth, C., Kalogerakis, E., Maji, S., Singh, K.: Visemenet: audio-driven animator-centric speech animation. ACM Trans. Graph. (TOG) **37**(4), 1–10 (2018)