3D Talking Face with Personalized Pose Dynamics

Anonymous cvm submission

Paper ID 150

Abstract

Recently, we have witnessed a booming growth in applications of 3D talking face generation. However, existing methods can only generate 3D faces with the static head pose, which is inconsistent with the human sense. In this paper, we propose a unified audio-inspired approach to endow 3D talking face with personalized pose dynamics. To achieve this goal, we establish an original person-specific dataset, providing corresponding head pose sequence and face shapes for each video. Our framework is composed of two separate modules, PoseGAN and PGFace. Given input audio, PoseGAN first produces head pose sequence for 3D head, then PGFace module utilizes the audio and pose information to generate natural face models. With the combination of these two parts, a 3D talking head with dynamic head movements can be constructed. To our best knowledge, this is the first audio-driven technique to automatically generate 3D talking faces with pose dynamics. Experimental evidences indicate our method generates preferable results and best matches with human experience.

1. Introduction

Talking face generation is an attractive research topic in computer vision and graphics. Aside from being interesting, it has a wide range of applications, *e.g.*, game animation, 3D video calls, and 3D avatars for AR/MR. Most of the existing works [11, 14, 25, 40, 45, 54, 32, 42, 47] have been proposed to generate talking faces from static images. Due to the lack of 3D face model datasets, there are only a few works [55, 16] being proposed to generate talking faces in 3D shapes.

The synthesized talking face from the state-of-the-art approaches usually has a static and fixed pose of the head model throughout the whole speech process. However, in any realistic talking scenario, the person's head will rotate and translate accordingly. If the 3D talking face cannot move reasonably, it will not seem authentic for the audience. We name the corresponding movement of the head as *head pose sequence* in this work. Convolutional Neural Network (CNN) has been adopted as an encoder for 3D face shape generation to achieve state of the art results [16]. VisemeNet [55] adopted Long Short-Term Memory (LSTM) network to generate 3D talking face without any head movement. It should be noted that all these conventional methods do not take head poses into consideration when generating 3D talking faces, which severely compromises the reality of the synthesized results. The head pose sequences vary in different video scenarios, but show strong correlations with the person's identities, as illustrated in Figure 2. Therefore, generating dynamic pose animations is a crucial step for realistic 3D talking head syntheses.

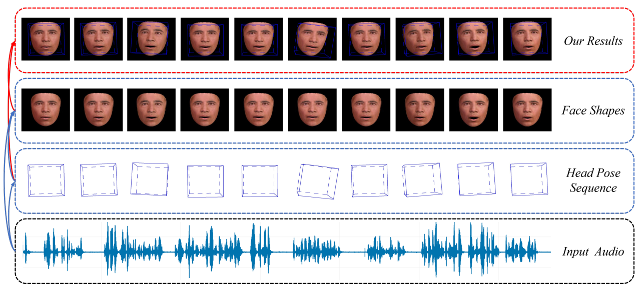

Figure 1. Pipeline to synthesize the talking face with pose dynamics. Given an input audio, we generate the corresponding sequence of 3D head pose and face shapes.

In this paper, we introduce a fully automatic generation framework for audio-driven 3D talking face with pose dynamics (see Figure 1). To assign different persons with individual head poses, we build a person-specific head motion dataset, providing corresponding head pose sequences and face shapes for each video. During the inference phase, the input audio is first encoded with deep speech [23] and the extracted features are then fed into two proposed modules, the head Pose Generative Adversarial Network (PoseGAN) module and Pose-Guided Face (PGFace) generation module. As shown in Figure 3, the PoseGAN module is used to extract the cross-modal head pose sequence with rotation and translation parameters. PGFace module with head pose parameters is applied to generate face shape parameters corresponding to the audio. With the combination of the audio, head pose sequence, and face shape parameters, the final 3D

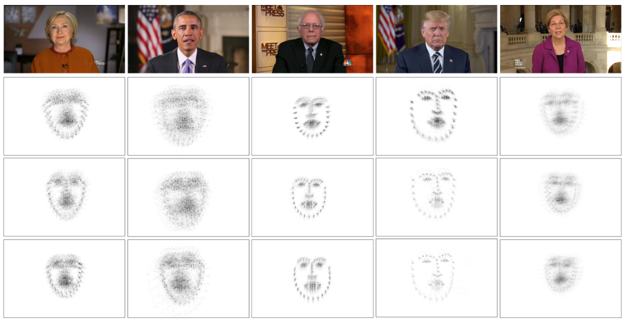

Figure 2. Our person-specific head motion dataset. Below each person are three heatmaps of face landmarks tracked from different videos, which depict the frequency of landmarks in different spatial locations. This visualization reveals the speaker's resting pose and their unique head movement style.

talking face with pose dynamics can be synthesized.

To the best of our knowledge, this work is the first audio-driven technique to automatically generate 3D talking faces with pose dynamics. Based on this person-specific head motion dataset, we propose an end-to-end unified approach to synthesize a natural 3D talking head. The main contributions of our work are three-fold:

- We introduce a new method to construct a person-specific head motion dataset, which includes over 535,400 frames from 450 video clips. Based on this dataset, a unified audio-driven framework is proposed to generate 3D talking faces with pose dynamics.
- Taking audio flows as input, a new cross-modal PoseGAN module is proposed to generate the dynamic head poses. A new loss function and initial poses are introduced to ensure the consistency of long-term generations. A PGFace module is designed for pose-dependent facial shape correction, which makes the face shape rendering results more realistic.
- Extensive ablation studies and comparisons with conventional methods indicate that our method is able to generate person-specific head pose sequence that is in sync with the input audio and best matches with the human expectation of talking heads.

2. Related Work

There has been a branch of researches in facial animation that focuses on synthesizing the facial motion from audios, and generating either 2D videos or 3D models as the results.

Audio-based 2D facial animation Chung *et al.* [14] proposed an encoder-decoder CNN model to generate synthesized talking face video frames. Deep bidirectional LSTM (BLSTM) was applied by Fan *et al.* [19] in their talking head system. Vougioukas *et al.* [45] used a temporal GAN with two discriminators to generate lip movements and facial expressions. Suwajanakorn *et al.* [40] proposed to learn the mapping from raw audio features to mouth shapes by a recurrent neural network. Chen *et al.* [11] devised a network to synthesize lip movements and proposed a correlation loss to synchronize lip changes and speech changes. Xie and Liu [48] used a dynamic Bayesian network to model the movements of articulators. Jalalifar *et al.* [25] produced realistic faces conditioned on landmarks using a recurrent neural network and a conditional GAN [31, 22]. The arbitrary subject talking face generation method is realized by Zhou *et al.* [54] using disentangled audio-visual representation with GANs.

It should be noted that none of these 2D facial video synthesis methods consider the personalized head motion. Our synthesized 3D talking head with personalized pose dynamics can serve as an important intermediate step for these 2D video synthesis methods, which we would like to explore in our future work.

Audio-based 3D facial animation A deep learning approach proposed by Taylor *et al.* [41] uses a sliding window predictor that learns mappings from phoneme label input sequences to mouth movements. Zhou *et al.* [55] proposed an automatic real-time lip-synchronization from audio solution based on LSTM network architecture. Karras *et al.* [26] presented real-time, low latency 3D facial animations based on speech audio input with emotional state. Liu *et al.* [30] employed a data-driven regressor for modeling the correlation between speech data and mouth shapes with a DNN acoustic model. The dynamic facial expressions of the source subject were transferred to the target subject in [52]. Face Transfer is based on a multilinear model [44] of 3D face meshes that separable parameterizes the space of geometric variations. Most recently, Cudeiro *et al.* [16] proposed Voice Operated Character Animation (VOCA), which takes a random speech signal as input and generates a wide range of adult faces realistically. VOCA first converts the input audio into DeepSpeech [23] features, then one-hot encoding with different subjects is used to train offsets of 3D face mesh. The FLAME [29] model is applied to generate their final face shape.

However, none of these works take the personalized head motions into consideration and the results from these works highly depend on the quality of 3D face dataset which is hard to collect in real life.

Text-based facial animation Relatively small amount of works have been proposed to generate face model directly from text input. Sako *et al.* [34] described a text-based technique to generate realistic auditory speech and lip image sequences using Hidden Markov Models (HMMs). The system for expressive Visual Text-To-Speech (VTTS) was presented by Anderson *et al.* [4] in which the face is

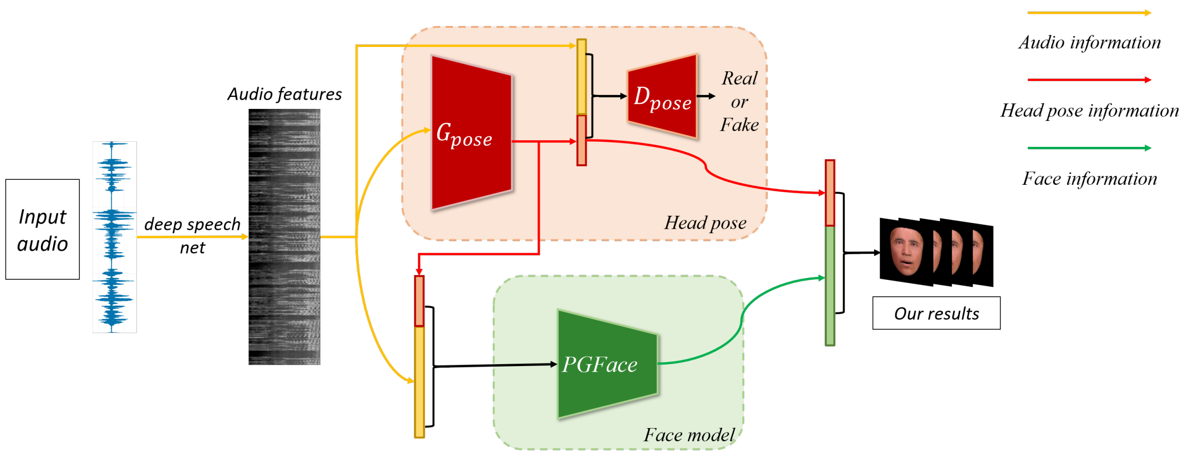

Figure 3. An overview of our unified framework. G_{pose} denotes the generator of 3D head pose sequence and D_{pose} is the discriminator. Face shape parameters are generated by PGFace.

modeled using an Active Appearance Model (AAM). Kumar *et al.* [28] presented a text-based lip-sync generation method that takes a time-delayed LSTM to generate mouth keypoints synced to the audio. Hong *et al.* [24] described a visual speech synthesizer that provides a form of virtual face-to-face communication using text streams.

While in this work we focus on the generation of 3D faces from audio, it is possible to convert our framework into a text-driven model by using a Text-to-Speech engine (e.g.,Tacotron 2 [37]), which we leave to our future work for further in-depth exploration.

3D face datasets On the one hand, Several datasets [8, 35, 50] are concerned with the static 3D face model analysis.On the other hand, some datasets [2, 9, 51, 15, 53] focus on dynamic 3D face models and expressions. In addition, there are several datasets containing scanned face models. Cheng *et al.* [13] published the 4DFAB dataset containing 4D captures of 180 subjects and Fanelli *et al.* [20] proposed a 3D audio-visual corpus, which contains a large set of audio-4D scan pairs using a real-time 3D scanner. The VOCASET presented by Cudeiro *et al.* [16] contains 3D scans of 255 sentences with the entire head and neck. Our approach in this paper is a novel dataset construction method. We generate a large number of face models and head pose sequences corresponding to speech.

3. Dataset

The motivation in this work is to learn and extract pose characteristics of human talking face from any data available in the wild. However, real-world 3D face data is labor-intensive to capture using high-speed facial scanners. Another disadvantage of such 3D capture is that this kind of data is typically captured by a well-designed environment with tens of cameras and projectors. Hence the participants may unintentionally suppress their natural head movements and facial expressions under such conditions. In contrast, in most videos of real-world scenarios available online, people usually perform more natural behaviors, which can serve our research purpose much better. To this end, we advocate collecting dynamic 3D talking data by analyzing the videos in the wild instead of the labor-intensive 3D facial capture.

The videos used in this paper has a total length of approximately 5 hours, collected from the videos used by Agarwal *et al.* [1] for their deepfake detection. Our dataset contains over 535,400 frames from 450 video clips along with the audios, 3D head pose parameters, and 3D face shape parameters.

Head pose parameters We adopt the OpenFace [3] to generate 3D head pose parameters. Head pose $\mathbf{p} \in \mathbb{R}^6$ is represented by Euler angles (pitch θ_x, yaw θ_y, roll θ_z) and a 3D translation vector $\mathbf{t}$. If we naively apply head pose sequences detected in the original video by OpenFace, it will cause unstable effects in some high-frequency regions and the head motion will look unsatisfying. Therefore, we propose a Gaussian filtering method that filters the head pose parameters throughout the time dimension and generates convincing results. Specifically, our Gaussian filtering method removes the abnormal head jitter effectively. As shown in Figure 4, the $pitch$ parameter of head pose is measured in the time dimension over the video clip. In the high-frequency region (e.g., the area in the red rectangle), the curve of the pitch parameter is smoothed as shown by the orange curve. The Gaussian density and head pose fil-

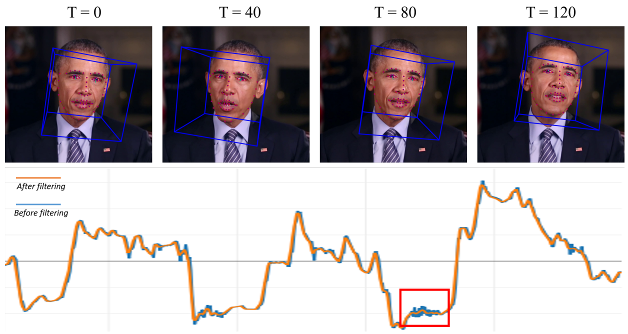

Figure 4. Gaussian Filtering. Blue curve denotes the original pitch parameter. Orange curve is for the smoothed pitch parameter.

tering functions are given as follows:

$$F(x) = \frac{1}{\sqrt{2\pi}\delta} e^{-\frac{1}{2\delta^2}x^2},$$
$$\mathbf{p}(i) = \sum_{k=i-m}^{i+m} \mathbf{p}(k)F(k-i), \tag{1}$$

where i is the frame index, $2m$ is the window size of the filter, and $\mathbf{p}(i)$ indicates the head pose of the ith frame.

The original videos are divided into small sets of video clips based on the camera parameters, the detection of the frame continuity, and the length of frames. The head pose is centralized and unified under the same coordinate system in every small video set.

3D face shape parameters The deep 3D face reconstruction method [17] achieves state-of-the-art performance on multiple datasets. Therefore, we apply this method to generate face shape parameters $[\alpha_{\text{id}}, \alpha_{\text{exp}}]$. The 3DMM [5, 8] face shape model is defined as:

$$S = \overline{S} + B_{\text{id}}\alpha_{\text{id}} + B_{\text{exp}}\alpha_{\text{exp}}, \tag{2}$$

where $\overline{S}$ is the averaged face shapes; B_{id} and B_{exp} are the PCA bases of identity and expression respectively; $\alpha_{\text{id}} \in \mathbb{R}^{80}$ and $\alpha_{\text{exp}} \in \mathbb{R}^{64}$ are the corresponding coefficients.

It is generally a non-trivial task to capture the 3D face models. We provide a unified framework to get precise 3D face models corresponding to video frames along with the head pose sequence. Such person-specific dataset supports our fully automatic framework for generating 3D talking face. The proposed method for data collection and preparation can be also easily extended to the videos of other person identities available online.

4. Methodology

4.1. Head Pose Sequence Generation Network

Generate a corresponding 3D head pose sequence from input audio is non-trivial. Depending on the speaking scenarios and individual speaking habits, people do not always exhibit the same head pose sequence when speaking the same words. Ginosar *et al.* [21] proposed an audio-based generation method for 2D body gestures. Specifically, they acquired the 2D landmarks of the character's arm and gesture from audio inputs, and demonstrated the effectiveness of GAN for cross-modal pose generation.

The generation of head pose sequence is also a cross-modal prediction task. Inspired by Ginosar *et al.* [21], we propose the PoseGAN to generate the corresponding head pose sequence. To ensure the correlation between the generated head pose sequence and the input audio, we introduce the conditional GAN to determine the output of the head pose sequence that belongs to the specific character and a discriminator to determine the authenticity of the head pose sequence. Here, we set 256 frames as a unit sequence.

We notice that the conventional pose loss cannot guarantee the consistency between neighboring sequences and the continuity of head poses in each sequence. To address these problems, an embedding method and a motion loss function are proposed. Experimental results show that with the initial pose loss constraint and the motion loss function, the two discontinuity problems are solved successfully.

4.1.1 Generator

As shown in Figure 5, we develop an enhanced CNN encoder before the U-net [33] structure to build the generator G and embed the initial head pose $\mathbf{p}$ into the input layer and the U-net output layer to constrain the initial position and orientation of the generated head pose sequence.

The initial head pose $\mathbf{p}$ and audio $\mathbf{x}$ are simultaneously input into the generator G, as shown in Figure 5. During the training stage, the pose of the first frame is adopted as the initial pose $\mathbf{p}$ in the head pose sequence. During the inference stage, the rest pose of the same identity is adopted as $\mathbf{p}$ for the generation of the first head pose sequence. The last pose of previous sequence is adopted as $\mathbf{p}$ for subsequent head pose sequence generation. The initial pose guarantees the consistency between neighboring sequences.

The output head pose sequence presents abnormal instability when directly using the L^2 norm of pose loss (defined in Equation 3), since there are no constraints for continuous motion between frames. We introduce the motion loss to ensure the motion continuity of the output head pose sequence.

The L^2 norm loss functions for pose and motion are defined as follows:

$$\mathcal{L}_{\text{pose}}(G) = \mathbb{E}_{\mathbf{x},\mathbf{y},\mathbf{p}}[\|\mathbf{y} - G(\mathbf{x},\mathbf{p})\|^2 + \|\mathbf{p} - G_0(\mathbf{x},\mathbf{p})\|^2],$$
$$\mathcal{L}_{\text{motion}}(G) = \mathbb{E}_{\mathbf{x},\mathbf{y},\mathbf{p}}[\|(\mathbf{y}_{t+1} - \mathbf{y}_t) - (G_{t+1}(\mathbf{x},\mathbf{p}) - G_t(\mathbf{x},\mathbf{p}))\|^2], \tag{3}$$

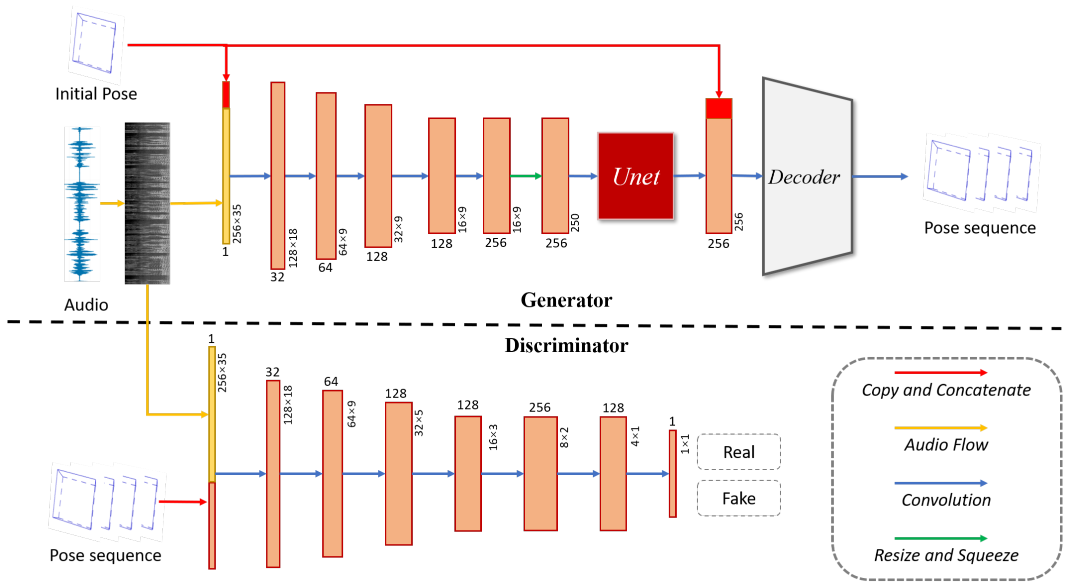

Figure 5. The architecture of our PoseGAN for head pose estimation from input audios.

where $\mathbf{x}$ is the input audio feature, $\mathbf{y}$ represents the corresponding head pose sequence with 256 frames, $0 \leq t < 256$ and $\mathbf{p}$ indicates the initial head pose. The generator takes $\mathbf{x}$ and $\mathbf{p}$ as inputs and predicts the head pose sequence. G_0 is the first frame in the generated head pose sequence.

The Generator's loss function is defined as:

$$\mathcal{L}_{L^2}(G) = \alpha \mathcal{L}_{\text{pose}}(G) + \beta \mathcal{L}_{\text{motion}}(G), \tag{4}$$

where α and β are weights to control the balance between the pose and motion losses.

4.1.2 Discriminator

A CNN structure is applied to discriminate the true and false head pose sequences, by taking the generated head pose sequence $G(\mathbf{x}, \mathbf{p})$ combined with audio $\mathbf{x}$ as input.

The loss function of discriminator D is defined as:

$$\begin{aligned} \mathcal{L}_{\text{GAN}} = \arg \min_{G} \max_{D} \mathbb{E}_{\mathbf{x},\mathbf{y}}[\log D(\mathbf{x}, \mathbf{y})] + \\ \mathbb{E}_{\mathbf{x},\mathbf{p}}[\log(1 - D(\mathbf{x}, G(\mathbf{x}, \mathbf{p}))], \end{aligned} \tag{5}$$

where the generator G tries to minimize this objective function, while the discriminator D tries to maximize it.

The final PoseGAN's loss function is then defined as:

$$\mathcal{L}_{\text{PoseGAN}}(G) = \lambda \mathcal{L}_{\text{GAN}} + \mathcal{L}_{L^2}, \tag{6}$$

where λ is a weight parameter, controlling the balance between the GAN loss and L^2 loss.

4.2. Pose-Guided Face Generation Network

The face shape parameters are generated by the deep 3D face reconstruction method [17]. The generated identity parameters α_{id} could be different for each frame. These differences are introduced by camera parameters, speaker position and inaccurate expression shape. The conventional methods, e.g., [16], only generated expression parameters α_{exp}, which are not suitable for our case. Inspired by the VOCA network [16], we propose a pose-guided face shape generation method (PGFace), which includes the head pose parameters as input for estimating the change of face shape to make up the difference. We concatenate audio features $\mathbf{x} \in \mathbb{R}^{29 \times 16}$ and head pose parameters $\mathbf{p} \in \mathbb{R}^6$ for each frame as input for the network. The network output is the corresponding face shape parameters $[\alpha_{\text{id}}, \alpha_{\text{exp}}]$. Please refer to the supplementary material for the details of our PGFace network.

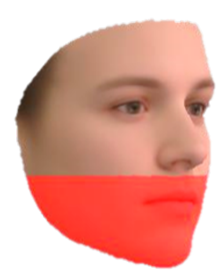

Figure 6. The lower part of the face, shown in red, is used to calculate higher weights for the vertex-level loss.

Based on our experiments, audio shows a higher correlation with the lower part of the face as shown in Figure 6. We employ a vertex-based loss function, which attaches a 10-times weight $\mathbf{m}$ on the lower part of the face model. The loss functions can be formally represented as:

$$\begin{aligned}\mathcal{L}_{\text{shape}} &= \mathbb{E}_{\mathbf{v},\mathbf{f}}[\|(\mathbf{v}-\mathbf{f})\odot\mathbf{m}\|^2],\\ \mathcal{L}_{\text{s-motion}} &= \mathbb{E}_{\mathbf{v},\mathbf{f}}[\|((\mathbf{v}_{\text{next}}-\mathbf{v})-(\mathbf{f}_{\text{next}}-\mathbf{f}))\odot\mathbf{m}\|^2],\end{aligned} \tag{7}$$

where $\mathbf{v}$ denotes the ground-truth face vertices, and $\mathbf{f}$ represents the generated face vertices; $\mathbf{v}_{\text{next}}$ and $\mathbf{f}_{\text{next}}$ indicate the values of $\mathbf{v}$ and $\mathbf{f}$ in the next frame; the mask $\mathbf{m}[i] = 10$ if the vertex i is in the lower part of the face, otherwise $\mathbf{m}[i] = 1$. The $\odot$ operation means element-wise product. The motion loss $\mathcal{L}_{\text{s-motion}}$ represents the vertex displacement between neighboring frames in sequence.

The PGFace's loss function is then defined as:

$$\mathcal{L}_{\text{PGFace}} = \mu_1\mathcal{L}_{\text{shape}} + \mu_2\mathcal{L}_{\text{s-motion}}, \tag{8}$$

where μ_1 and μ_2 balance the shape and motion losses.

4.3. Implementation Details

The networks for head pose and face shapes are trained on an Nvidia GTX 1080 Ti using Adam [27] with a batch size of 64 and a learning rate of 10^{-4}. We divide our dataset using a train-val-test split of 7-1-2. In PoseGAN training section, we first centralize and normalize the head poses as described in our dataset section. The frame rate of our video is 30fps. We use a 256-frame sliding window as a training sample and the output is 256-frame head pose sequence. The sliding distance between neighbors is 5 frames. During training, α and β are set to 1 and 10. The value of λ is 0.01. A total of 150 epochs are trained. The best performing model on the validation set is selected. In PGFace training section, the network is learned from audio features and head pose parameters with 100 epochs. The window size used for PGFace is 16 and the output is the face shape in the 8th frame. The values of μ_1 and μ_2 are 1 and 10, respectively.

5. Experimental Results

5.1. Evaluation of Feasibility: Correlation Verification

Since our goal is to generate the head pose sequence from speech, we first verify that there is a correlation between a person's speech and his/her head pose. DeepSpeech is used to extract the speech feature for each frame and OpenFace is used to extract the corresponding head pose. Each frame corresponds to 29 speech features and 6 values of head pose. We calculate the correlation between the speech and head pose sequence on 256 frames by Pearson's correlation function, to obtain the 29×6 features for each 256-frames clip:

$$F(i,j) = \frac{\sum_{k=0}^{255}(S_{ik}-\bar{S}_i)(H_{jk}-\bar{H}_j)}{\sqrt{\sum_{k=0}^{255}(S_{ik}-\bar{S}_i)^2}\sqrt{\sum_{k=0}^{255}(H_{jk}-\bar{H}_j)^2}}, \tag{9}$$

where $i \in [0,5], j \in [0,28]$. S_{ik} and H_{jk} are ith speech feature and jth head pose value in the kth frame. $\bar{S}_i$ and $\bar{H}_j$ are their average values across 256 frames, respectively.

We then train a one-class Support Vector Machine (SVM) [36] with 29×6 features on real data samples. As shown in Table 1, we replace the head pose sequence in the test dataset of each person with a random head pose sequence. The results of one-class SVM are reduced when replacing the original head pose sequence, which indicates the existence of correlation between the head pose sequence and the speech of a particular person. Furthermore, other works [7, 49] have also verified the direct correlation between audio and pose.

5.2. Quantitative Evaluation

We compare our PoseGAN to the following four head pose generation methods.

The mean head pose: Most of 2D talking face videos [6, 12, 14, 18, 38, 39, 45, 46, 54] and 3D talking faces [41, 55, 26, 30, 52, 52, 44, 16] can only generate fixed head pose now. In most of the time, the head is in a resting position and orientation during speech (see Figure 2). Thus we use mean pose to compare with these 2D and 3D methods.

Randomly chosen head pose sequence: Another simple way to quickly generate the head pose sequence is to randomly select a head pose sequence from the dataset. Such choice is somehow reasonable since they are true head poses. This random method is widely used in 2d talking face methods [40, 32, 42, 47]. Although the re-timing technique is used in [40] to increase the authenticity, this method is still a random pose sequence and cannot generate new head poses based on speech. Therefore, such a randomly selected head pose sequence does not correspond to the input audio.

Nearest neighboring (NN) pose: The head pose chosen by this method is close to the real head pose in the audio feature space. For each test audio, the head pose sequence with the closest audio feature in the training set is selected as the final output.

Convolutional neural network (CNN): Conventional CNN [16] achieved state-of-the-art results with 3D face shape generation. Few 2D talking face methods [49, 10] also use CNNs to generate head pose in videos. For example, Yi *et al.* [49] used LSTM to generate head pose sequences in their talking face video. However, the head pose estimation is a cross-modal prediction task. We find that the head pose sequence generated without GAN tends to be close to a static head pose. It is hard to consider the results of CNN as realistic head pose sequences.

Table 1. One-class SVM results for verifying the correlation between the speech and head pose sequence.

Audio Feature	Corresponding Head Pose	Random Head Pose				
		Clinton	Obama	Sanders	Trump	Warren
Clinton	**0.90**	0.75	0.74	0.72	0.73	0.72
Obama	**0.88**	0.47	0.52	0.44	0.46	0.46
Sanders	**0.83**	0.72	0.72	0.71	0.71	0.73
Trump	**0.85**	0.74	0.76	0.74	0.73	0.72
Warren	**0.80**	0.60	0.59	0.57	0.55	0.59

Table 2. L^2 distance with head pose and motion on the test set.

Method	$\mathcal{L}_{pose}$	$\mathcal{L}_{motion}$
Mean	0.90	0.12
Random	1.21	0.15
NN	1.18	0.14
CNN	0.82	0.11
Our PoseGAN	0.89	0.12

5.2.1 L^2 Distance Comparison

To compare our PoseGAN architecture to all these four baselines, we select 10 videos in the test dataset and calculate the L^2 pose distance and motion distance of each method. In Table 2, the random method and nearest neighbor are performing significantly worse in accuracy. This is because those two methods have no constraints on the head pose. The distance of the mean head pose method is low because the speaker is mostly in a static head pose while speaking. The distance with CNN is lowest because only the pose loss and motion loss are used for training. As discussed before, the generated head pose sequence with CNN tends to be static. The L^2 distance results of our PoseGAN outperforms most of the baseline methods except for CNN. This is expected, as we add GAN loss to our generator to produce more realistic and reasonable head pose sequences.

5.2.2 Head Pose Classifier

A head pose classifier is optimized on our training set with 5 identities in order to evaluate the head pose results obtained by different methods. Classic CNN and dense layer structure were used to implement the head pose classifier, where the output of the last fully connected layer was set to 5. The input is the head pose motion on 256 frames. We choose the best performance on the validation set, which has an accuracy rate of 92% in test set. As shown in Table 3, the result of our method is closest to the true head pose distribution. If the confidence value is greater than 0.5, most of the data in this category are correctly classified. The results of Mean and CNN methods are close to a random distribution (0.2), which deviate from the true head pose distribution.

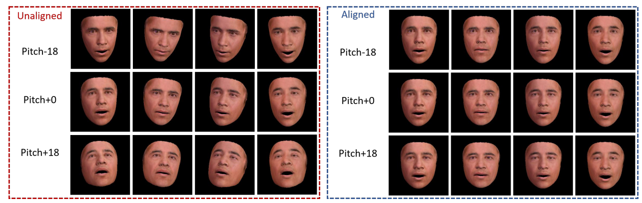

Figure 7. The rendering results of face shape under different head poses with the same audio.

Figure 8. Comparison to state-of-the-art 3D face generation methods including VOCA [16] and Karras *et al.* [26].

5.3. User Study

5.3.1 Head Pose

One user study is designed to compare our method to the ground truth and all baselines. We prepared 100 pairs of videos. Each of them includes two videos: one is the talking face with ground truth head pose sequence; another is generated by one of the four baselines or our method. Three ground truth videos are given to participants to learn before the task. Participants are required to select the better one from each pair. Among the 100 pairs, 60 sets of videos are 4 seconds in length, 25 sets of videos are 8 seconds, and 15 sets of videos are 12 seconds. 50 persons participate in the study to evaluate the rationality and authenticity of the synthesized 3D talking faces.

We present the results in Table 4. For each video pair (synthesized and ground truth) of different lengths, we measure the probability of selecting the face model generated by the method as the better one. Intuitively, a higher prob-

Table 3. The result of the head pose classifier. Each value represents the confidence of correct classification.

Method	Clinton	Obama	Sanders	Trump	Warren	**Avg**
Mean	0.51	0.01	0.21	0.27	0.00	0.20
CNN	0.13	0.06	0.61	0.30	0.01	0.22
Our PoseGAN	**0.86**	**0.87**	**0.70**	**0.52**	**0.65**	**0.72**

Figure 9. Results of our framework. From the input audio, we generate the 3D talking head with personalized pose dynamics by comparison methods and our method. The head pose and face result are sampled in every 60 frames (2 seconds).

Table 4. User study results. Each value (%) represents the probability that the user selected the generated pose (the true pose is not selected). A larger value indicates that the result is more realistic.

Method	4 seconds	8 seconds	12 seconds
Mean	14.2	14.8	12.0
Random	27.3	20.4	21.3
NN	20.3	16.0	16.7
CNN	16.5	18.8	12.7
Our PoseGAN	**34.3**	**28.4**	**30.0**

ability means the better performance for that method. We found that CNN performs poorly in the user study, while the random method performs relatively better on the 4-second videos but poorly on videos of longer times. It is shown that our method works well on all videos of different lengths.

5.3.2 Face Shape

Our second user study is to show the comparison between our pose-corrected face shape with fixed identity shape. Participants select more realistic videos among three groups of 50-second video pairs. Most of them think our results are more realistic (73%) than the fixed identity method (27%). Detailed results are provided in the supplementary material.

5.4. Qualitative Evaluation

5.4.1 Pose-Dependent Facial Shape Correction

We propose a face shape generation method to complement the face shape rendering result with head pose information. To show the influence of head poses on face shapes, we conduct three experiments using different head pose parameters: i) use the normal head pose sequence ($Pitch + 0$); ii) increase the pitch angle by 18 degrees ($Pitch{+}18$); iii) control the pitch angle downward by 18 degrees ($Pitch - 18$). Results are shown in Figure 7. To visualize the results in a clear way, we also align the face shapes. Observing that in both cases, the head pose has a noticeable effect on producing more reasonable face shape with the same input audio.

5.4.2 Ablation Studies

Different variants are compared for head pose generation including no-motion loss and our methods. No-motion loss results in jitter problems and no-initial pose leads to discontinuities. In contrast, our proposed PoseGAN generates realistic head pose sequences. More results can be found in the supplementary video. In the supplementary video, we show that our method is still applicable under different noises. Although our training language is based on English, we also show that the method applies to multiple language environments.

5.4.3 Comparison with Other Methods

In the supplementary video, we compare our results with state-of-the-art 3D face generation methods including VOCA [16] and Karras *et al.* [26]. In figure 8, we show a representative frame of results for generating the corresponding 3D faces based on input audio.

5.4.4 More Visualization Results

Figure 9 shows the visualization results of our framework. Given input audio, we generate the 3D talking face with personalized pose dynamics. From top to bottom, they are input audio, head pose sequence, and face shape with head pose. We can see that the head pose sequence of the mean method remains the same. The head pose sequence of the CNN method tends to be close with the mean pose and changes slightly. The head poses generated by Random and NN methods change sharply. However, the head pose sequence generated by our method changes stably and reasonably. Please refer to the supplementary video for the detailed results.

6. Conclusion and Future Work

To the best of our knowledge, this is the first work to generate 3D talking face with personalized pose dynamics based on audio. Our 3D face database includes audio, head pose sequence, and face shape parameters. The PoseGAN is trained to generate the head pose sequence, with the initial head pose loss constraint and motion loss function, which guarantees the continuity of head pose sequence in long term. The PGFace network is designed for pose-dependent facial shape correction, which makes the face shape rendering results more realistic. Our experiments verify the effectiveness of our approach, and our synthesized 3D talking head looks more realistic than other baselines.

As mentioned in Section 2, we would like to integrate our pose-dynamics-empowered 3D talking head as a basic building block for synthesizing audio-driven 2D videos of facial reenactment [43], to further improve the realism of head motion in the synthesized videos, as well as extending it for text-based facial animation in our future work.

References

[1] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li. Protecting world leaders against deep fakes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 38–45, 2019. 3

[2] T. Alashkar, B. B. Amor, M. Daoudi, and S. Berretti. A 3d dynamic database for unconstrained face recognition. In *Proceedings of 5th International Conference on 3D Body Scanning Technologies*, pages 357–364, 2014. 3

[3] B. Amos, B. Ludwiczuk, and M. Satyanarayanan. Openface: A general-purpose face recognition library with mobile ap-

plications. Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016. 3

[4] R. Anderson, B. Stenger, V. Wan, and R. Cipolla. Expressive visual text-to-speech using active appearance models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3382–3389, 2013. 2

[5] V. Blanz, T. Vetter, et al. A morphable model for the synthesis of 3d faces. In *Siggraph*, volume 99, pages 187–194, 1999. 4

[6] C. Bregler, M. Covell, and M. Slaney. Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 353–360, 1997. 6

[7] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan. Rigid head motion in expressive speech animation: Analysis and synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1075–1086, 2007. 6

[8] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013. 3, 4

[9] Y. Chang, M. Vieira, M. Turk, and L. Velho. Automatic 3d facial expression analysis in videos. In *International Workshop on Analysis and Modeling of Faces and Gestures*, pages 293–307. Springer, 2005. 3

[10] L. Chen, G. Cui, C. Liu, Z. Li, Z. Kou, Y. Xu, and C. Xu. Talking-head generation with rhythmic head motion. *arXiv preprint arXiv:2007.08547*, 2020. 6

[11] 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)*, pages 520–535, 2018. 1, 2

[12] 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*, pages 7832–7841, 2019. 6

[13] S. Cheng, I. Kotsia, M. Pantic, and S. Zafeiriou. 4dfab: A large scale 4d database for facial expression analysis and biometric applications. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5117–5126, 2018. 3

[14] J. S. Chung, A. Jamaludin, and A. Zisserman. You said that? *arXiv preprint arXiv:1705.02966*, 2017. 1, 2, 6

[15] D. Cosker, E. Krumhuber, and A. Hilton. A facs valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling. In *2011 International Conference on Computer Vision*, pages 2296–2303. IEEE, 2011. 3

[16] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10101–10111, 2019. 1, 2, 3, 5, 6, 7, 9

[17] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 4, 5

[18] T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. *ACM Transactions on Graphics (TOG)*, 21(3):388–398, 2002. 6

[19] B. Fan, L. Wang, F. K. Soong, and L. Xie. Photo-real talking head with deep bidirectional lstm. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4884–4888. IEEE, 2015. 2

[20] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. Van Gool. A 3-d audio-visual corpus of affective communication. *IEEE Transactions on Multimedia*, 12(6):591–598, 2010. 3

[21] 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*, pages 3497–3506, 2019. 4

[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[23] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014. 1, 2

[24] P. Hong, Z. Wen, and T. S. Huang. iface: a 3d synthetic talking face. *International Journal of Image and Graphics*, 1(01):19–26, 2001. 3

[25] S. A. Jalalifar, H. Hasani, and H. Aghajan. Speech-driven facial reenactment using conditional generative adversarial networks. *arXiv preprint arXiv:1803.07461*, 2018. 1, 2

[26] 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)*, 36(4):94, 2017. 2, 6, 7, 9

[27] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[28] R. Kumar, J. Sotelo, K. Kumar, A. de Brébisson, and Y. Bengio. Obamanet: Photo-realistic lip-sync from text. *arXiv preprint arXiv:1801.01442*, 2017. 3

[29] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4d scans. *ACM Transactions on Graphics (TOG)*, 36(6):194, 2017. 2

[30] Y. Liu, F. Xu, J. Chai, X. Tong, L. Wang, and Q. Huo. Video-audio driven real-time facial animation. *ACM Transactions on Graphics (TOG)*, 34(6):182, 2015. 2, 6

[31] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[32] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 484–492, 2020. 1, 6

[33] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4

[34] S. Sako, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Hmm-based text-to-audio-visual speech synthesis. In *Sixth International Conference on Spoken Language Processing*, 2000. 2

[35] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun. Bosphorus database for 3d face analysis. In *European Workshop on Biometrics and Identity Management*, pages 47–56. Springer, 2008. 3

[36] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001. 6

[37] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE, 2018. 3

[38] S. Sinha, S. Biswas, and B. Bhowmick. Identity-preserving realistic talking face generation. *arXiv preprint arXiv:2005.12318*, 2020. 6

[39] Y. Song, J. Zhu, D. Li, X. Wang, and H. Qi. Talking face generation by conditional recurrent adversarial network. *arXiv preprint arXiv:1804.04786*, 2018. 6

[40] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017. 1, 2, 6

[41] 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)*, 36(4):93, 2017. 2, 6

[42] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *European Conference on Computer Vision*, pages 716–731. Springer, 2020. 1, 6

[43] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *Proceedings of the European Conference on Computer Vision*, 2020. 9

[44] D. Vlasic, M. Brand, H. Pfister, and J. Popovic. Face transfer with multilinear models. In *ACM SIGGRAPH 2006 Courses*, pages 24–es. 2006. 2, 6

[45] K. Vougioukas, S. Petridis, and M. Pantic. End-to-end speech-driven realistic facial animation with temporal gans. In *CVPR Workshops*, pages 37–40, 2019. 1, 2, 6

[46] K. Vougioukas, S. Petridis, and M. Pantic. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, pages 1–16, 2019. 6

[47] X. Wen, M. Wang, C. Richardt, Z.-Y. Chen, and S.-M. Hu. Photorealistic audio-driven video portraits. *IEEE Transactions on Visualization and Computer Graphics*, 2020. 1, 6

[48] L. Xie and Z.-Q. Liu. Realistic mouth-synching for speech-driven talking face using articulatory modelling. *IEEE Transactions on Multimedia*, 9(3):500–510, 2007. 2

[49] 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

[50] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 211–216. IEEE, 2006. 3

[51] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2013. 3

[52] Y. Zhang and W. Wei. A realistic dynamic facial expression transfer method. *Neurocomputing*, 89:21–29, 2012. 2, 6

[53] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3438–3446, 2016. 3

[54] 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*, volume 33, pages 9299–9306, 2019. 1, 2, 6

[55] 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)*, 37(4):161, 2018. 1, 2, 6