AnyoneNet

Synchronized Speech and Talking Head Generation for Arbitrary Persons

Wang, Xinsheng; Xie, Qicong; Xie, Lei; Zhu, Jihua; Scharenborg, Odette

DOI
10.1109/TMM.2022.3214100

Publication date
2023

Document Version
Final published version

Published in
IEEE Transactions on Multimedia

Citation (APA)
Wang, X., Xie, Q., Xie, L., Zhu, J., & Scharenborg, O. (2023). AnyoneNet: Synchronized Speech and Talking Head Generation for Arbitrary Persons. *IEEE Transactions on Multimedia*, *25*, 6717-6728. https://doi.org/10.1109/TMM.2022.3214100

AnyoneNet: Synchronized Speech and Talking Head Generation for Arbitrary Persons

Xinsheng Wang, Qicong Xie, Jihua Zhu, *Member, IEEE*, Lei Xie, *Senior Member, IEEE*, and Odette Scharenborg, *Senior Member, IEEE*

***Abstract*—Automatically generating videos in which synthesized speech is synchronized with lip movements in a talking head has great potential in many human-computer interaction scenarios. In this paper, we present an automatic method to generate synchronized speech and talking-head videos on the basis of text and a single face image of an arbitrary person as input. In contrast to previous text-driven talking head generation methods, which can only synthesize the voice of a specific person, the proposed method is capable of synthesizing speech for any person. Specifically, the proposed method decomposes the generation of synchronized speech and talking head videos into two stages, i.e., a text-to-speech (TTS) stage and a speech-driven talking head generation stage. The proposed TTS module is a face-conditioned multi-speaker TTS model that gets the speaker identity information from face images instead of speech, which allows us to synthesize a personalized voice on the basis of the input face image. To generate the talking head videos from the face images, a facial landmark-based method that can predict both lip movements and head rotations is proposed. Extensive experiments demonstrate that the proposed method is able to generate synchronized speech and talking head videos for arbitrary persons, in which the timbre of the synthesized voice is in harmony with the input face, and the proposed landmark-based talking head method outperforms the state-of-the-art landmark-based method on generating natural talking head videos.**

***Index Terms*—Avatar, facial landmark, speech synthesis, talking head generation.**

I. Introduction

AUTOMATICALLY generating videos in which synthesized speech is synchronized with lip movements in a talking head [1] has great potential in many human-computer interaction scenarios, e.g., computer games and virtual reality, and in the field of entertainment, e.g., visual dubbing and short video creation. Intuitively, the synchronized speech and facial animation should not only be dynamically consistent, i.e., the lip and jaw movements should be synchronized to the produced speech, but also perceptively consistent, i.e., the voice should sound like it could be uttered by the person in the video. Otherwise, the generated video would be perceived as unreal and strange. One way to generate talking head videos is to train a model with paired talking head videos and speech, similar to that in ObamaNet [2]. However, a model trained in this fashion can only be used for those persons/faces that are part of the training process, and such a method thus has very limited generalization. In contrast, in this paper, we present a method that uses a still face image of any person and text as input to generate a talking head video with a voice that could have been that of the person in the input face image. This method thus works for anyone.

In terms of input (driven) information, the talking head generation methods can be categorized into speech-driven, text-driven, and video-driven [3], [4] methods, i.e., taking audio, text, or video as input to guide the lip movements. Compared to the speech-driven and video-driven methods, the text-driven method is more flexible, as it allows users to create any new content because it is not dependent on an existing speech corpus or source videos. Although there are several text-driven methods that directly use textual phonetic labels to predict the visual speech [5], most of the recent text-driven methods [2], [6], [7] decompose the text-to-video process into separate text-to-speech (TTS) and speech-to-video processes, i.e., 1) synthesize speech with text as input using the TTS module and 2) perform the speech-driven talking head generation with synthesized speech as input. As no extra alignment between synthesized speech and the generated video is required in the text-to-speech-to-video method, this text-to-speech-to-video strategy is adopted in the current work. This method allows us to use the intermediate representation of synthesized speech, e.g., spectrograms, to generate the synchronized video.

A high-level overview of our system is illustrated in Fig. 1. The input face image not only provides identity information for the video generation but also for the TTS system. Specifically, the TTS module tries to synthesize speech with a voice that sounds like it could have been uttered by the person in the input image. Note, however, that unlike research on the reconstruction of face images conditioned on the voice [8], [9], [10], we do not

Manuscript received 11 December 2021; revised 19 February 2022, 27 May 2022, 7 August 2022, and 21 August 2022; accepted 3 October 2022. Date of publication 12 October 2022; date of current version 1 November 2023. This work was supported in part by the National Key R&D Program of China under Grants 2018AAA0102504 and 2020AAA0108600, and in part by the Key Research and Development Program of Shaanxi under Grants 2021GY-025 and 2021GXLH-Z-097. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xavier Alameda-Pineda. *(Corresponding authors: Jihua Zhu; Lei Xie.)*

Xinsheng Wang is with the School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China, also with the School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China, and also with the Multimedia Computing Group, Delft University of Technology, 2628CD Delft, The Netherlands (e-mail: wangxinsheng@stu.xjtu.edu.cn).

Qicong Xie and Lei Xie are with the School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: xieqicong@mail.nwpu.edu.cn; lxie@nwpu.edu.cn).

Jihua Zhu is with the School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: zhujh@xjtu.edu.cn).

Odette Scharenborg is with the Multimedia Computing Group, Delft University of Technology, 2628CD Delft, The Netherlands (e-mail: O.E.Scharenborg@tudelft.nl).

Digital Object Identifier 10.1109/TMM.2022.3214100

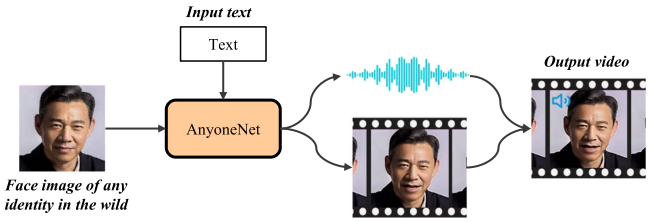

Fig. 1. Illustration of generating a talking head video with synchronized speech. The input is text and a still face image, while the output is a talking head video with synchronized speech in which the synthesized voice is in harmony with the person's portrait in the input image and output video.

argue that there is a strong relationship between one's portrait and his or her voice. Here, our goal is simply to synthesize a voice that is in harmony with the face in the still image, in order to make the generated voice and the face in the video look natural.

In terms of the video generation process, both the lip movements and the head movements are predicted using facial landmarks. Different from the state-of-the-art landmark-based method of [11], in which the head movements are treated as the shift of facial landmarks, here, the head orientation is presented as quaternions, which allows us to predict head rotations.

To sum up, the main contribution of this paper is the proposed method that is able to generate a talking head video with synchronized speech in a personalized voice, using text and a face image of an arbitrary person as input. The proposed method deviates from previous work which either cannot produce personalized voices for arbitrary persons [11], [12], or the voiced talking head video generation can only be used for a single person [2].

The rest of this paper is organized as follows: Section II reviews related work on TTS and speech-driven talking head generation. Section III describes the proposed method. Section IV introduces the databases that are used to train the different modules and presents extensive experimental results. Section V discusses the limitations of the proposed method and the ethical considerations are also discussed here. Finally, the paper concludes in Section VI. Demos of AnyoneNet can be found on the website.[1]

II. Related Work

A. Text-to-Speech Synthesis

Similar to some recent text-driven talking head generation methods [6], [7], [13], our method uses TTS to synthesize the speech signal. The goal of a TTS system is to synthesize human-like speech from a natural language text input.

Most of the recent neural-based TTS methods consist of two stages. The first stage is to predict low resolution intermediate audio features, typically Mel-spectrograms [14], [15], [16], vocoder features [17], or linguistic features [18], from a text input. The second stage is to synthesize the raw waveform audio from the predicted intermediate representation [19], [20], [21], [22], [23]. Most recently, some end-to-end TTS models also have been proposed [24], [25], [26] without using the intermediate features. However, for the talking head generation task, the intermediate representations of the two-stage approach are useful as intermediate representation Mel-spectrograms are used in the video generation process. Therefore, a typical two-stage TTS system is adopted in the proposed method.

TTS systems can be categorized into single-speaker TTS and multi-speaker TTS systems. The single speaker TTS systems are trained on a speech corpus recorded by a single person, e.g., LJspeech [27], and thus can only synthesize speech with a specific voice. In contrast, the multi-speaker TTS systems are able to produce multiple voices by training with speech corpus recorded by many speakers. In early research, a multi-speaker TTS model was typically trained as an average voice model using all speakers' data, which was then adapted to an individual speaker [28], [29], [30]. In the recent neural-based methods, conditioning on speaker embeddings has been a popular strategy. Typically, a speaker representation is extracted by a speaker embedding model and is then used as the conditional attribute in a TTS model [31], [32], [33], [34]. For instance, in [31], speaker embedding vectors are obtained from a separately trained speaker verification model, and the TTS model Tacotron2 [14] conditioned on the speaker embeddings is used for multi-speaker speech synthesis.

An advantage of the embedding-based multi-speaker TTS is that speaker embeddings can be extracted from any speaker, also speakers who do not exist in the training set, making it possible for the multi-speaker TTS to be used for any person. To build a talking head generation model that can be used for any person, the embedding-based multi-speaker TTS method is adopted in our TTS module. Different from existing multi-speaker TTS systems, in which the reference speaker embedding is obtained from speech recorded by this speaker, in our method the speaker embedding is based on a person's face image.

B. Speech-Driven Talking Head Generation

The goal of speech-driven talking head generation is to create a talking head from a still face image in which lip movements are synchronized with the speech signal. Early methods in this field were usually based on a pre-defined dictionary of visemes, which typically refers to the mouth shape of a given phoneme, and the model's task was to learn the mappings between the speech signal and the lip articulations [35], [36]. There are also many efforts from computer graphics to construct 3D models to predict the relationships between speech and 3D facial parameters. [37], [38], [39], [40], [41]. However, these models heavily depend on the captured 3D facial graphic parameters, which is much more challenging to be obtained than 2D video recording. Therefore, normally, only one or several persons' videos are recorded with 3D facial graphic parameters, making them hard to be used to build a model that can be used for arbitrary persons that are not seen during the training process.

Compared to 3D facial graphic parameters, facial landmarks, or facial key-points, are simpler representations that can be used to present the face and mouth shape. Moreover, benefiting from the recently developed robust and efficient off-the-shelf landmark detectors [42], [43], facial landmarks can be easily

[1]Demos can be found at https://www.youtube.com/watch?v=jTb9pyzlHaU

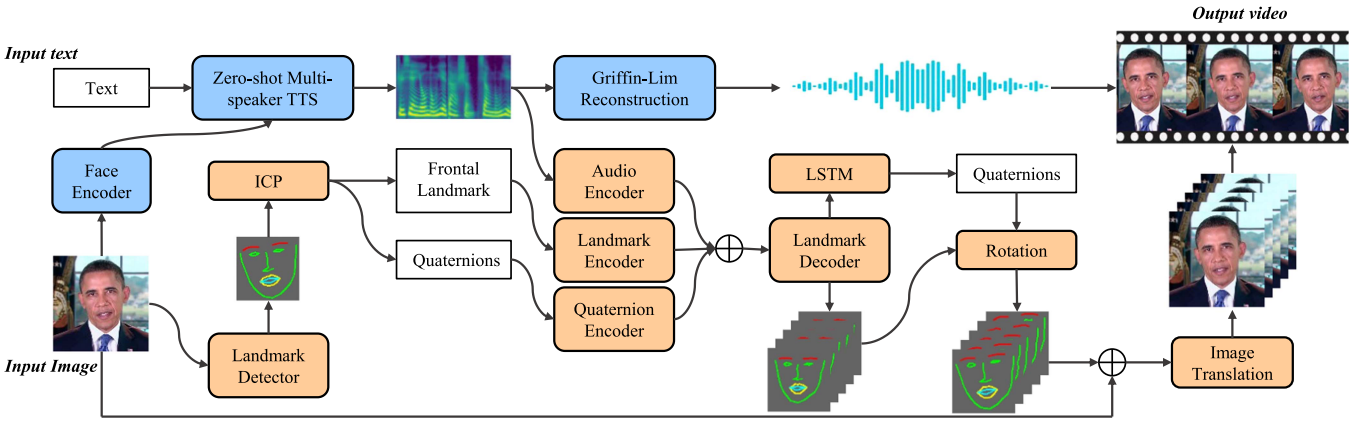

Fig. 2. Overall framework of the proposed method. The input is a single still face image and some text. ICP (Iterative Closest Point) is used to register the facial landmarks to a front-facing standard facial template, and the resulting rotational parameters are presented as quaternions. ⊕ indicates concatenation.

obtained. A face landmark is the position of key points on a face, such as the tip of the nose and the center of the eye. Facial landmarks can be used to represent facial-related characteristics, e.g., face shapes, head poses, and mouth shapes, and can easily be used to build mappings between facial landmarks and the facial expression in a photo. Recently, compact facial landmarks have been popular intermediate representations to bridge the gap between the raw speech signal and photo-realistic videos [6], [12], [44], [45], [46], [47]. However, these methods suffer from several limitations, e.g., they can only be used for the person that is used as training data [44] thus not for arbitrary persons, they depend on reference videos to provide pose information [44], [46], [47], [48], or they cannot predict head movement but only static head poses [12]. To address these issues, MakeItTalk [11] was developed. MakeItTalk disentangles linguistic information and information about the identity of the speaker in the input speech signal. Subsequently, the linguistic information is used to guide (drive) the lip movements while speaker identity information is used to drive the facial expressions and head poses. By predicting shifts of landmarks rather than landmarks with specific shapes, MakeItTalk can be easily used for arbitrary identities.

Recently, end-to-end models have shown promising results in generating accurate lip movements [49], [50], [51], [52], [53], [54]. However, these methods can only generate a talking head that has a fixed head pose, which limits the naturalness of the generated videos. In order to generate a talking head with more natural head movement, in [55], a source video is used to provide head pose information which is used to give the predicted talking head the same pose movements. However, this method also suffers from the limitation that can only be used for identities from the training database rather than for an arbitrary person. Some other end-to-end cross-modal generation tasks between audio and visual, such as musical performance video generation based on audio [56], are also limited to the scene or identity from the training database.

With the goal to build talking head videos with natural head movements for an arbitrary person, we follow the basic idea in [11] and take landmarks as the intermediate representation to present the lip movement and head pose. Following [11], the lip movements are represented as shifts of key points. Different from [11] which represents the head movement as key points' shifts, we treat head movements as rotations, which allows the model to predict more natural head poses. Furthermore, as both speech synthesis and video generation are considered in this work, to simplify the pipeline, the input speech in the landmark prediction module is presented as Mel-spectrograms which is the same as the intermediate representation in the TTS system.

III. Method

A. Overall Framework

The overall framework of the proposed method is shown in Fig. 2. The input of this framework is a text and a still face image of a person. The output is a talking head video of this person where the person speaks the text with a voice that is conditioned on the face image. The proposed framework consists of two sub-modules, i.e., a speech synthesis module (top flow in Fig. 2) and a video generation module (bottom flow of Fig. 2). The speech synthesis module is a zero-shot multi-speaker TTS model, with text and a face embedding vector as input. This face embedding vector, which is to provide speaker identity information, is obtained via a pre-trained face encoder (see Section III-B).

The video generation module is a speech-driven talking-head video generation module, which is decomposed into two steps: landmark prediction and video generation. In the first step, the goal is to generate a sequence of facial landmarks with the input of synthesized speech intermediate representations, i.e, Mel-spectrograms, and the initial facial landmarks extracted from the input image. Then, with the generated landmarks and the input face image, we can generate a sequence of photo-realistic images, and the image sequence is then converted into the talking head video. Here, an off-the-shelf face 3D landmark detector [43] is used to extract the facial landmarks for the training video data and also for the still image in the reference stage.

B. Face Encoder

The face encoder is to encode the face image into an embedding vector that provides speaker information to the multi-speaker TTS system (see Section III-C). Training such a face

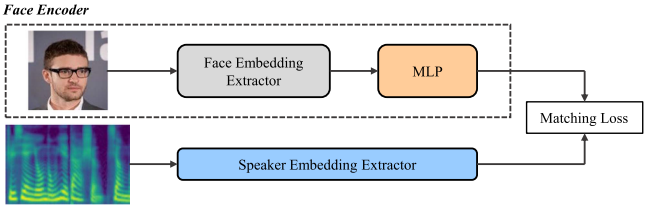

Fig. 3. Framework to train the face encoder.

encoder could intuitively be done together with the whole multi-speaker TTS system; however, in order to do so, a speech database paired with the speaker's face images is needed. Unfortunately, no such database for multi-speaker TTS is available. Fortunately, several speech-visual (talking video with face image frames) paired databases exist, which were originally collected for, e.g., speaker verification [57], [58] or lip-reading [59], [60]. These databases allow us to train the face encoder separately from the TTS system.

A multi-speaker TTS model is typically based on the speaker embedding that is extracted from speech of one speaker to provide this speaker's identity information. The basic motivation to train the face encoder is that if we can map the face feature into the speaker embedding space, in which the speaker embedding and the face embedding from the same speaker ideally same to each other, then the speaker embedding can be replaced by the face embedding to provide the speaker identity information in the inference stage, achieving the goal of face-based multi-speaker TTS.

To this end, the face encoder can be trained supervised by a pre-trained speaker embedding network with paired face-speech database. As shown in Fig. 3, the face encoder is trained under a typical teacher-student framework, where the teacher is the pre-trained speaker embedding extractor and the student is the face encoder.

Architecture: The speaker encoder, which works as the teacher and is named as speaker embedding extractor in Fig. 3, is based on the ResNet-34 [61] as that in [58]. Here, the last fully connected layer is dropped, and the output speaker embedding is represented as a 1024-D vector with L2 normalization.

The face encoder consists of an off-the-shelf face feature extractor that is trained for face recognition with face image as input[2] [62] and an MLP block with two linear transformation layers. The output of the face feature extractor is a 512-D vector. The hidden unit size of the MLP is 2048, and the output size is the same as that of the speaker embedding vector which is 1024. The L2 normalization layer is also added after the MLP in the face encoder as in the speaker embedding extractor.

Training: To train the face encoder, we have to first pre-train the speaker embedding extractor. Here, we trained this speech-based speaker encoder in a speaker verification task with the large margin softmax loss [63]. The parameters of this speaker encoder would not be updated further during the training of the face encoder.

[2]https://github.com/timesler/facenet-pytorch

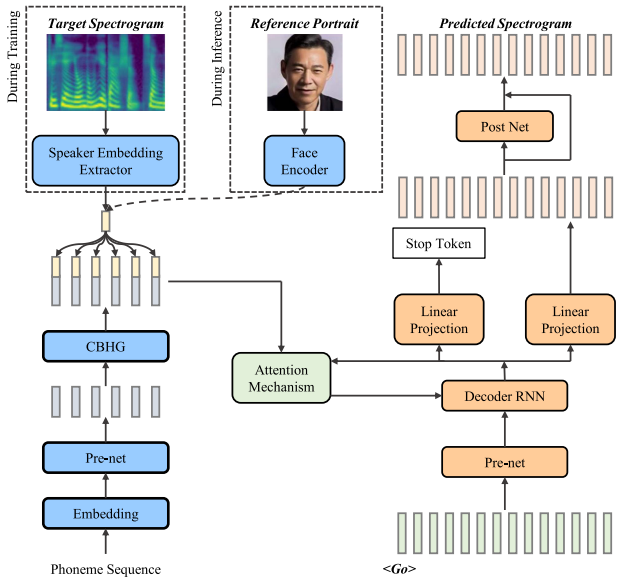

Fig. 4. Framework of the face-conditioned multi-speaker TTS.

To project a face image feature into the matched speech embedding space, i.e., to minimize the distance between a matched face embedding and speech embedding pair, the Masked Margin Softmax (MMS) [64] that is designed for visually grounded speech representation learning is adopted as the matching loss. During the training process, parameters of the face feature extractor and speaker embedding extractor are fixed, and only parameters of the MLP are updated. With the trained image encoder, which consists of the off-the-shelf face feature extractor and trained MLP layers, we can obtain the final face embedding that is used to replace the speaker embedding in the TTS system.

C. Face-Conditioned Multi-Speaker TTS

Due to the lack of a speech-face paired database, we cannot train a face-conditioned multi-speaker TTS directly. Instead, we use the standard method to train a multi-speaker TTS model with a text-speech paired database [65]. Specifically, during training, the speaker information is provided through the speaker embedding extracted by the same speaker encoder used in training the face encoder (see bottom flow in Fig. 3). During inference, the speaker embedding is replaced with the face embedding, which results in the face-conditioned multi-speaker TTS.

The proposed face-conditioned multi-speaker TTS framework is shown in Fig. 4. The Tacotron-based model [14], [66] is used as the Mel-spectrogram prediction model. Specifically, the multi-speaker TTS model has a typical attention mechanism-guided encoder-decoder architecture. The encoder (the left part in Fig. 4) follows Tacotron's [66] encoder that consists of a pre-net and a CBHG block. The text that works as the input to the encoder is represented as a sequence of phonemes that are then embedded into a vector sequence. The decoder (the right part in Fig. 4) follows Tacotron2's [14] decoder that consists of a pre-net, RNN decoder, and a post-net. Additionally, between the RNN decoder and post-net, two linear projection layers are used to predict Mel-spectrograms and stop tokens, respectively.

The attention mechanism is to provide a soft alignment between the encoder states and the target Mel-spectrograms. Here, the GMMV2b attention mechanism [67], which shows better robustness on inferring long utterances than the location-sensitive attention mechanism adopted in Tacotron2 [14], is adopted. The predicted Mel-spectrograms from the decoder are then fed to the Griffin-Lim reconstruction algorithm [68] to synthesize the waveform.

Following the speaker embedding-based multi-speaker TTS model [65], the speaker embedding in the training phase is engaged after the CBHG block. It works as a speaker attribute to provide speaker information to the TTS system. During the inference phase, the speaker attribute is provided by the face embedding, so that we can synthesize speech guided by the face image. The standard training method of Tacotron2 [14] is adopted to train the face-conditioned multi-speaker TTS.

D. Talking Head Generation

The talking head generation module generates the talking head video given the Mel-spectrograms synthesized by the TTS module. This process consists of two steps: 1) Mel-spectrograms-to-facial landmark sequence generation, and 2) landmark sequence-to-video generation (see Section III-E).

The facial landmark sequence generation module follows the basic idea of MakeItTalk [11], i.e., separately predicting the lip movement and head movement, and combining them to generate the final facial landmarks. As also shown in Fig. 2, the landmark sequence generation module consists of three encoders: an audio encoder, a landmark encoder, and a quaternion encoder, which are used for the encoding of the synthesized Mel-spectrograms, the facial landmarks of the input face image, and the orientation of the face in the input image, respectively. The output from these three encoders is concatenated and subsequently used as input to the decoder to generate the facial landmark sequence. By connecting consecutive key points of facial landmarks in each frame with pre-defined colors, i.e., using different colors to distinguish different parts like that in [11], we obtain a sequence of facial sketches. These facial sketches are then concatenated with the input face image, resulting in a sequence of 6-channel images used to generate photo-realistic frames in the final video with an image-to-image translation way.

A vivid talking head should not only have lip movements synchronized to the speech but also natural head pose movements. While lip movements and facial expressions, e.g., the jaw and eye movements, of course, occur in 3D, the final facial landmarks are drawn on a 2D plane, i.e., a facial sketch from which the photo-realistic facial image is rendered. Therefore, movements in the direction that is perpendicular to the face are not important for the landmark-to-image generation. In contrast, rotations of the head (referred to as head pose) lead to different head sketches on the 2D plane. To effectively model facial expressions and head poses, we decompose the landmark prediction into landmark shifts in a 2D plane and head rotations in a 3D space with the help of a 3D facial landmark detector [43] that can detect 3D coordinates of landmarks from images (video frames).

Given an input face image I, the extracted facial landmarks consist of 68 key points, each of which is represented by three-dimensional coordinate values. To capture the facial expression-related movements, such as the lip and jaw movements in the same plane, we first map the facial landmarks to a front-facing standard facial template as done in [11] with the ICP method proposed in [69]. The orientation of the original face is represented as a set of quaternion numbers $q \in R^4$. The landmark of the input image is important conditional information for the prediction of landmarks. As shown in Fig. 2, Mel-spectrograms, the frontal facial landmarks, and quaternions are encoded by the audio encoder, landmark encoder, and quaternion encoder, respectively. Outputs from these three encoders are concatenated to work as input to the landmark decoder that generates the new landmarks.

We denote the sequence of Mel-spectrograms as $S = \{s_1, s_2, \ldots, s_T\}$, where T is the sequence length. The goal is to generate a sequence of frontal landmarks $\hat{P} = \{\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_T\}$ and a sequence of quaternions $\hat{Q} = \{\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_T\}$. The corresponding ground-truth landmark sequence and quaternion sequence are $P = \{p_1, p_2, \ldots, p_T\}$ and $Q = \{q_1, q_2, \ldots, q_T\}$, respectively. In practice, we drop the dimension of the landmarks in the depth direction (z-axis value), and only 2D landmarks are used to present the landmark displacements, which means each point is represented by two-dimensional coordinates, i.e., *<x-axis value, y-axis value>*. Considering the 68 key points of the facial landmarks, one landmark frame can be represented as $p_t \in R^{136}$ by concatenating the coordinate values of the x-axis and the y-axis.

Different persons have different face shapes. This makes it challenging to predict landmarks for a new person that was never seen during the training process. [11] tackled this problem by predicting the displacements of landmarks instead of directly predicting the landmarks. They then added these displacements to the base landmarks' coordinates. Here, we follow the same approach. To this end, we have to choose a frame to provide the base landmarks and quaternions, which are referred to as base landmarks and base quaternions hereafter. During training, the training samples pair is a sequence of Mel-spectrograms and a sequence of landmarks extracted from the video, so that we can randomly choose one landmark frame to provide the base facial landmarks and quaternions. During the inference process, only one input image is available, and the base facial landmarks and quaternions are provided by this input image. The prediction of the landmarks' displacements can be formulated as:

$$
\begin{aligned}
s_i' &= AE\left(s_i; \mathrm{w}_{AE}\right) \\
p' &= LE\left(p_{init}; \mathrm{w}_{LE}\right) \\
q' &= QE\left(q_{init}; \mathrm{w}_{QE}\right) \\
f_i &= \mathrm{concat}\left(s_i', p,' q'\right), F = \{f_1, f_2, \ldots, f_t\} \\
\Delta P, \Delta Q' &= LD\left(F; \mathrm{w}_{LD}\right) \\
\Delta Q &= LSTM\left(\Delta Q'; \mathrm{w}_{LSTM}\right), \qquad (1)
\end{aligned}
$$

where AE, LE, QE, and LD are the audio encoder, landmark encoder, quaternion encoder, and landmark decoder,

respectively. w_{AE}, w_{LE}, w_{QE}, w_{LD}, and w_{LSTM} are learnable parameters. F is a sequence of features, each of which is obtained by concatenating the frame-level speech embedding vector s'_i, base landmark embedding vector p', and base quaternion embedding vector q'. The output of the landmark decoder is a sequence of concatenated landmark displacements and preliminary quaternion changes in the frame level. Specifically, each frame of the decoder's output is a 140 dimensional vector that consists of 136 dimensions for landmark displacements and the other 4 dimensions for quaternion changes. After segmenting each frame, we can get a sequence of landmark displacements ΔP, and a sequence of quaternion changes $\Delta Q'$. Compared to lip movements, the head pose changes more slowly and smoothly. To make the predicted head pose changes be smooth, a further LSTM is adopted to deal with the predicted quaternion changes $\Delta Q'$, which results in the final quaternion changes ΔQ. The predicted frame-level frontal landmarks and quaternions can be obtained via:

$$\begin{aligned} \bar{p}_i &= p_{init} + \Delta p_i \\ \bar{q}_i &= q_{init} + \Delta q_i. \end{aligned} \tag{2}$$

With the predicted quaternions, we get the rotation matrix M, with which we obtain the final rotated landmarks:

$$\hat{p}_i = \bar{p}_i \cdot M \tag{3}$$

Model architecture: All the designed encoders, i.e., audio encoder, landmark encoder, and quaternion encoder, are multilayer perceptrons (MLP) with two linear transformation layers, where the first linear transformation layer is followed by a layer normalization [70] and an activation function of LeakyReLu [71]. The hidden unit sizes of AE, LE, and QE are 512, 256, and 64, respectively. The vector dimensions of s'_i, p', and q' are 512, 128, and 4 respectively. Therefore, the dimension of f_i is 644.

The landmark decoder LD consists of a 1D-CNN block, a bidirectional LSTM block, and an MLP. The 1D-CNN consists of six 1-D convolutional layers with unit sizes of 512, 512, 1024, 1024, 1024, and 2048, respectively. Instance normalization is used after the first convolutional layer, while the other convolutional layers are followed by batch normalization. The MLP follows the same structure as those encoders, with a hidden unit size of 512.

Objective function: The objective functions for the displacement prediction consist of an L2 regression loss and a pairwise inter-frame loss. Specifically, the L2 regression loss is defined as:

$$\mathcal{L}_d = \sum_{t=1}^{T} \sum_{i=1}^{N} \|p_{i,t} - \hat{p}_{i,t}\|_2^2 \tag{4}$$

where N is the batch size. The pairwise inter-frame loss is defined as:

$$\mathcal{L}_{in} = \sum_{t=2}^{T} \sum_{i=1}^{N} \|(p_{i,t} - p_{i,t-1}) - (\hat{p}_{i,t} - \hat{p}_{i,t-1})\|_2^2. \tag{5}$$

The objective function for the quaternion prediction is a L1 regression loss:

$$\mathcal{L}_q = \sum_{t=2}^{T} \sum_{i=1}^{N} \|q_{i,t} - \hat{q}_{i,t}\|. \tag{6}$$

The total loss function of the landmark prediction is

$$\mathcal{L}_L = \mathcal{L}_d + \mathcal{L}_{din} + \mathcal{L}_q. \tag{7}$$

E. Landmark to Photo-Realistic Image

In the generated landmark sequence, each frame consists of facial landmarks with a special head pose and lip shape. With the facial landmarks of each frame, we can generate the photo-realistic face image by the face generator in Fig. 2. Here we take the UNet architecture from [11], [72], [73] as the face generator to perform this landmark-to-image translation. The landmarks of each frame are drawn as a portrait sketch on a 2D plane by connecting the key points with pre-defined colorful lines, as shown in Fig. 2. Then this portrait sketch is concatenated with the input image, resulting in a 6-channel image with a resolution of 256×256 which will work as the input to the face generator. The output is a photo-realistic face image with the same facial key points as input landmarks.

To train the image generator, in addition to minimizing the L1 pixel-level distance and perceptual feature distance between the reconstructed face and the training target face as in [11], conditional generative adversarial training loss in [74] is also used. Following [74], the discriminator is a patch-based fully convolutional network. The input of the discriminator is also the channel-wise concatenation of the portrait sketch and the input image (real) or the generated image (fake).

IV. Experiments and Results

A. Database

Table I lists the various databases that were used to train the different modules of the proposed method. In addition to these databases, we also collected data to evaluate the face-conditioned multi-speaker TTS and the final generated talking head video. These collected data will be introduced in Section IV-B. The databases listed in Table I will be introduced below grouped by the corresponding module.

1) Database for the TTS: In order to be able to make both Mandarin and English speaking talking head videos, databases of both languages, i.e., AISHELL-3[3] and VCTK[4] were adopted to train the multi-speaker TTS model. AISHELL-3 is a multi-speaker Mandarin speech database with speech by 218 native Chinese Mandarin speakers with a total of 88,035 utterances. VCTK is a multi-speaker English speech database with speech from 110 English speakers with various accents, where each speaker reads out around 400 sentences. The multi-speaker model is trained with these two databases together, which allows the trained model to be used for both Chinese and English.

[3]http://www.aishelltech.com/aishell_3
[4]https://datashare.ed.ac. UK/handle/10283/3443

TABLE I
DATABASES THAT WERE USED TO TRAIN THE DIFFERENT MODULES

Database	Adopted Modality	Language	Speaker number	Used for module
AISHELL-3	Text-Audio	Mandarin	218	TTS
VCTK	Text-Audio	English	110	TTS
Aidatatang-200zh	Audio	Mandarin	600	Speaker embdding extractor
VoxCeleb2 subset [58]	Audio-Video	English	433	Face encoder; Image translation model
Cn-Celeb subset [75]	Audio-Image	Mandarin	313	Face encoder
Obama Weekly Address [44]	Audio-Video	English	1	Landmark prediction model

Note that these two databases are only used for the training of the multi-speaker TTS model. In the final speech synthesis for the talking head video, the speaker identity is provided by the face image. However, no paired face image exists in AISHELL-3 and VCTK. Therefore, only 100 transcriptions are randomly selected as the test sentences for the whole talking head generation task, and their paired utterances are not used.

2) Database for Face Encoder: The speaker embedding extractor which works as the teacher to train the face encoder is trained with the database of Aidatatang-200zh.[5] This is a Chinese Mandarin speech corpus that contains 200 hours of speech data from 600 speakers. After obtaining the pre-trained speech embedding extractor, databases that paired speech and faces are needed to train the face encoder. We use two subsets from VoxCeleb2[6] and Cn-Celeb[7] to train the face encoder.

Both VoxCeleb2 and CN-Celeb were originally designed for the task of speaker verification. VoxCeleb2 is an audio-visual database, which consists of short clips of human speech extracted from interview videos uploaded to YouTube. The associated video track provides us with the matched face images to the corresponding utterances. Here, a subset of VoxCeleb2 [76] is adopted. This subset consists of 16,128 English utterances uttered by 433 speakers. Following [76], 422 speakers with 15,729 utterances are used as training data and the other 11 speakers are used as the test set to provide the speaker image in the talking head generation task during evaluation. For each speaker, we randomly extracted 50 frames from their talking videos to build a paired face database.

The original Cn-Celeb contains more than 130,000 utterances from 1,000 Chinese celebrities, but without face information. To obtain the speech-image pairs, we collected a face image database of a part of the speaker identities in the Cn-Celeb. Specifically, this collected face database consists of 313 speakers and each speaker has 40 to 100 face images downloaded from Baidu Image.[8] Therefore, the final database to train the face encoder consists of 735 speakers and 28,450 utterances.

3) Database for the Talking Head Generation: Following [11], the Obama Weekly Address database [77], which contains around 6 hours of Obama's speeches, is used to train the landmark prediction model. We cut the audio signals into fixed-length utterances with the duration of 3 s. Subsequently, the total set of utterances is split into three subsets: 90%, 5%, and 5% of the total set are used for training, validation, and test, respectively.

The database to train the image translation model is the subset of VoxCeleb2 introduced in Section IV-A2. Different from the data pairs used in Section IV-A2, which are speech-image pairs, here, speech-video pairs are used.

4) Data Processing: In all proposed modules, speech is represented as Mel-spectrograms with the same parameters. Specifically, the Mel-spectrograms are computed through a short-time Fourier transform (STFT) with 50 ms frame size and 12.5 ms frame hop, resulting in a frame frequency of 80 Hz. The frame rate of the videos from the Obama Weekly Address database is 25 fps. To align the Mel-spectrograms and video frames, we up-sample the video frame rate to 80 fps. This up-sampling is performed on the landmark features instead of on the raw video frames.

B. Evaluation

Our goal is to generate talking head videos with text and a face image as input. A well-generated result should consist of: 1) speech that is likely produced by the person in the given face image, and 2) a video that is synchronized with synthetic speech. Therefore we have to evaluate the synthetic speech and the generated video respectively.

1) Face-Conditioned Multi-Speaker TTS: The goal of the face-conditioned multi-speaker TTS is to synthesize speech that sounds like it could be produced by the given face image. This makes the evaluation a subjective task. To that end, we carried out two A/B test experiments in which participants were asked to indicate which of two synthesized speech samples (A or B) was more likely to have been produced by the person in a given image. In the first A/B test experiment, we investigated whether the face image-based synthetic speech achieved comparable results to speech-based synthetic speech which was created from the real speech of this person to provide the speaker information to our system (also referred to as the "reference speech-based synthetic speech"). In the second A/B test experiment, we compared the face image-based synthetic speech to synthesized speech that is conditioned on reference speech that is randomly selected from the training set with the same gender.

Both VoxCeleb2 and Cn-Celeb databases that were used to train the face encoder contain speech and images from celebrities, which might be familiar to the participants and subsequently influence the participants' responses. Therefore, we collected 16 (8 men and 8 women) videos recorded by 16 unknown native Chinese-speaking YouTubers from YouTube,[9] which were used as an evaluation set. The input text was taken from the test set of AISHELL-3.

[5]http://www.openslr.org/62/
[6]http://www.openslr.org/49/
[7]http://www.openslr.org/82/
[8]https://image.baidu.com/
[9]https://www.youtube.com/

Each A/B test experiment consisted of 16 trials, where each trial consisted of a face image shown on the screen, a speech sample synthesized by our face image-based multi-speaker TTS, and a compared speech sample synthesized by the reference speech-based multi-speaker TTS (the reference is ground-truth speech of the person in this given image or randomly selected from training set but with the same gender). Both synthetic speech samples were created from the same text input and were presented auditorily. A total of 27 participants (8 females and 19 males; age range 18-40; Chinese native speaker) participated in both of the A/B test experiments. In the experiments, they were asked to choose the one speech sample that they thought was more likely to have been produced by the identity in the given face image or indicate that "They are similar" when they can not tell which one is better. Each participant attended the experiment individually to make sure the independence of the evaluation.

2) Talking Head Generation: The two stages of the speech-to-video generation module, i.e., speech-to-landmark generation and landmark-to-video generation, are evaluated separately. For the speech-to-landmark generation module, the ground-truth landmark sequences are available from the test set of Obama Weekly Address database. Objective evaluations are performed to compare the generated landmark sequences and the ground-truth landmark sequences. The final generated videos are evaluated using a human perceptual rating experiment, which evaluates the naturalness of the video and the synchronization between the lip movements and the speech.

Evaluation Metrics: Following [11], we evaluate the speech-to-landmark generation, and particularly the accuracy of the lip movements, using the landmark distance for lips (**D-LL**), landmark velocity difference for lips (**D-VL**), and difference in the open mouth area (**D-A**) as evaluation metrics. D-LL is the average Euclidean distance between the predicted lip landmarks and the ground-truth ones. D-VL is the average Euclidean distance between the predicted lip landmark velocities and that of the ground-truth ones. D-A represents the average difference between the area of the predicted mouth shape and the ground-truth one.

Human rating experiment: Twenty-two participants (5 females and 17 males; age range: 18-40; first language: Mandarin; second language: English) were asked to rate 20 generated videos in terms of 1) the synchronization of the lip movements and speech, and 2) the overall realness of the video, respectively, on a 5-point scale using a slider. A score of 1 means "very bad" and a score of "5" means excellent. To create the 20 videos, twenty face images were randomly collected using Google Image to generate the talking head videos. Ten of the face images were used to create Chinese talking head videos and the 10 other face images were used to create the English talking head videos. For both languages, two sentences were randomly selected from AISHELL-3 or VCTK, which were used as the input sentences. Note that, the cross-lingual task is not considered in this paper. During the inference process, the language is manually defined based on whether the person in the image is Chinese(-looking) or not.

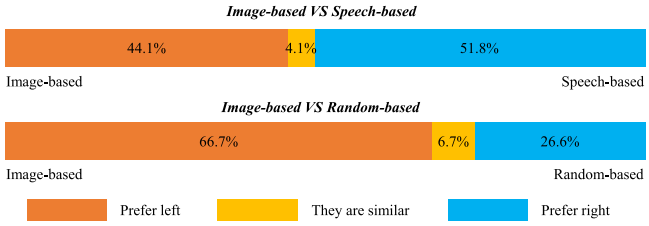

Fig. 5. The results of the human study on the evaluation of the face-conditioned multi-speaker TTS.

C. Results

This section presents the evaluation results for the 1) face-conditioned multi-speaker TTS, and the 2) talking head generation.

1) Face-Conditioned Multi-Speaker TTS: The human A/B experiment (see Section IV-B1) results are shown in Fig. 5, which displays the percentage of the total votes for the two speech samples averaged over all listeners and all trials in the two A/B test experiments. The upper bar shows the percentage of the votes for the face image-based method (orange) and that for the reference speech-based method (blue). Note that the reference speech-based method can be treated as an upper boundary in the multi-speaker TTS task because the speech embedding vectors are extracted from the ground-truth speech signals. While a 7.7% difference exists between the face image-based method and the reference speech-based method, a one-way ANOVA shows that no significant difference exists between the image-based results and speech-based results ($F = 0.17$, $p = 0.68$). The lower bar shows the percentage of the votes for the face image-based method (orange) and that for the reference speech-based method with random reference speech with the same gender (blue). The results clearly show that the participants more often chose the speech generated using the proposed face image-based method as being produced by the face image than the speech that was produced with random selected reference speech. A one-way ANOVA confirmed these results: there is a significant difference between the face image-based method and the randomly selected speech-based results ($F = 7.25$, $p = 01$). Taking the result of the two A/B experiments together, we can conclude that our face image-based method produces speech in a voice that is in accordance with the face in the image. Fig. 6 shows a few examples of the synthesized results of our face-image based method. The left side of the figure shows the input images, while the right side shows the corresponding synthesized spectrograms, and the fundamental frequency of the synthesized speech (the red line). As can be seen, using the same sentence (top of the right side of the figure) but different face images as input, the generated spectrograms and also fundamental frequencies are significantly different, indicating that the face image indeed can provide discriminative identity information.

2) Talking Head Generation: Lip movement prediction. We first evaluate how well the predicted lip landmarks are synchronized with the ground-truth lip landmarks and compare the performance of our lip movement prediction to that of the

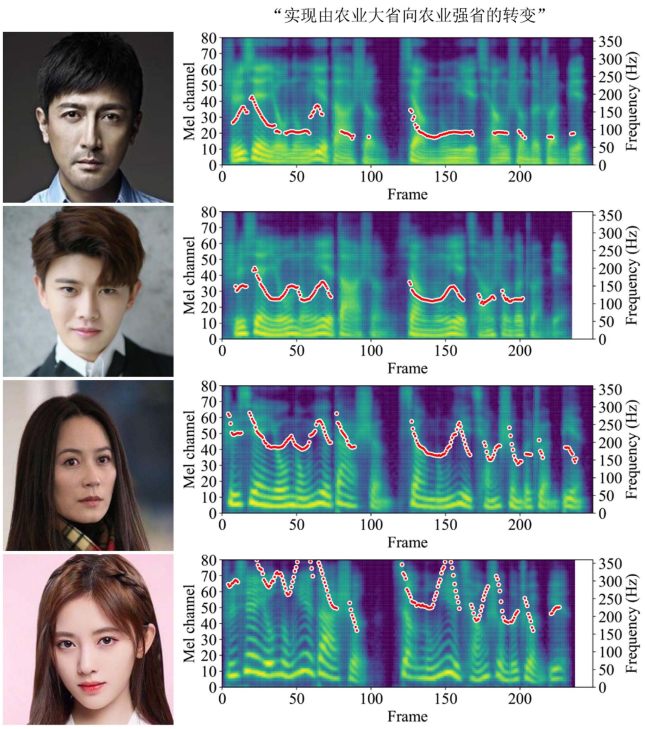

Fig. 6. Examples of the synthesized speech (right panels) after conditioning on the face image (left panels).

TABLE II
QUANTITATIVE EVALUATION OF THE LIP LANDMARK PREDICTIONS. FOR ALL EVALUATION METRICS, A LOWER VALUE MEANS BETTER PERFORMANCE. BOLD INDICATES THE BEST RESULT OF EACH METRIC

Method	D-LL↓	D-VL↓	D-A↓
MakeItTalk	0.143	0.036	0.143
TDLSTM	0.105	0.027	0.130
BLSTM	0.101	0.027	0.115
Our approach	**0.095**	**0.026**	**0.105**

state-of-the-art landmark-based talking head method for arbitrary persons MakeItTalk [11]. MakeItTalk is based on the same 3D landmark extractor as our model is, and its landmark prediction model is also trained on the Obama weekly talking database, just as our model. This allows for a fair comparison between our method and that of MakeItTalk. Moreover, to evaluate the proposed CNN-LSTM-based landmark decoder, our approach is compared to two variants of the proposed methods: the TDLSTM approach and the BLSTM approach, in which the landmark decoder in Fig. 2 is replaced by a time-delay LSTM and bi-directional LSTM, respectively, both of which are popular architectures in related work [11], [44], [46]. The results are shown in Table II. Bold indicates the best result. As can be seen, our method outperforms MakeItTalk in terms of all evaluation metrics. The proposed method also outperforms the TDLSTM-based and BLSTM-based methods, indicating the superiority of the proposed CNN-BLSTM landmark decoder over the TDLSTM and the BLSTM decoders.

TABLE III
HUMAN RATINGS FOR THE VIDEO EVOLUTION. LARGER IS BETTER, AND THE MAXIMUM VALUE IS 5. BOLD INDICATES THE BEST RESULT. ALL $P < .001$ IN A ONE-WAY ANOVA

Method MOS	Lip Sync Quality	Overall Realness
ATVGnet	2.92	2.60
MakeItTalk	2.55	2.68
Ours	**3.16**	**3.17**

Video generation: Several frames of the final generated photo-realistic videos are presented in Fig. 7. We compare our method and MakeItTalk with another talking head generation method: ATVGnet [12] which is also designed for arbitrary persons. Compared to MakeItTalk and our proposed method, ATVGnet crops the face region of the input image and no head pose is considered. While the amplitude of the lip movements is larger for the ATVGnet generated talking faces than that for those generated by MakeItTalk and our method, many of them are unnatural, such as those lip regions circled by blue circles. Compared to the input image (left-most column), obvious distortions appear in the results generated by MakeItTalk. For instance, in the first case, the generated results of MakeItTalk (the second row in Fig. 7) show a thinner facial shape than the original facial shape. Besides, there is a loss of the facial details in these generated frames, which reduces the sharpness of the generated faces. In contrast, the facial details are preserved well in the frames generated by our method, and no distortion appears in our generated frames.

Table III shows the results from the human rating experiment, where 5 is the highest rating and bold indicates the best results. As shown in this Table, the human raters gave higher rates to the lip sync quality and the overall realness of our method compared to those of ATVGnet and MakeItTalk. A one-way ANOVA with the method as a factor (three levels of the factor are included, i.e., ATVGnet, MakeitTalk, and our method) and rating score as dependent variable showed that significant differences exist between the ratings for the different methods, indicating that our method had significantly better lip sync quality and overall realness than ATVGnet and MakeItTalk.

V. DISCUSSION

In this paper, a talking head generation method is proposed which generates the speech in a voice consistent with the input face image taking only text and a face image as input. This extends previous work which either was able to generate talking head videos for only specific persons or is designed for arbitrary persons but depends on existing speech.

Due to the lack of paired face image and high-quality speech database, the face embedding, which provides the identity information, is obtained by a pre-trained face encoder. It would be more intuitive to optimize the face encoder with the TTS system. To this end, a high-quality multi-speaker TTS corpus paired with face images should be collected. Alternatively, on the basis of the system provided by this work, the face encoder could be

Fig. 7. Comparison of generated talking head video frames by different methods.

further fine-tuned in the TTS system with existing paired image and speech databases.

In MakeItTalk [11], the talking head movements are treated as shifts of key points. In contrast, we treat the head movements as rotations instead of shifts. In the proposed method, quaternions are used to represent the rotations of the talking head. The human rating results show that our method can generate more natural talking head videos than MakeItTalk. However, similar to MakeItTalk, our method also suffers from the pose limitation that the predicted poses are small. The main reason is that there is no explicit correlation between speech and head pose, and the latter is more random and is, therefore, harder to predict. Therefore, using generative adversarial learning strategies can be considered in the future to predict random head movements that look natural.

While the recently proposed method [55] cannot be applied to an arbitrary person, this end-to-end method shows obvious superiority in generating more accurate lip movements compared to landmark-based methods. It is caused by the landmarks' low dimensionality which suppresses details, which subsequently can lead to mismatches between the landmarks and the photo-realistic face image. Fig. 8 shows some examples of generated images conditioned on landmarks where errors occurred. In this figure, the landmarks are extracted from the real target image (second column), which means that these landmarks are ground-truth landmarks. However, even with these ground-truth landmarks, there are still differences between the generated images (right-most column) and the real target images, because

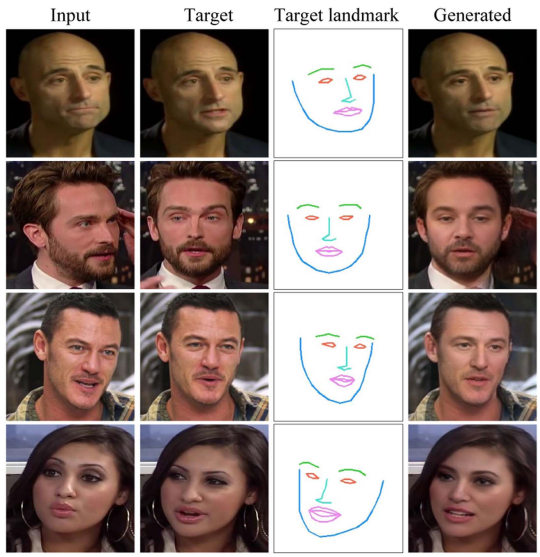

Fig. 8. Generated images based on ground-truth landmarks.

different lip shapes from the photo-realistic images could lead to the same sketch on a 2D plane due to reduced dimensions. However, the landmarks show the superiority in presenting the head pose, making the proposed method can predict the head pose automatically. For future research, using head pose information provided by the landmarks can be considered in the end-to-end way to predict the head pose instead of using the head poses from a reference video as in [55].

Ethical consideration: While the proposed method can synthesize speech based on the input face image for any person from that input face image, we do not argue that there is an inevitable relation between a face and a voice. The proposed face-conditioned multi-speaker TTS module is not created to reconstruct someone's real voice but rather to give a face in a photo a voice that sounds as if the person in the photo could have produced speech with that voice. The proposed method could be used in many scenarios, e.g., film making, video editing, and human-computer interaction. However, such forward-looking technology could also have the potential to be misused or abused for various malevolent purposes, such as spreading false statements or misinformation. To prevent our released code from being abused, a watermark is included in this code to make the generated videos. We also encourage the public to report any suspicious videos to the appropriate authorities.

VI. Conclusion

This paper presented a method, which we called AnyoneNet, for the first time, can synthesize a talking head video with synchronized speech for an arbitrary person with only text and a face image as input in a voice that is consistent with the input face image. The voice of the talking head is created on the basis of the face image. The proposed method consists of two main modules, i.e., a face-conditioned multi-speaker TTS module and a speech-driven talking head video generation module. The results of several experiments showed that the proposed face-conditioned multi-speaker TTS can synthesize voices in harmony with the face in the given face image, and the proposed speech-driven talking head video generation method has the state-of-the-art performance on the task of landmark-based talking head generation.

Acknowledgment

The authors thank Dong Wang and colleagues who built the CN-Celeb database for providing us the real speaker identities in this database, so that we can obtain the face-speech pairs for CN-Celeb.

References

[1] L. Chen, G. Cui, Z. Kou, H. Zheng, and C. Xu, "What comprises a good talking-head video generation?: A survey and benchmark," 2020, *arXiv:2005.03201*.

[2] R. Kumar, J. Sotelo, K. Kumar, A. de Brébisson, and Y. Bengio, "ObamaNet: Photo-realistic lip-sync from text," 2017, *arXiv:1801.01442*.

[3] P. Garrido et al., "Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track," in *Computer Graphics Forum*, vol. 34, no. 2. Hoboken, NJ, USA: Wiley, 2015, pp. 193–204.

[4] J. Charles, D. Magee, and D. Hogg, "Virtual immortality: Reanimating characters from TV shows," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 879–886.

[5] B. Fan, L. Xie, S. Yang, L. Wang, and F. K. Soong, "A deep bidirectional LSTM approach for video-realistic talking head," *Multimedia Tools Appl.*, vol. 75, no. 9, pp. 5287–5309, 2016.

[6] Y. Hati, F. Rousseaux, and C. Duhart, "Text-driven mouth animation for human computer interaction with personal assistant," in *Proc. Int. Conf. Auditory Display Dept. Comput. Inf. Sci.*, 2019, pp. 75–82.

[7] W. Chae and Y. Kim, "Text-driven speech animation with emotion control," *KSII Trans. Internet Inf. Syst.*, vol. 14, no. 8, pp. 3473–3487, 2020.

[8] T.-H. Oh et al., "Speech2Face: Learning the face behind a voice," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7539–7548.

[9] H.-S. Choi, C. Park, and K. Lee, "From inference to generation: End-to-end fully self-supervised generation of human face from speech," in *Proc. Int. Conf. Learn. Representations*, 2019.

[10] Y. Sun, H. Zhou, Z. Liu, and H. Koike, "Speech2talking-face: Inferring and driving a face with synchronized audio-visual representation," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, 2021, pp. 1018–1024.

[11] Y. Zhou, et al., "Makelttalk: Speaker-aware talking-head animation," *ACM Trans. Graph.*, vol. 39, no. 6, pp. 1–15, 2020.

[12] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7832–7841.

[13] S. Zhang, J. Yuan, M. Liao, and L. Zhang, "Text2video: Text-driven talking-head video synthesis with personalized phoneme-pose dictionary," in *Proc. ICASSP 2022-2022 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 2659–2663.

[14] J. Shen et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 4779–4783.

[15] Y. Ren et al., "FastSpeech: Fast, robust and controllable text to speech," in *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[16] C. Yu et al., "Durian: Duration informed attention network for speech synthesis," in *Proc. InterSpeech*, 2020, pp. 2027–2031.

[17] J. Sotelo et al., "Char2wav: End-to-end speech synthesis," in *Proc. Int. Conf. Learn. Representations*, 2017.

[18] A. van den Oord et al., "WaveNet: A generative model for raw audio," in *Proc. 9th ISCA Speech Synth. Workshop*, 2016, pp. 125.

[19] A. v. d. Oord et al., "WaveNet: A generative model for raw audio," in *Proc. 9th ISCA Speech Synth. Workshop*, 2016, pp. 125.

[20] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 3617–3621.

[21] K. Kumar et al., "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[22] J. Kong, J. Kim, and J. Bae, "Hifi-Gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 17022–17033, 2020.

[23] G. Yang et al., "Multi-band melgan: Faster waveform generation for high-quality text-to-speech," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2021, pp. 492–498.

[24] W. Ping, K. Peng, and J. Chen, "ClariNet: Parallel wave generation in end-to-end text-to-speech," in *Proc. Int. Conf. Learn. Representations*, 2018.

[25] J. Donahue, S. Dieleman, M. Bińkowski, E. Elsen, and K. Simonyan, "End-to-end adversarial text-to-speech," in *Proc. Int. Conf. Learn. Representations*, 2021.

[26] R. J. Weiss, R. Skerry-Ryan, E. Battenberg, S. Mariooryad, and D. P. Kingma, "Wave-tacotron: Spectrogram-free end-to-end text-to-speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 5679–5683.

[27] K. Ito and L. Johnson, "The LJ speech dataset," 2017. [Online]. Available: https://keithito.com/LJ-Speech-Dataset/

[28] J. Yamagishi et al., "Robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 6, pp. 1208–1230, Aug. 2009.

[29] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 4475–4479.

[30] S. Yang, Z. Wu, and L. Xie, "On the training of DNN-based average voice model for speech synthesis," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2016, pp. 1–6.

[31] Y. Jia et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[32] E. Nachmani, A. Polyak, Y. Taigman, and L. Wolf, "Fitting new speakers based on a short untranscribed sample," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 3683–3691.

[33] E. Cooper et al., "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6184–6188.

[34] E. Casanova et al., "Sc-Glowtts: An efficient zero-shot multi-speaker text-to-speech model," 2021, *arXiv:2104.05557*.

[35] O. Schreer, R. Englert, P. Eisert, and R. Tanger, "Real-time vision and speech driven avatars for multimedia applications," *IEEE Trans. Multimedia*, vol. 10, pp. 352–360, 2008.

[36] P. Edwards, C. Landreth, E. Fiume, and K. Singh, "Jali: An animator-centric viseme model for expressive lip synchronization," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–11, 2016.

[37] L. Yu, J. Yu, and Q. Ling, "BLTRCNN-based 3-D articulatory movement prediction: Learning articulatory synchronicity from both text and audio inputs," *IEEE Trans. Multimedia*, vol. 21, no. 7, pp. 1621–1632, Jul. 2019.

[38] X. Wen, M. Wang, C. Richardt, Z.-Y. Chen, and S.-M. Hu, "Photorealistic audio-driven video portraits," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 12, pp. 3457–3466, Dec. 2020.

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

[40] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner, "Neural voice puppetry: Audio-driven facial reenactment," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 716–731.

[41] A. Richard et al., "Audio-and gaze-driven facial animation of codec avatars," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 41–50.

[42] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, 2009.

[43] 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 *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1021–1030.

[44] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: Learning lip sync from audio," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–13, 2017.

[45] F. Fang, X. Wang, J. Yamagishi, and I. Echizen, "Audiovisual speaker conversion: Jointly and simultaneously transforming facial expression and acoustic characteristics," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6795–6799.

[46] L. Yu, J. Yu, M. Li, and Q. Ling, "Multimodal inputs driven talking face generation with spatial-temporal dependency," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 203–216, Jan. 2021.

[47] X. Ji et al., "Audio-driven emotional video portraits," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14080–14089.

[48] L. Chen et al., "Talking-head generation with rhythmic head motion," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 35–51.

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

[50] L. Chen, Z. Li, R. K. Maddox, Z. Duan, and C. Xu, "Lip movements generation at a glance," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 520–535.

[51] K. Vougioukas, S. Petridis, and M. Pantic, "End-to-end speech-driven facial animation with temporal GANs," *CVPR Workshops*, pp. 37–40, 2019.

[52] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, "Talking face generation by adversarially disentangled audio-visual representation," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, pp. 9299–9306.

[53] Y. Song, J. Zhu, D. Li, A. Wang, and H. Qi, "Talking face generation by conditional recurrent adversarial network," in *Proc. 28th Int. Joint Conf. Artif. Intell., IJCAI-19. Int. Joint Conf. Artif. Intell. Org.*, 2019, pp. 919–925. [Online]. Available: https://doi.org/10.24963/ijcai.2019/129

[54] 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 *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 484–492.

[55] H. Zhou, et al., "Pose-controllable talking face generation by implicitly modularized audio-visual representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4176–4186.

[56] L. Chen, S. Srivastava, Z. Duan, and C. Xu, "Deep cross-modal audio-visual generation," in *Proc. Thematic Workshops ACM Multimedia*, 2017, pp. 349–357.

[57] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Comput. Speech Lang.*, vol. 60, 2020, Art. no. 101027.

[58] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[59] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proc. Asian Conf. Comput. Vis.*, Springer, 2016, pp. 87–103.

[60] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3444–3453.

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[62] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.

[63] Y. Liu, L. He, and J. Liu, "Large margin softmax loss for speaker verification," in *Proc. Interspeech*, 2019, pp. 2873–2877.

[64] G. Ilharco, Y. Zhang, and J. Baldridge, "Large-scale representation learning from visually grounded untranscribed speech," in *Proc. 23rd Conf. Comput. Natural Lang. Learn.*, 2019, pp. 55–65.

[65] Z. Cai, C. Zhang, and M. Li, "From speaker verification to multispeaker speech synthesis, deep transfer with feedback constraint," in *Proc. Interspeech*, 2020, pp. 3974–3978.

[66] Y. Wang et al., "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017, pp. 4006–4010.

[67] E. Battenberg et al., "Location-relative attention mechanisms for robust long-form speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6194–6198.

[68] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, Apr. 1984.

[69] A. Segal, D. Haehnel, and S. Thrun, "Generalized-ICP," in *Robotics: science and systems*, vol. 2, no. 4. Seattle, WA, 2009, p. 435.

[70] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[71] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier Nonlinearities Improve Neural Network Acoustic Models," in *Proc. Int. Conf. Mach. Learn.*, 2013, vol. 30, no. 1, pp. 3–8.

[72] P. Esser, E. Sutter, and B. Ommer, "A variational U-Net for conditional appearance and shape generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8857–8866.

[73] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9459–9468.

[74] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1125–1134.

[75] L. Li et al., "Cn-celeb: Multi-genre speaker recognition," *Speech Commun.*, vol. 137, pp. 77–91, 2022.

[76] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," *Adv. Neural Inf. Process. Syst.*, vol. 32, pp. 7137–7147, 2019.

[77] N. S. Nourabadi, "Synthesizing naturalistic and meaningful speech-driven behaviors," Ph.D. dissertation, Dept. Elect. Eng., The Univ. Texas at Dallas, USA, 2017.