Chapter 8

Talking Faces: Audio-to-Video Face Generation

Yuxin Wang, Linsen Song, Wayne Wu, Chen Qian, Ran He,
and Chen Change Loy

Abstract Talking face generation aims at synthesizing coherent and realistic face sequences given an input speech. The task enjoys a wide spectrum of downstream applications, such as teleconferencing, movie dubbing, and virtual assistant. The emergence of deep learning and cross-modality research has led to many interesting works that address talking face generation. Despite great research efforts in talking face generation, the problem remains challenging due to the need for fine-grained control of face components and the generalization to arbitrary sentences. In this chapter, we first discuss the definition and underlying challenges of the problem. Then, we present an overview of recent progress in talking face generation. In addition, we introduce some widely used datasets and performance metrics. Finally, we discuss open questions, potential future directions, and ethical considerations in this task.

Y. Wang · W. Wu · C. Qian
SenseTime Research, Beijing, China
e-mail: wangyuxin1@sensetime.com

W. Wu
e-mail: wuwenyan@sensetime.com

C. Qian
e-mail: qianchen@sensetime.com

L. Song · R. He
NLPR, CASIA, Beijing, China
e-mail: songlinsen2018@ia.ac.cn

R. He
e-mail: rhe@nlpr.ia.ac.cn

C. C. Loy (✉)
S-Lab, Nanyang Technological University, Singapore, Singapore
e-mail: ccloy@ntu.edu.sg

8.1 Introduction

Talking face generation aims at synthesizing a realistic target face, which talks in correspondence to the given audio sequences. Thanks to the emergence of deep learning methods for content generation [1–3], talking face generation has attracted significant research interests from both computer vision [4–8] and computer graphics [9–14].

Talking face generation has been studied since 1990s [15–18], and it was mainly used in cartoon animation [15] or visual-speech perception experiments [16]. With the advancement of computer technology and the popularization of network services, new application scenarios emerged. First, this technology can help multimedia content production, such as making video games [19] and dubbing movies [20] or TV shows [21]. Moreover, the animated characters enhance human perception by involving visual information, such as video conferencing [22], virtual announcer [23], virtual teacher [24], and virtual assistant [12]. Furthermore, this technology has the potential to realize the digital twin of real person [21].

Talking face generation is a complicated cross-modal task, which requires the modeling of complex and dynamic relationships between audio and face. Existing methods typically decompose the task into subproblems, including audio representation, face modeling, audio-to-face animation, and post-processing. As the source of talking face generation, voice contains rich content and emotional information. To extract essential information that is useful for talking face animation, one would require robust methods to analyze and comprehend the underlying speech signal [7, 12, 22, 25–28]. As the target of talking face generation, face modeling and analysis are also important. Models that characterize human faces have been proposed and applied to various tasks [17, 22, 23, 29–33]. As the bridge that joins audio and face, audio-to-face animation is the key component in talking face generation. Sophisticated methods are needed to accurately and consistently match a speaker’s mouth movements and facial expressions to the source audio. Last but not least, to obtain a natural and temporally smooth face in the generated video, careful post-processing is inevitable.

Toward conversational human-computer interaction, talking face generation requires techniques that could generate realistic digital talking faces that make human observers feeling comfortable. As highlighted in Uncanny Valley Theory [34], if an entity is anthropomorphic but imperfect, its non-human characteristics will become the conspicuous part that creates strangely familiar feelings of eeriness and revulsion in observers. The requirement poses stringent requirements on the talking face models, demanding realistic fine-grained facial control, continuous high-quality generation, and generalization ability for arbitrary sentence and identity. In addition, this also prompts researchers to build diverse talking face datasets and establish fair and standard evaluation metrics.

8.2 Related Work

In this session, we discuss relevant techniques employed to address the four sub-problems in talking face generation, namely, audio representation, face modeling, audio-to-face animation, and post-processing.

8.2.1 Audio Representation

It is generally believed the high-level content information and emotional information in the voice are important to generate realistic talking faces. While original speech signal can be directly used as the input of the synthesis model [25], most methods prefer more representative audio features [7, 12, 22, 26–28]. A pre-defined analysis method or a pre-trained model is often used to extract audio features from the original speech, and then the obtained features are used as the input to the face generation system. Four typical audio features are illustrated in Fig. 8.1.

Mel-spectrum features are commonly used in speech-related multimodal tasks, such as speech recognition. Considering that human auditory perception is only concentrated on specific frequencies, methods can be designed to selectively filter the audio frequency spectrum signal to obtain Mel-spectrum features. Mel-frequency

Fig. 8.1 Illustration of four commonly used audio features. **a** Original speech signal, **b** spectrum feature, **c** phoneme (English International Phonetic Alphabet (IPA)), and **d** Mel-frequency cepstrum coefficients (MFCC)

cepstrum coefficients (MFCC) can be obtained by performing Cepstrum analysis on the Mel-spectrum features. Prajwal et al. [26] used Mel-spectrum features as audio representation to generate talking face, while Song et al. [7] used MFCC.

Noise in the original audio signal could corrupt the MFCC. Some methods thus prefer to extract text features that are related to the content of the speech. These methods often borrow models from specific speech signal processing tasks such as automatic speech recognition (ASR) or voice conversion (VC). The automatic speech recognition (ASR) task aims at converting speech signals into corresponding text. For example, DeepSpeech [35, 36] is a speech-to-text model, which can transform an input speech frequency spectrum to English strings. Das et al. [27] took DeepSpeech features as the input to the talking face generation model. From the perspective of acoustic attributes, a phoneme is the smallest speech unit. It is considered that each phoneme is bounded to a specific vocalization action. For example, there are 48 phonemes in the English International Phonetic Alphabet (IPA), corresponding to 48 different vocal patterns. Quite a few methods [12, 28] use phoneme representation to synthesize talking face. The voice conversion (VC) task aims at converting non-verbal features such as accent, timbre, and speaking style between speakers while retaining the content features of the voice. Zhou et al. [22] used a pre-trained VC model to extract text features to characterize the content information in the speech.

8.2.2 Face Modeling

The human's perception of the quality of talking videos is mainly constituted by their visual quality, lip-sync accuracy, and naturalness. To generate high-quality talking face videos, the synchronization of 2D/3D facial representations with the input audios play an important role. Many geometry representations of human faces have been explored in recent years, including 2D/3D face modeling.

2D Models. 2D facial representations like 2D landmarks [17, 22, 29–31], action units (AUs) [32], and reference face images [23, 31, 33] are commonly used in talking face generation. Facial landmark detection is defined as the task of localizing and representing salient regions of the face. As shown in Fig. 8.2, facial landmarks are usually composed of points around eyebrows, eyes, nose, mouth, and jawline. As a shape representation of the face, facial landmark is a fundamental component in many face analysis and synthesis tasks, such as face detection [37], face verification [38], face morphing [39], facial attribute inference [40], face generation [41], and face reenactment [29]. Chen et al. [37] showed that aligned face shapes provide better features for face classification. The proposed joint learning of face detection and alignment greatly enhances the capability of real-time face detection. Chen et al. [38] densely sampled multi-scale descriptors centered at dense facial landmarks and used the concatenated high-dimensional feature for efficient face verification. Seibold et al. [39] presented an automatic morphing pipeline to generate morphing attacks, by warping images according to the corresponding detected facial landmarks

Fig. 8.2 Illustration of 106 facial landmarks. The landmarks are detected and marked in green. Best viewed zoomed in. The original pictures are obtained from the Internet

and replacing the inner part of the original image. Di et al. [41] presented that the information preserved by landmarks (gender in particular) can be further accentuated by leveraging generative models to synthesize corresponding faces. Lewenberg et al. [40] proposed an approach that incorporates facial landmark information for input images as an additional channel, helping a convolutional neural network (CNN) to learn face-specific features for predicting various traits of facial images. Automatic face reenactment [42] learns to transfer facial expressions from the source actor to the target actor. Wayne et al. proposed ReenactGAN [29] to reenact faces by the facial boundaries constituted by facial landmarks. Conditioned on the facial boundaries, the reenacted face images become more robust to challenging poses, expressions, and illuminations.

Action units are the fundamental actions of facial muscles defined in the Facial Action Coding System (FACS) system [32]. The combination of AUs can characterize comprehensive facial expression features, which can be used in expression-related face analysis and synthesis, e.g., facial expressions recognition [32], and facial animation [43]. For example, Pumarola et al. [43] introduced a generative adversarial network (GAN) [1] conditioning on action units annotations to realize controllable facial animation with robust expressions and lighting conditions.

3D Models. Some exiting methods exploit the 3D geometry of human faces like 3D landmarks [44, 45], 3D point cloud [46], facial mesh [47], facial rigs [13], and facial blendshapes [48–50] to generate talking face videos with diverse head gestures and movements.

Before the emergence of deep convolution networks (DCN) and GAN [1] in face image generation, 3D morphable face model (3DMM) is commonly deployed as a general face representation and a popular tool to model human faces. In 1999, Blanz and Vetter [44] proposed the first 3DMM that shows impressive performance. In

Fig. 8.3 Illustration of sampled faces of Basel face model (BFM) proposed by Paysan et al. [45]. The mean together with the first three principle components of the shape (left) and texture (right) PCA model. The figure shows the mean shape and texture, along with their components with plus/minus five standard deviations σ . A mask with four manually chosen segments (eyes, nose, mouth, and rest) is used in the fitting to extend the flexibility. The image is adopted from Paysan et al. [45]

2009, the first publicly available 3DMM model, also known as Basel face model (BFM), is released by Paysan et al. [45]. These face models inspire the research of 3DMM and its applications on many computer vision tasks related to human faces. For instance, Cao et al. [51] proposed the FaceWarehouse model and Bolkart and Wuhler [52] proposed the Multilinear face model. Both models capture the geometry of facial shapes and expressions. Cao et al. [51] released a RGBD dataset of 150 subjects, each with 20 expressions. Bolkart et al. [52] released a dataset of 100 subjects, each with 25 expressions. Sampled faces of Basel Face Model (BFM) are shown in Fig. 8.3.

These methods model the facial shapes and expressions in a linear space, neglecting the nonlinear transformation of facial expressions. Li et al. [48] proposed the FLAME model that enables the nonlinear control on 3D face model by incorporating the linear blendshapes with eyes, jaw, and neck joints. To tackle the challenges like large head poses, appearance variations, inference speed, and video stability in 3D face reconstruction, Guo et al. proposed 3DFFA [49] and its improved variant, 3DFFA_V2 [50]. Apart from the 3D Face Morphable Model, other 3D models like face rigs [13], 3D point cloud [46], facial mesh [47], and customized computer graphic face model are also applied in 3D face representation.

With the advances of 3D face models, a variety of applications are enabled, such as face recognition [53], face reenactment [42], face reconstruction [54], face rotation [55], visual dubbing [56], and talking face generation [57]. Blanz et al. [53] showed that the cosine distance between two face images' shape and color coefficients estimated by a 3D face model can be used for identification. Thies et al. [58] proposed the first real-time face reenactment system by transferring the expression coefficients of a source actor to a target actor while preserving person-specificity. Gecer et al. [54] employed a large-scale face model [59] and proposed a GAN-based method for high-fidelity 3D face reconstruction. Zhou et al. [55] developed a face

rotation algorithm by projecting and refining the rotated 3D face reconstructed from the input 2D face image by 3DDFA [49]. Kim et al. [56] presented a visual dubbing method that enables a new actor to imitate the facial expressions, eye movements, and head movements of one actor only from its portrait video. Thies et al. [57] presented that the learned facial expression coefficients from speech audio features extracted by DeepSpeech [36] can animate a 3D face model uttering the given speech audio. In general, these 3D models are not publicly released due to copyright restrictions.

8.2.3 Audio-to-Face Animation

To synthesize realistic and natural talking faces, it is crucial to establish the correspondence between the audio signal and the synthesized face. To improve the visual quality, lip-sync accuracy, and naturalness of talking videos, different methods have been explored in recent years, including 2D/3D-based models and video frame selection algorithm.

Audio-Visual Synchronization. Quite a few methods construct the correspondence between phonemes and visemes and use search algorithms to map audios to mouth shapes during the testing phase [12, 17, 28]. The pipeline is illustrated in Fig. 8.4. Specifically, they divide speech into pre-defined minimum audio units (phonemes), which naturally correspond to the smallest visual vocalization methods (visemes). In this way, a repository of phoneme-viseme pairs can be established from training data. After that, each sentence can be decomposed into a sequence of phonemes, correspond to a sequence of visemes during the testing phase. The video will be further synthesized from visemes by generation or rendering. The visemes here can be the facial landmarks related to the vocalization [17], or the pre-defined 3D face model controller coefficients [12, 28]. In this framework, defining phoneme-viseme pairs and finding a search-stitching algorithm are two critical steps. Considering coarticulation, Bregler et al. [17] split each word into a sequence of triphones, and established a corresponding relationship with the eigenpoint position of the lips and chin. Yao et al. [28] established the relationship between the phonemes obtained by the p2fa [60] algorithm and the controller coefficients obtained by the parameterized human head model [61]. They proposed a new phoneme search algorithm to quickly find the best phoneme subsequence combination and stitch the corresponding expression coefficients to synthesize the speaking video.

Other researchers designed an encoder-decoder structure, taking in audio and speaker images, outputting the generated target faces [5, 25, 62]. Specifically, as shown in Fig. 8.5a, the designed model is a combination of two encoders taking in audio and face images as input for two different modalities, and a decoder generating an image synchronized with the audio while preserving the identity information of the input images. In this system, two encoders are, respectively, responsible for encoding the audio content information and the facial identity information. The decoder following is responsible for decoding the fused multi-modality features into

Fig. 8.4 Pipeline of phoneme-viseme correspondence method for talking face generation. Phoneme is firstly mapped to viseme according to an established phoneme-viseme correspondence. Images are further synthesized based on visemes

Fig. 8.5 Pipelines of two methods for talking face generation. As shown in (a), an encoder-decoder structure is used to generate the target face by taking in the audio features and images. As shown in (b), the relationship between the audio features and specific intermediate features is established first, and then the corresponding face is generated based on the intermediate features

a face image with the corresponding mouth shape and face identity. The encoders and the decoder are usually trained end-to-end simultaneously. This kind of method makes full use of encoder-decoder structure and multimodal fusion to generate target images. In this way, researchers often design specific models and losses to realize the disentanglement of speaking content and speaker identity. For example, Zhou et al. [5] used a pre-trained word classifier to force the content information to be forgotten in the identity encoding process, and the content information obtained from images and audio were constrained as close as possible.

Other methods choose to first establish the relationship between audio features and intermediate features pre-defined by face modeling methods and then generate the corresponding faces from the intermediate features [22, 31], as shown in Fig. 8.5b. The intermediate features mentioned here can be the pre-defined facial landmarks or the expression coefficients of the 3D face model.

For 2D-based generation methods, facial landmarks are often used as sparse shape representation. Suwajanakorn et al. [31] used a recurrent neural network (RNN) to map the MFCC features to the PCA coefficients of the facial landmarks. The corresponding face image is thus generated from the reconstructed facial landmarks with the texture information provided by the face images. Zhou et al. [22] mapped the voice content code and the identity code to the offset of the facial landmarks relative to a face template, and then generated the target image through an image-to-image network. For 3D-based methods, the facial expression parameters are often used as the intermediate representation. Fried et al. [12] used the facial expression parameters of the human head model as intermediate features and designed a neural renderer to generate the target video. Wiles et al. [63] established a mapping from audio features to the latent code in a pre-trained face generation model to achieve audio-driven facial video synthesis. Guo et al. [64] used a conditional implicit function to generate a dynamic neural radiance field from the audio features, and then synthesized video using volume rendering. The main difference between these methods (as shown in Fig. 8.5b) and the aforementioned phoneme-viseme search methods (as shown in Fig. 8.4) is the use of regression models for replacing pre-constructed phoneme-viseme correspondence. The former can obtain more consistent correspondence in the feature space.

Some researchers designed specific models to ensure audio-visual synchronization. Chung et al. [65] proposed a network, as shown in Fig. 8.6, taking in audio features and face images sequence as input, outputting the lip-sync error. This structure is often used in talking face model training [26, 66] or evaluation [25, 27]. A specific model was designed by Agarwal et al. [67] to detect the mismatch between phoneme and viseme to determine whether the video has been modified.

Synthesis Based on 2D Models. At the early stage of 2D-based talking face generation, videos are generated based on a pre-defined face model or composition of background portrait video and mouth images. Lewis [15] associated the recognized phonemes from synthesized speeches with mouth positions to animate a face model. Bregler et al. [17] designed the first automatic facial animation system that automatically labels phonemes in the training data and morphs these mouth gestures with the background portrait video. Cosatto and Graf [23] described a system to animate lip-synced head model from the phonetic transcripts by retrieving images of facial parts and blend them onto a whole face image.

With the popularity of the multidimensional morphable model (MMM), Ezzat et al. [68] designed a visual speech model to synthesize a speaker's mouth trajectory in MMM space from the given utterance and an algorithm to re-composite the synthesized mouths onto the portrait video with natural head and eye movement.

Fig. 8.6 Illustration of the pipeline of SyncNet [65]. The network predicts whether the input audio and face images are synchronized

Chang and Ezzat [69] animated a novel speaker with only a small video corpus (15 s) by transferring an MMM trained on a different speaker with a large video corpus (10–15 min).

Inspired by the successful application of the hidden Markov model (HMM) in speech recognition, many HMM-based methods, such as R-HMM [18], LMS-HMM [70], and HMMI [71], were proposed since talking face generation can be seen as an audio-visual mapping problem. Different from these HMM-based methods that use a single-state chain, a coupled hidden Markov model (CHMM) approach [19] was used to model the subtle characteristics of audio and video modalities. To exploit the capability of HMM in modeling the mapping from the audio to visual modalities, Wang et al. [72] proposed a system to generate talking face videos guided by the visual parameter trajectory of lip movements produced from the trained HMM according to the given speech audio.

Due to the advancement of using RNN and long short term memory (LSTM), HMM is gradually replaced by LSTM in learning the mapping from the audio to the visual modality. For instance, Fan et al. [24] trained a deep bidirectional LSTM to learn the regression model by minimizing the error of predicting visual sequence from audio/text sequence, outperformed their previous HMM-based models. Suwajanakorn et al. [31] trained a time-delayed LSTM model to learn the mapping from the Mel-frequency cepstral coefficients (MFCC) features of an audio sequence to the mouth landmarks of a single frame.

The quality of human face synthesis improves dramatically with the recent advances of GAN-based image generator, such as DCGAN [2], PGGAN [73], CGAN [3], StyleGAN [74] and StyleGAN2 [75]. In 2014, Goodfellow et al. pro-

posed GAN [1] and demonstrated its ability in image generation by generating low-resolution images after training on datasets like MNIST [76], TFD [77] and CIFAR-10 [78]. Then, DCNs with different architectures are developed in GAN to generate images of higher resolution for specific domains. For instance, DCGAN [2] applied the layered deep neural network and PGGAN [73] learned to generate images in a coarse-to-fine manner by gradually increasing the resolution of generated images. In the context of image generation of human faces, a conditional CycleGAN [79] and FCENet [80] were developed to generate face images with controllable attributes like hair and eyes. While facial attributes can be precisely controlled by input condition codes, the image resolution is not high (128×128) and many facial details are missing. To generate high-resolution face images, Karras et al. proposed StyleGAN [74] and StyleGAN2 [75] to generate face images with a resolution up to 1024×1024 pixels, where coarse-grained style (e.g., eyes, hair, lighting) and fine-grained style (e.g., stubble, freckles, skin pores) are editable. To edit facial attributes more precisely, some GAN-based models [79, 80] were proposed to modify the generated high-resolution face images where fine-grained attributes like eyes, nose size, and mouth shape can be controlled by input condition codes. The design of 2D-based talking face video synthesis models is inspired by some related synthesis tasks like image-to-image translation [81, 82], high-resolution face image generation [74], face reenactment [29], and lip reading [65].

Inspired by GAN [1], many methods [4–8] improve the generated video quality from different aspects. Chen et al. designed a correction loss [4] to synchronize changes of lip and speech. Zhou et al. [5] proposed an adversarial network to disentangle the speaker identity from input videos and the word identity from input speeches to enable arbitrary-speaker talking face generation. To improve both the image and video realism, Chen et al. [6] designed a dynamic adjustable pixel-wise loss to eliminate the temporal discontinuities and subtle artifacts in generated videos. Song et al. [7] proposed a conditional recurrent generation network and a pair of spatial-temporal discriminators that integrate audio and image features for video generation. These GAN-based studies mainly concentrate on the talking face video generation of the frontal face and neutral expressions. The development of GAN-based human face generation and editing methods on head poses [83] and facial emotions [84] influences the research in talking face generation. For instance, Zhu et al. [8] employed the idea of mutual information to capture the audio-visual coherence and design a GAN-based framework to generate talking face videos that are robust to pose variations. Taking into account the speaker's emotions and head poses, Wang et al. [85] released an audio-visual dataset that contains various head poses, emotion categories, and intensities. They also proposed a baseline to demonstrate the feasibility of controlling emotion categories and intensities in talking face generation.

Synthesis Based on 3D Models. In the early days of the talking face generation, 3D representation is often used to represent the mouth or face of the driven speaker. For instance, in 1996, a 3D model of lips with only five parameters was developed to adapt lip contours of various speakers and any speech gesture [16]. Wang et al. [88]

proposed to control a 3D face model with the head trajectory and articulation movement predicted by an HMM model. Interestingly, after several years' exploration of applying deep learning in talking face generation, especially the recent advances of DCN and GAN, many methods return to 3D representation by integrating 3DMM and other 3D face models. For instance, Pham et al. [9] introduced a 3D blendshape model animated by 3D rotation and expression coefficients predicted only from the input speech. Karras et al. [10] presented a network that animates the 3D vertex coordinates of a 3D face model with different emotions from the input speech and emotional codes. Taylor et al. [11] developed a real-time system that animates active appearance model (AAM), CG characters, and face rigs by retargeting the face rig movements predicted from the given speech. Fried et al. [12] proposed a parametric head model to provide the position of retargeting the mouth images to the background portrait video. Edwards et al. [13] presented a face-rig model called JALI that mainly concentrates on the JAW and LIp movements. By making use of JALI, Zhou et al. [14] proposed a deep learning method to drive the JALI or standard FACS-based face rigs by the JALI and viseme parameters predicted from a 3-stage LSTM network. Recently, a series of methods explore the potential of deep learning techniques in learning the nonlinear mapping from audio features to facial movement coefficients of 3DMM. For instance, Thies et al. [57] introduced a small convolutional network to learn the expression coefficients of 3DMM from the speech features extracted by the DeepSpeech [35]. This method does not pay much attention to large head poses, head movements and requires a speaker-specific video renderer. Song et al. [33] presented an LSTM-based network to eliminate speaker information and predict expression coefficients from input audios. This method is robust to large pose variations and the head movement problem is tackled by the designed frame selection algorithm. Different from this method that retrieves head poses from existing videos, Yi et al. [89] tried to solve the head pose problem by directly predicting the pose coefficients of 3DMM from the given speech audio. Chen et al. [90] introduced a head motion learner to predict the head motion from a short portrait video and the input audio. To eliminate the visual discontinuity brought by the apparent head motion, a 3D face model is used due to its stability. In Fig. 8.7, representative works of talking face generation in recent years are listed in chronological order.

Video Frame Selection Algorithm. Note that the mouth texture in the training videos is abundant, video frame selection algorithms are designed to facilitate the synthesis of talking face videos by selecting frames from existing videos according to the input audios or mouth motion representations. The selected video frames can provide the texture of the whole face [31, 33] or only mouth areas [23].

Currently, generation based on 2D face representation (e.g., DCN and GAN) and 3D face representation (e.g., 3DMM) dominates the field of talking face synthesis. Before the emergence of these techniques, talking face generation mainly rely on 3D models and select video frames with matched mouth shapes. For instance, Cosatto et al. [23] introduced a flexible 3D head model used to composite facial parts' images retrieved by sampled mouth trajectories. Chang et al. [69] proposed a matching-by-synthesis algorithm that selects new multidimensional morphable model (MMM)

Fig. 8.7 Representative works of talking face generation in recent years (since 2017). The methods above the timeline are based on 2D models, from left to right are [4–6, 25, 26, 31, 62, 86]. The methods below the timeline are based on 3D models, from left to right are [10–12, 14, 22, 33, 57, 87]. The generated images of these methods are adopted from corresponding papers. Best viewed by zooming on the screen

prototype images from driving speaker’s videos. Wang et al. [72] introduced an HMM trajectory-guided approach as a guide to select an optimal mouth sequence from the training videos. Liu and Ostermann [91] presented a unit selection algorithm to retrieve mouth images from a speaker’s expressive database characterized by phoneme, viseme, and size.

The research on frame selection algorithms is still active even with the impressive talking face generation performance brought by deep learning and 3DMM techniques. For example, Fried et al. [28] introduced a dynamic programming method to retrieve expressions in the parameter space by visemes inferred from the input transcript. Suwajanakorn et al. [31] designed a dynamic programming algorithm to retrieve background video frames according to the input audio. How well the input audio volume matches the eye blink as well as head movement is considered in the frame selection algorithm.

8.2.4 Post-processing

The generated talking faces may not be of high quality or natural enough due to various reasons. This requires researchers to do introduce post-processing steps, such as refinement and blending, to further enhance the naturalness of the videos. For instance, Jamaludin et al. [62] first obtained a talking face generation model that produced blurred faces and then trained a separate CNN to sharpen the blurred images. Bregler et al. [17] pointed out the necessity to blend the generated faces into a natural background, and the importance of animate the chin and jawlines, not

just the mouth region to improve realism. There exist many methods that apply a static video background [4, 5, 8, 68]. For some news program translation or movie dubbing applications [21, 26], the natural video results can be obtained by blending the generated face back into the original background.

8.3 Datasets and Metrics

8.3.1 Dataset

In recent years, increasingly more audio-visual datasets have been released, promoting the development of talking face generation. These datasets can be used for lip reading, speech reconstruction, and talking face generation. We divide these datasets into two categories according to the collection environment. (1) Indoor environment, where speakers recite the specified words or sentences. (2) In-the-wild environment, where speakers talk in the scene closer to the actual applications, such as speech video and news program video. In this section, we summarize commonly used audio-visual datasets and their characteristics.

Indoor Environment. Datasets collected in the indoor environment often exhibit consistent settings and lighting conditions, when the speakers read the specified words or sentences.

GRID [92] is a multi-speaker audio-visual corpus consisting of audio and video recordings of 1000 sentences spoken by each of 34 speakers. TCD-TIMIT [93] consists of audio and video footages of 62 speakers reading a total of 6913 phonetically rich sentences. Three of the speakers are professionally-trained lip speakers, with the assumption that trained speakers can read better than ordinary speakers. Video footage was recorded from the frontal view and 30° pitch angle. CREMA-D [94] is a dataset of 7442 original clips from 91 actors. The speakers are composed of 48 men and 43 women from different races and nationalities, ranging in age from 20 to 74 years old. They speak 12 sentences using one of six different emotions and four different emotion levels. However, all the datasets mentioned above do not consider emotional information. Wang et al. [85] released a high-quality audio-visual dataset that contains 60 actors and actresses talking with eight different emotions at three different intensity levels. All clips in MEAD are captured at seven different view angles in a strictly controlled environment.

In-the-Wild Environment. Other datasets are often derived from news program videos or speech videos. They are closer to actual application scenarios, with more abundant words, more natural expressions, and more speakers.

Suwajanakorn et al. [31] downloaded 14 h of Obama weekly address videos from YouTube for experiments. LRW [95], LRS2 [96], LRS3 [97] datasets are all designed for research on lip reading. Lip reading is defined as understanding speech content by visually interpreting the movements of the lips, face, and tongue when normal sound

Table 8.1 Summary of audio-visual datasets commonly used for talking face generation

Dataset	Environment	Identity	Hours	Year
GRID [92]	Indoor	34	27.5	2006
CREMA-D [94]	Indoor	91	11.1	2014
TCD-TIMIT [93]	Indoor	62	11.1	2015
MEAD [85]	Indoor	60	40	2020
Obama weekly address [31]	Wild	1	14	2017
LRW [95]	Wild	Hundreds of	173	2016
LRS2 [96]	Wild	Hundreds of	224.5	2018
LRS3 [97]	Wild	Thousands of	438	2018
VoxCeleb1 [98]	Wild	1251	352	2017
VoxCeleb2 [99]	Wild	6112	2400	2018

is not available. This is similar to the inverse task of talking face generation. LRW consists of about 1000 utterances of 500 words, spoken by hundreds of speakers. All videos are about 1.16 s in length, and the duration of each word is also given. LRS2 expands the content of the speech from words to sentences, consisting of thousands of spoken sentences from BBC television, where each sentence is up to 100 characters in length. LRS3 contains thousands of spoken sentences from TED and TEDx speech videos.

VoxCeleb1 [98] collects celebrity videos uploaded by users from YouTube, which contains over 100,000 utterances for 1251 celebrities. VoxCeleb2 [99] further expands the data volume, which contains over 1 million utterances for 6112 celebrities. VoxCeleb2 can be used as a supplement for VoxCeleb1 because it has no overlap with the identities in the VoxCeleb1. Datasets mentioned in this section are summarized in Table 8.1.

8.3.2 Metrics

It is challenging to evaluate the naturalness of generated talking faces. People often have very strict requirements on the quality and naturalness of the generated talking face. A slight flaw will be regarded as obviously unreal. On the one hand, this puts high demands on the models of talking face generation. On the other hand, it is crucial to develop comprehensive evaluation metrics for talking face generation. Evaluation metrics can be divided into objective quantitative evaluation and subjective qualitative evaluation.

Quantitative Evaluation. As mentioned above, people can easily find out when the generated talking faces do not speak like real people from various aspects. Thus, the quantitative evaluation also needs to measure from several different angles. In gen-

eral, existing quantitative evaluation metrics mainly focus on the following aspects of the generated video. (1) The generated videos should be of high quality. (2) The mouth shape of the generated speaker should match the audio. (3) The speaker in the synthesized video should be the same as the target person. (4) Eye blinking when speaking should be natural.

Image quality evaluation metrics are commonly used in face generation tasks. Peak signal-to-noise ratio (PSNR), defined via mean squared error, can reflect the pixel-level difference between two images. However, there is still a considerable gap between human perception and PSNR. Structural Similarity (SSIM) [100] measures the similarity of two images in terms of illuminance, contrast, and structure. To evaluate the diversity of the generative model, Inception Score (IS) [101] is introduced. Fréchet inception distance (FID) [102] is calculated by comparing the mean and standard deviation of the two features produced by a pre-trained Inception-v3 model. However, these methods require reference images for evaluation. Cumulative probability blur detection (CPBD) [103] is a non-reference image evaluation metric used to evaluate the sharpness of images, while frequency domain blurriness measure (FDBM) [104] evaluates frequency domain blurriness based on the image spectrum.

Audio-lip synchronization is also an important indicator to measure the naturalness of talking face generation. Landmark distance (LMD) is defined as the mouth landmark distance between generated and real reference images to measure the generated mouth shape. As mentioned in Sect. 8.2.3, the lip reading task learns the mapping from face images to the corresponding text. Thus, the pre-trained lip reading model can be used to calculate the word error rate (WER). For example, Vougioukas et al. [105] calculated WER based on a LipNet [106] model pre-trained on GRID [92]. Syncnet [65], the model specifically designed to judge audio-visual synchronization, can also be borrowed [25, 27] to calculate Audio-Visual synchronization metrics (AV Offset and AV confidence). A lower AV offset with higher AV confidence indicates better lip synchronization. Recently, Chen et al. [107] proposed a new lip-synchronization evaluation metric lip-reading similarity distance (LRSD) from the perspective of human perception. Based on a newly proposed lip reading model, they use the distance between features of generated video clips and ground truth video clips to measure the audio-visual synchronization.

Some methods suffer from wrong or lost of speaker identity, that is, the generated speaker and the target speaker do not seem to be the same person. Therefore, some metrics that measure identity preservation are also applied in the talking face generation task. Often, a pre-trained face recognition model [108, 109] is used as an identity feature extractor. Identity preservation is quantified by measuring the distance between features. For instance, average content distance (ACD) [25, 27] is calculated by measuring the similarity between FaceNet [108] features of the reference identity image and the predicted image. Chen et al. [90] used cosine similarity (CSIM) between embedding vectors of ArcFace [109] for measuring identity mismatch.

Finally, the realisticness of blinking should also be considered. Vougioukas et al. [25] proposed that the average blink duration and blink frequency from the generated video should be similar to that of natural human blinks. In specific, they

Table 8.2 Summary of quantitative talking face metrics via four different degrees. The upward arrows ($\uparrow$) indicate higher values are better for that metric, while downward arrows ($\downarrow$) mean lower values are better

Degree	Metrics
Image quality	PSNR $\uparrow$ SSIM $\uparrow$ IS $\uparrow$ FID $\downarrow$ CPBD $\uparrow$ FDBM $\uparrow$
Audio-lip synchronization	LMD $\downarrow$ WER $\downarrow$ AV (Offset) $\downarrow$ AV (Confidence) $\uparrow$ LRSD $\downarrow$
Identity-preserved	ACD $\downarrow$ CSIM $\uparrow$
Blink	Duration Frequency

calculated the average duration and frequency to evaluate the naturalness of blinking. Quantitative evaluation metrics mentioned in this section are summarized in Table 8.2.

Qualitative Evaluation. Although the quantitative evaluation mentioned above can provide a reference and filter out some obvious artifacts, the ultimate goal of talking face is to fool real people. Therefore, the generated talking face still needs some subjective feedback from people. Generally speaking, researchers usually design user studies to allow real users to judge the quality of the generated videos.

8.4 Discussion

8.4.1 *Fine-Grained Facial Control*

Even if the speaker’s mouth movements naturally match the audio, one wishes to establish the relationship between audio and other facial components, such as chins, jawlines, eyes, head movements, and even teeth.

In fact, most of the current talking face generation methods do not consider the correlation between audio and eyes. Vougioukas et al. [25] designed a blink generation network, using Gaussian noise vectors as input to generate eyes keypoints, which can generate blinks of similar duration and frequency to real videos. Zhang et al. [110] took eye blink signal and audio signal together as input to generate the corresponding talking face. Zhou et al. [22] learned mapping from the audio information to facial landmarks where eye landmarks are excluded. These methods are based on the assumption that blinking is a random signal unrelated to the input audio. However, according to Karson et al. [111], listeners’ blink duration is related to talking and thinking. Hömke et al. [112] also proposed that blinks are meaningfully rather than randomly paced, although no visual information is processed. When it comes to generation techniques, the movements of the eyes are generally modeled as part of the emotional coefficients in 3D-based methods and as eye landmarks in 2D-based methods. Shu et al. [113] leveraged user’s photo collections to find a set of reference eyes and transfer them onto a target image. However, for now, it is still difficult to

model the relationship between audio and eye movements. In other words, how to generate more flexible and informative eyes is still an open question in talking face generation task.

Another question is whether the teeth generation is related to the input audio. From the perspective of phoneme-viseme correspondence, each phoneme corresponds to a set of teeth and tongue movements. However, as described in [31], the teeth are sometimes hidden behind lips when speaking, which makes synthesis challenging. There are also no teeth landmarks in 2D landmark definition. Even in most 3D head models, the teeth are not explicitly modeled. Some researchers copy teeth texture from other frames [58] or use teeth proxy [20, 114]. However, these methods may cause blur or artifacts. Suwajanakorn et al. [31] achieved decent teeth generation by combining low-frequency median texture and high-frequency details from a teeth proxy image. Recently, some more accurate teeth models have been established, for example, Velinov et al. [115] established an intra-oral scan system for capturing the optical properties of live human teeth. Some new teeth editing methods have also been proposed. For example, Yang et al. [116] realized an effective disentanglement of an explicit representation of the teeth geometry from the in-mouth appearance, making it easier to edit teeth.

The lips, eyes, and teeth mentioned above are all part of the human face. One would also need to consider the generation of natural head movements. Most talking face methods do not consider the problem of generating controllable head movements without a pre-defined 3D model. Jamaludin et al. [62] only generated aligned faces while Zhang et al. [110] took the head pose signal as the input signal explicitly. Wiles et al. [63] can generate talking faces with different poses, but the head motion is not decoupled from other facial expression attributes. Recently, some researchers have proposed methods to generate controllable head poses. Chen et al. [90] designed a head motion disentangler to decouple the head movement in the 3D geometry space and used the head motion and audio information of the current frame to predict the head motion of the next frame. Similarly, Wang et al. [117] realized the decoupling of motion-related information and identity-specific information by learning 3D key-point representation. Zhou et al. [86] modularized audio-visual representations by devising an implicit low-dimension pose code to generate pose-controllable talking face videos.

For a realistic talking face, the emotion of the speaker should also match the voice. For example, a voice with an angry tone should correspond to an angry face. But how to manipulate the emotion in 2D-based talking face generation is still an open question. Some researchers exploit expression information from the voice to generate talking faces [25, 118]. But they cannot explicitly control the emotional intensity of the video. MEAD [85] is a talking face dataset featuring 60 people talking with eight different emotions at three different intensity levels, which provides data support for the generation of emotional talking faces. Ji et al. [87] decomposed speech into emotion space and content space. With the disentangled features, emotional facial landmarks and videos can be deduced.

In Sect. 8.3.2, we mentioned several evaluation metrics for talking face generation, but these quantitative indicators still have limitations from the perspective of human perception. We believe that talking face integrating eyes, teeth, head pose, and emotion will be a more natural and human-like virtual person.

8.4.2 Generalization

The model generalization of a talking face system is mainly determined by the dataset used to build the system and the applied techniques in designing the modules of the system. The audio-visual datasets are contributed by two essential factors, the phonetic dictionary size of the corpus and the diversity of speakers such as gender, age, language, accent, and the speaker number. In the following, the model generalization of recent talking face generation methods is analyzed by key factors, e.g., corpus and speaker.

The small corpus size and speaker number of many audio-visual datasets might limit the model generalization. For example, the GRID dataset [92] contains very few words. Although it is designed to cover the pronunciations of every single phoneme, the limited vocabulary still lacks diverse diphones and triphones that encode surrounding phonemes. Many audio-visual datasets contain very limited speaker diversity, i.e., the speaker number of GRID [92] and RAVDESS [119] is fewer than 100 and these datasets do not contain diverse accents, head pose, movements, and emotions. To alleviate the poor model generalization brought by audio-visual datasets, Wang et al. [85] collected a large-scale dataset with different skin colors, emotions and head poses.

With the development of GAN-based image generation methods, recent methods can generate photo-realistic talking face videos with fewer and fewer portrait videos. For instance, generating a high-fidelity fake video of Barack Obama requires massive training footage up to 14h in [31]. Though the generated videos of [31] are hard to distinguish from the real ones, the requirement on training data is inapplicable in many real-world application scenarios. Thus, many methods circumvent the training data burden at the cost of generated video quality. For example, Thies et al. [57] presented that transferring a trained model to an unseen speaker requires about only 2 min of footage. Zhou et al. [22] presented that even a single static face image is sufficient for generating talking videos with diverse head movements.

Another aspect of model generalization is the speaker's identity. Suwajanakorn et al. [31] built a speaker-specific 3D face model and trained a speaker-specific network for Barack Obama to synthesize his forged videos. The applied speaker-specific 3D face model limits its generalization for other speakers. Then, Thies et al. [57] proposed an audio-to-video pipeline that consists of a speaker-generalized network to learn the mapping from the audio to expression parameters and a speaker-specific video renderer to render photo-realistic video according to the 3D head model and

learned expressions. To an unseen speaker, it still requires a 2 min portrait video to fine-tune the speaker-specific renderer. The renderer parameters only optimize for a specific speaker since it refines the speaker-specific texture rendered by 3DMM. Recent methods [22, 33, 117] can generate talking video for unseen speakers without any further finetuning and the testing set can even be as small as a short footage [33] or a single image [22, 117]. Such model generalization is realized since these methods do not optimize based on any speaker-specific prior knowledge.

8.5 Conclusion

With the advancement of face modeling methods and deep learning techniques, especially generation models, academic researchers make it possible to generate realistic talking faces. In turn, considering a wide range of practical applications, talking face generation has also attracted increasing interest from industrial developers. This chapter has summarized the development of talking face generation from different perspectives. Related work and recent progresses are discussed from the perspectives of audio representation, face modeling, audio-to-face animation, and post-processing. We have also listed commonly used public datasets and evaluation metrics. Finally, we discussed some open questions in the task of talking face generation.

Talking face generation techniques may be misused or abused for various malevolent purposes, e.g., fraud, aspersion, and dissemination of malicious propaganda. Out of ethical considerations, the government and researchers should jointly detect and combat harmful edited videos, and apply this technology without harming the public interest. We believe that with the dual attention of academia and industry, the generated videos will become more realistic with newly proposed models. In the future, there will also be more practical applications conducive to the public.

8.6 Further Reading

Interested readers are referred to the following further readings:

- Chen et al. [107] for a benchmark designed for evaluating talking-head video generation.
- Zhu et al. [120] for a survey on deep audio-visual learning.

References

1. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: *Proceedings of the advances in neural information processing systems*, vol 27
2. Radford A, Metz L, Chintala S (2016) Unsupervised representation learning with deep convolutional generative adversarial networks. In: *Proceedings of the international conference on learning representations*
3. Mirza M, Osindero S (2014) Conditional generative adversarial nets. *CoRR arXiv:abs/1411.1784*
4. Chen L, Li Z, Maddox RK, Duan Z, Xu C (2018) Lip movements generation at a glance. In: *Proceedings of the European conference on computer vision*, pp 520–535
5. Zhou H, Liu Y, Liu Z, Luo P, Wang X (2019) Talking face generation by adversarially disentangled audio-visual representation. In: *Proceedings of the AAAI conference on artificial intelligence*, vol 33, no 1, pp 9299–9306
6. Chen L, Maddox RK, Duan Z, Xu C (2019) Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 7832–7841
7. Song Y, Zhu J, Li D, Wang A, Qi H (2019) Talking face generation by conditional recurrent adversarial network. In: Kraus S (ed) *Proceedings of the international joint conference on artificial intelligence*, pp 919–925
8. Zhu H, Huang H, Li Y, Zheng A, He R (2020) Arbitrary talking face generation via attentional audio-visual coherence learning. In: *Proceedings of the international joint conference on artificial intelligence*, pp 2362–2368
9. Pham HX, Cheung S, Pavlovic V (2017) Speech-driven 3d facial animation with implicit emotional awareness: a deep learning approach. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp 80–88
10. Karras T, Aila T, Laine S, Herva A, Lehtinen J (2017) Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Trans Graph* 36(4):1–12
11. Taylor S, Kim T, Yue Y, Mahler M, Krahe J, Rodriguez AG, Hodgins J, Matthews I (2017) A deep learning approach for generalized speech animation. *ACM Trans Graph* 36(4):1–11
12. Fried O, Tewari A, Zollhöfer M, Finkelstein A, Shechtman E, Goldman DB, Genova K, Jin Z, Theobalt C, Agrawala M (2019) Text-based editing of talking-head video. *ACM Trans Graph* 38(4):1–14
13. Edwards P, Landreth C, Fiume E, Singh K (2016) Jali: an animator-centric viseme model for expressive lip synchronization. *ACM Trans Graph* 35(4):1–11
14. Zhou Y, Xu Z, Landreth C, Kalogerakis E, Maji S, Singh K (2018) Visemenet: audio-driven animator-centric speech animation. *ACM Trans Graph* 37(4):1–10
15. Lewis J (1991) Automated lip-sync: background and techniques. *J Visualization Comput Animat* 2(4):118–122
16. Guiard-Marigny T, Tsingos N, Adjoudani A, Benoit C, Gascuel M-P (1996) 3d models of the lips for realistic speech animation. In: *Proceedings of the computer animation*, pp 80–89
17. Bregler C, Covell M, Slaney M (1997) Video rewrite: driving visual speech with audio. In: *Proceedings of the annual conference on computer graphics and interactive techniques*, pp 353–360
18. Brand M (1999) Voice puppetry. In: *Proceedings of the annual conference on computer graphics and interactive techniques*, pp 21–28
19. Xie L, Liu Z-Q (2007) A coupled HMM approach to video-realistic speech animation. *Pattern Recogn* 40(8):2325–2340
20. Garrido P, Valgaerts L, Sarmadi H, Steiner I, Varanasi K, Perez P, Theobalt C (2015) Vdub: modifying face video of actors for plausible visual alignment to a dubbed audio track. *Comput Graph Forum* 34(2):193–204
21. Charles J, Magee D, Hogg D (2016) Virtual immortality: reanimating characters from TV shows. In: *Proceedings of the European conference on computer vision*, pp 879–886

22. Zhou Y, Han X, Shechtman E, Echevarria J, Kalogerakis E, Li D (2020) Makelttalk: speaker-aware talking-head animation. *ACM Trans Graph* 39(6):1–15
23. Cosatto E, Graf HP (2000) Photo-realistic talking-heads from image samples. *IEEE Trans Multimedia* 2(3):152–163
24. Fan B, Wang L, Soong FK, Xie L (2015) Photo-real talking head with deep bidirectional LSTM. In: *Proceedings of the IEEE international conference on acoustics, speech and signal processing*, pp 4884–4888
25. Vougioukas K, Petridis S, Pantic M (2019) Realistic speech-driven facial animation with gans. *Int J Comput Vis* 1–16
26. Prajwal KR, Mukhopadhyay R, Namboodiri VP, Jawahar CV (2020) A lip sync expert is all you need for speech to lip generation in the wild. In: *Proceedings of the ACM international conference on multimedia*, pp 484–492
27. Das D, Biswas S, Sinha S, Bhowmick B (2020) Speech-driven facial animation using cascaded gans for learning of motion and texture. In: *Proceedings of the European conference on computer vision*, pp 408–424
28. Yao X, Fried O, Fatahalian K, Agrawala M (2020) Iterative text-based editing of talking-heads using neural retargeting. *arXiv preprint [arXiv:2011.10688](https://arxiv.org/abs/2011.10688)*
29. Wu W, Zhang Y, Li C, Qian C, Loy CC (2018) ReenactGAN learning to reenact faces via boundary transfer. In: *Proceedings of the European conference on computer vision*, pp 603–619
30. Song L, Wu W, Fu C, Qian C, Loy CC, He R (2021) Everything’s talkin’: Pareidolia face reenactment. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*
31. Suwajanakorn S, Seitz SM, Kemelmacher-Shlizerman I (2017) Synthesizing Obama: learning lip sync from audio. *ACM Trans Graph* 36(4):1–13
32. Friesen E, Ekman P (1978) Facial action coding system: a technique for the measurement of facial movement. *Palo Alto* 3(2):5
33. Song L, Wu W, Qian C, He R, Loy CC (2020) Everybody’s talkin’: let me talk as you want. *arXiv [arXiv:abs/2001.05201](https://arxiv.org/abs/2001.05201)*
34. Mori M, MacDorman KF, Kageki N (2012) The uncanny valley [from the field]. *IEEE Robot Autom Mag* 19(2):98–100
35. Hannun AY, Case C, Casper J, Catanzaro B, Diamos G, Elsen E, Prenger R, Satheesh S, Sengupta S, Coates A, Ng AY (2014) Deep speech: scaling up end-to-end speech recognition. *CoRR abs/1412.5567*
36. Amodei D, Ananthanarayanan S, Anubhai R, Bai J, Battenberg E, Case C, Casper J, Catanzaro B, Cheng Q, Chen G et al (2016) Deep speech 2: end-to-end speech recognition in English and mandarin. In: *Proceedings of the international conference on machine learning*, pp 173–182
37. Chen D, Ren S, Wei Y, Cao X, Sun J (2014) Joint cascade face detection and alignment. In: *Proceedings of the European conference on computer vision*, pp 109–122
38. Chen D, Cao X, Wen F, Sun J (2013) Blessing of dimensionality: high-dimensional feature and its efficient compression for face verification. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 3025–3032
39. Seibold C, Samek W, Hilsman A, Eisert P (2017) Detection of face morphing attacks by deep learning. In: *Proceedings of the international workshop on digital watermarking*, pp 107–120
40. Lewenberg Y, Bachrach Y, Shankar S, Criminisi A (2016) Predicting personal traits from facial images using convolutional neural networks augmented with facial landmark information. In: *Proceedings of the AAAI conference on artificial intelligence*, vol 30, no 1
41. Di X, Sindagi VA, Patel VM (2018) GP-GAN: gender preserving GAN for synthesizing faces from landmarks. In: *Proceedings of the international conference on pattern recognition*, pp 1079–1084
42. Garrido P, Valgaerts L, Rehmsen O, Thormahlen T, Perez P, Theobalt C (2014) Automatic face reenactment. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 4217–4224

43. Pumarola A, Agudo A, Martinez AM, Sanfeliu A, Moreno-Noguer F (2018) Ganimation: anatomically-aware facial animation from a single image. In: Proceedings of the European conference on computer vision, pp 818–833
44. Blanz V, Vetter T (1999) A morphable model for the synthesis of 3d faces. In: Proceedings of the annual conference on computer graphics and interactive techniques, pp 187–194
45. Paysan P, Knothe R, Amberg B, Romdhani S, Vetter T (2009) A 3d face model for pose and illumination invariant face recognition. In: Proceedings of the IEEE international conference on advanced video and signal-based surveillance, pp 296–301
46. Besl PJ, McKay ND (1992) Method for registration of 3-d shapes. In: Sensor fusion IV: control paradigms and data structures, vol 1611. International Society for Optics and Photonics, pp 586–606
47. Kalogerakis E, Hertzmann A, Singh K (2010) Learning 3d mesh segmentation and labeling. *ACM Trans Graph* 29(4):1–12
48. Li T, Bolkart T, Black MJ, Li H, Romero J (2017) Learning a model of facial shape and expression from 4d scans. *ACM Trans Graph* 36(6):1–17
49. Zhu X, Liu X, Lei Z, Li SZ (2017) Face alignment in full pose range: a 3d total solution. *IEEE Trans Pattern Anal Mach Intell* 41(1):78–92
50. Guo J, Zhu X, Yang Y, Yang F, Lei Z, Li SZ (2020) Towards fast, accurate and stable 3d dense face alignment. *Proceedings of the European conference on computer vision* 12364:152–168
51. Cao C, Weng Y, Zhou S, Tong Y, Zhou K (2013) Facewarehouse: a 3d facial expression database for visual computing. *IEEE Trans Visualization Comput Graph* 20(3):413–425
52. Bolkart T, Wuhler S (2015) A groupwise multilinear correspondence optimization for 3d faces. In: Proceedings of the IEEE international conference on computer vision, pp 3604–3612
53. Blanz V, Romdhani S, Vetter T (2002) Face identification across different poses and illuminations with a 3d morphable model. In: Proceedings of the IEEE international conference on automatic face and gesture recognition, pp 202–207
54. Gecer B, Ploumpis S, Kotsia I, Zafeiriou S (2019) Ganfit: generative adversarial network fitting for high fidelity 3d face reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1155–1164
55. Zhou H, Liu J, Liu Z, Liu Y, Wang X (2020) Rotate-and-render: unsupervised photorealistic face rotation from single-view images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5911–5920
56. Kim H, Garrido P, Tewari A, Xu W, Thies J, Niessner M, Pérez P, Richardt C, Zollhöfer M, Theobalt C (2018) Deep video portraits. *ACM Trans Graph* 37(4):1–14
57. Thies J, Elgharib M, Tewari A, Theobalt C, Nießner M (2020) Neural voice puppetry: audio-driven facial reenactment. In: Proceedings of the European conference on computer vision, pp 716–731
58. Thies J, Zollhofer M, Stamminger M, Theobalt C, Nießner M (2016) Face2face: real-time face capture and reenactment of RGB videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2387–2395
59. Booth J, Roussos A, Zafeiriou S, Ponniah A, Dunaway D (2016) A 3d morphable model learnt from 10,000 faces. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5543–5552
60. Rubin S, Berthouzoz F, Mysore GJ, Li W, Agrawala M, Content-based tools for editing audio stories. In: Proceedings of the ACM symposium on user interface software and technology, pp 113–122
61. Garrido P, Zollhöfer M, Casas D, Valgaerts L, Varanasi K, Pérez P, Theobalt C (2016) Reconstruction of personalized 3d face rigs from monocular video. *ACM Transactions on Graphics* 35(3):1–15
62. Jamaludin A, Chung JS, Zisserman A (2019) You said that?: synthesising talking faces from audio. *Int J Comput Vision* 127(11):1767–1779
63. Wiles O, Koepke A, Zisserman A (2018) X2face: a network for controlling face generation using images, audio, and pose codes. In: Proceedings of the European conference on computer vision, pp 670–686

64. Guo Y, Chen K, Liang S, Liu Y, Bao H, Zhang J (2021) Ad-nerf: audio driven neural radiance fields for talking head synthesis. arXiv preprint [arXiv:2103.11078](https://arxiv.org/abs/2103.11078)
65. Chung JS, Zisserman A (2016) Out of time: automated lip sync in the wild. In: Proceedings of the Asian conference on computer vision, pp 251–263
66. Prajwal KR, Mukhopadhyay R, Philip J, Jha A, Nambodiri V, Jawahar CV (2019) Towards automatic face-to-face translation. In: Proceedings of the ACM international conference on multimedia, pp 1428–1436
67. Agarwal S, Farid H, Fried O, Agrawala M (2020) Detecting deep-fake videos from phoneme-viseme mismatches. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 660–661
68. Ezzat T, Geiger G, Poggio T (2002) Trainable videorealistic speech animation. *ACM Trans Graph* 21(3):388–398
69. Chang Y-J, Ezzat T (2005) Transferable videorealistic speech animation. In: Proceedings of the ACM SIGGRAPH/Eurographics symposium on computer animation, pp 143–151
70. Chen T (2001) Audiovisual speech processing. *IEEE Signal Process Mag* 18(1):9–21
71. Choi K, Luo Y, Hwang J-N (2001) Hidden markov model inversion for audio-to-visual conversion in an mpeg-4 facial animation system. *J VLSI Signal Process Syst Signal Image Video Technol* 29(1):51–61
72. Wang L, Qian X, Han W, Soong FK (2010) Synthesizing photo-real talking head via trajectory-guided sample selection. In: Proceedings of the annual conference of the international speech communication association
73. Karras T, Aila T, Laine S, Lehtinen J (2018) Progressive growing of gans for improved quality, stability, and variation. In: Proceedings of the international conference on learning representations
74. Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4401–4410
75. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T (2020) Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8110–8119
76. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. *Proc IEEE* 86(11):2278–2324
77. Devries T, Biswaranjan K, Taylor GW (2014) Multi-task learning of facial landmarks and expression. In: Proceedings of the Canadian conference on computer and robot vision, pp 98–103
78. Krizhevsky A et al (2009) Learning multiple layers of features from tiny images. Master's thesis, University of Tront
79. Lu Y, Tai Y-W, Tang C-K (2018) Attribute-guided face generation using conditional cyclegan. In: Proceedings of the European conference on computer vision, pp 282–297
80. Song L, Cao J, Song L, Hu Y, He R (2019) Geometry-aware face completion and editing. *Proc AAAI Conf Artif Intell* 33(1):2506–2513
81. Huang X, Belongie S (2017) Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision, pp 1501–1510
82. Shen Z, Huang M, Shi J, Xue X, Huang TS (2019) Towards instance-level image-to-image translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3683–3692
83. Yin Y, Jiang S, Robinson JP, Fu Y (2020) Dual-attention GAN for large-pose face frontalization. In: Proceedings of the IEEE international conference on automatic face and gesture recognition, pp 24–31
84. Qiao F, Yao N, Jiao Z, Li Z, Chen H, Wang H (2018) Geometry-contrastive GAN for facial expression transfer. arXiv preprint 1802.01822
85. Wang K, Wu Q, Song L, Yang Z, Wu W, Qian C, He R, Qiao Y, Loy CC (2020) Mead: a large-scale audio-visual dataset for emotional talking-face generation. In: Proceedings of the European conference on computer vision, pp 700–717

86. Zhou H, Sun Y, Wu W, Loy CC, Wang X, Liu Z (2021) Pose-controllable talking face generation by implicitly modularized audio-visual representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition
87. Ji X, Zhou H, Wang K, Wu W, Loy CC, Cao X, Xu F (2021) Audio-driven emotional video portraits. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition
88. Wang L, Han W, Soong FK (2012) High quality lip-sync animation for 3d photo-realistic talking head. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing, pp 4529–4532
89. Yi R, Ye Z, Zhang J, Bao H, Liu Y-J (2020) Audio-driven talking face video generation with natural head pose. arXiv preprint [arXiv:2002.10137](https://arxiv.org/abs/2002.10137)
90. Chen L, Cui G, Liu C, Li Z, Kou Z, Xu Y, Xu C (2020) Talking-head generation with rhythmic head motion. In: Proceedings of the European conference on computer vision, pp 35–51
91. Liu K, Ostermann J (2011) Realistic facial expression synthesis for an image-based talking head. In: IEEE international conference on multimedia and expo, pp 1–6
92. Cooke M, Barker J, Cunningham S, Shao X (2006) An audio-visual corpus for speech perception and automatic speech recognition. *J Acoust Soc America* 120(5):2421–2424
93. Harte N, Gillen E (2015) TCD-TIMIT: an audio-visual corpus of continuous speech. *IEEE Trans Multimedia* 17(5):603–615
94. Cao H, Cooper DG, Keutmann MK, Gur RC, Nenkova A, Verma R (2014) Crema-d: crowd-sourced emotional multimodal actors dataset. *IEEE Trans Affect Comput* 5(4):377–390
95. Chung JS, Zisserman A (2016) Lip reading in the wild. In: Proceedings of the Asian conference on computer vision, pp 87–103
96. Chung JS, Senior A, Vinyals O, Zisserman A (2017) Lip reading sentences in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3444–3453
97. Chung JS, Zisserman A (2017) Lip reading in profile. In: Proceedings of the British machine vision conference
98. Nagrani A, Chung JS, Xie W, Zisserman A (2020) Voxceleb: large-scale speaker verification in the wild. *Comput Speech Lang* 60:101027
99. Chung JS, Nagrani A, Zisserman A (2018) Voxceleb2: deep speaker recognition. In: In proceedings of the annual conference of the international speech communication association, pp 1086–1090
100. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process* 13(4):600–612
101. Salimans T, Goodfellow IJ, Zaremba W, Cheung V, Radford A, Chen X (2016) Improved techniques for training GANs. In: Proceedings of the neural information processing systems
102. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Proceedings of the international conference on neural information processing systems, pp 6629–6640
103. Narvekar ND, Karam LJ (2011) A no-reference image blur metric based on the cumulative probability of blur detection (CPBD). *IEEE Trans Image Process* 20(9):2678–2683
104. De K, Masilamani V (2013) Image sharpness measure for blurred images in frequency domain. *Procedia Eng* 64:149–158
105. Vougioukas K, Petridis S, Pantic M (2019) End-to-end speech-driven realistic facial animation with temporal GANs. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 37–40
106. Assael YM, Shillingford B, Whiteson S, De Freitas N (2016) Lipnet: end-to-end sentence-level lipreading. arXiv preprint [arXiv:1611.01599](https://arxiv.org/abs/1611.01599)
107. Chen L, Cui G, Kou Z, Zheng H, Xu C (2020) What comprises a good talking-head video generation?: a survey and benchmark. arXiv preprint [arXiv:2005.03201](https://arxiv.org/abs/2005.03201)
108. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 815–823

109. Deng J, Guo J, Xue N, Zafeiriou S (2019) Arcface: additive angular margin loss for deep face recognition. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 4690–4699
110. Zhang J, Zeng X, Xu C, Chen J, Liu Y, Jiang Y (2020) Apb2facev2: real-time audio-guided multi-face reenactment. *arXiv preprint [arXiv:2010.13017](https://arxiv.org/abs/2010.13017)*
111. Karson CN, Berman KF, Donnelly EF, Mendelson WB, Kleinman JE, Wyatt RJ (1981) Speaking, thinking, and blinking. *Psychiatry Res* 5(3):243–246
112. Hömke P, Holler J, Levinson SC (2018) Eye blinks are perceived as communicative signals in human face-to-face interaction. *PloS one* 13(12):e0208030
113. Shu Z, Shechtman E, Samaras D, Hadap S (2016) Eyeopener: editing eyes in the wild. *ACM Trans Graph* 36(1):1–13
114. Thies J, Zollhöfer M, Nießner M, Valgaerts L, Stamminger M, Theobalt C (2015) Real-time expression transfer for facial reenactment. *ACM Trans Graph* 34(6):1–14
115. Velinov Z, Papas M, Bradley D, Gotardo PFU, Mirdehghan P, Marschner S, Novák J, Beeler T (2018) Appearance capture and modeling of human teeth. *ACM Trans Graph* 37(6): 207:1–207:13
116. Yang L, Shi Z, Wu Y, Li X, Zhou K, Fu H, Zheng Y (2020) Iorthopredictor: model-guided deep prediction of teeth alignment. *ACM Trans Graph* 39(6):1–15
117. Wang T-C, Mallya A, Liu M-Y (2021) One-shot free-view neural talking-head synthesis for video conferencing. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*
118. Sadoughi N, Busso C (2019) Speech-driven expressive talking lips with conditional sequential generative adversarial networks. *IEEE Trans Affect Comput*
119. Livingstone SR, Russo FA (2018) The Ryerson audio-visual database of emotional speech and song (ravdess): a dynamic, multimodal set of facial and vocal expressions in north American English. *PloS one* 13(5):e0196391
120. Zhu H, Luo M-D, Wang R, Zheng A-H, He R (2021) Deep audio-visual learning: a survey. *Int J Autom Comput* 1–26

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.