electronics

MDPI

Review

Human-Computer Interaction System: A Survey of Talking-Head Generation

Rui Zhen [1], Wenchao Song [1], Qiang He [1,2], Juan Cao [1,*], Lei Shi [1,*] and Jia Luo [3]

[1] State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China
[2] State Key Laboratory of Media Convergence Production Technology and Systems, Beijing 100803, China
[3] College of Economics and Management, Beijing University of Technology, Beijing 100124, China
[*] Correspondence: caojuan@cuc.edu.cn (J.C.); leiky_shi@cuc.edu.cn (L.S.)

Abstract: Virtual human is widely employed in various industries, including personal assistance, intelligent customer service, and online education, thanks to the rapid development of artificial intelligence. An anthropomorphic digital human can quickly contact people and enhance user experience in human–computer interaction. Hence, we design the human–computer interaction system framework, which includes speech recognition, text-to-speech, dialogue systems, and virtual human generation. Next, we classify the model of talking-head video generation by the virtual human deep generation framework. Meanwhile, we systematically review the past five years' worth of technological advancements and trends in talking-head video generation, highlight the critical works and summarize the dataset.

Keywords: talking-head generation; virtual human generation; human–computer interaction

check for updates

Citation: Zhen, R.; Song, W.; He, Q.; Cao, J.; Shi, L.; Luo, J. Human-Computer Interaction System: A Survey of Talking-Head Generation. *Electronics* **2023**, *12*, 218. https://doi.org/10.3390/electronics12010218

Academic Editors: Peiying Zhang, Haotong Cao and Keping Yu

Received: 18 November 2022
Revised: 22 December 2022
Accepted: 27 December 2022
Published: 1 January 2023

1. Introduction

With the rapid development of artificial intelligence technology, virtual humans have been continuously applied in various scenarios, including virtual anchors, virtual customer service, and online education. In human–computer interaction, there is an anthropomorphic digital human who can quickly establish contact with users and improve user experience. Simultaneously, multimodal human–computer interaction is one of the application directions of virtual humans. The system aims to generate interactive objects with natural characteristics using deep learning models, including speech recognition, dialogue system, text-to-speech, and virtual human video synthesis. Among them, virtual human video generation is mainly divided into 2D/3D facial reconstruction, talking-head generation, body movements, and human movements. Meanwhile, in the talking-head generation task, we need to consider the audition consistency of lip shapes and facial attributes, such as facial expressions and eye movements.

In the research of talking-head generation, audio-driven lip synthesis is a popular research direction by inputting the corresponding audio and any mesh vertex, facial image, or video to synthesize the lip-synced talking-head video. In other words, the model dynamically maps the lower-dimensional speech or text signal to the higher-dimensional video signal. Note that text-driven lip synthesis is a natural extension of the task.

Prior to the popularity of deep learning, many researchers mainly adopted cross-modal retrieval methods [1–4] and Hidden Markov Model (HMM) to solve this problem [5]. However, cross-modal retrieval methods based on the mapping relationship between morphemes and visemes do not consider the contextual semantic information of the speech. Similarly, many factors, such as prior assumptions, also limit the application of HMM-based methods.

With the rapid improvement of computing power, the task of talking-head generation based on deep learning has attracted widespread attention, promoting the vigorous development of this field. The paper mainly makes a systematic review of the talking-head video synthesis model based on deep learning in the past five years. Figure 1 shows the literature map for talking-head generation. Along the timeline, the number of works has increased dramatically in recent years. According to the content of the model input, we can divide talking-head generation models into 2D-based methods and those based on a 3D approach. According to the method structure of the model, we can divide the talking-head generation technology into a pipeline and end-to-end types, as shown in Figure 2. However, in synthesizing talking-head video, most models take a relatively long time to generate video, and only a small part of models, such as DCK [6], can output results in a short time. More details are discussed in the third part of this paper.

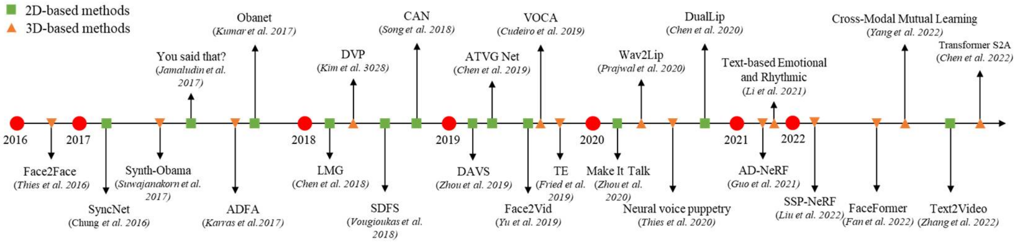
🟩 2D-based methods
▲ 3D-based methods
2016
2017
2018
2019
2020
2021
2022
Face2Face (Thies et al. 2016)
SyncNet (Chung et al. 2016)
Synth-Obama (Suwajanakorn et al. 2017)
You said that? (Jamaludin et al. 2017)
ADFA (Karras et al.2017)
Obanet (Kumar et al. 2017)
LMG (Chen et al. 2018)
DVP (Kim et al. 3028)
SDFS (Vougioukas et al. 2018)
CAN (Song et al. 2018)
DAVS (Zhou et al. 2019)
ATVG Net (Chen et al. 2019)
Face2Vid (Yu et al. 2019)
VOCA (Cudeiro et al. 2019)
TE (Fried et al. 2019)
Make It Talk (Zhou et al. 2020)
Wav2Lip (Prajwal et al. 2020)
Neural voice puppetry (Thies et al. 2020)
DualLip (Chen et al. 2020)
AD-NeRF (Guo et al. 2021)
Text-based Emotional and Rhythmic (Li et al. 2021)
SSP-NeRF (Liu et al. 2022)
FaceFormer (Fan et al. 2022)
Cross-Modal Mutual Learning (Yang et al. 2022)
Text2Video (Zhang et al. 2022)
Transformer S2A (Chen et al. 2022)

Figure 1. An overview of recent works on talking-head generation [3,7–31].

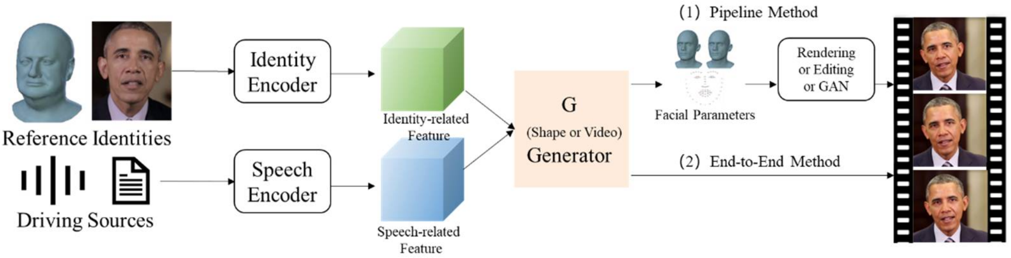

Figure 2. Classification of talking-head video generation methods.

In the research process of the whole task, datasets and evaluation metrics are an indispensable part of training the talking-head generative model. During the widespread application of deep learning methods, the emergence of large-scale datasets has driven the further development of models for talking-head video generation and served as a general platform for measuring and comparing different algorithms. However, science and technology have two sides, such as deepfake. In order to prevent technology from being used to harm the country and society, only a small part of the data set is completely open-sourced, and some are obtained by application (Note: The application authority for this part of data is only open to researchers, teachers, and engineers from universities, research institutes, and enterprises. Students are prohibited from applying.). In the fourth part, we review commonly used datasets, including statistics, highlights, and download links.

Now, we can summarize our main contributions in this paper:

1. We present a systematic framework for multimodal human–computer interaction, which provides a new idea for the application of talking-head generation models.
2. We propose two taxonomies to group methods with important reference significance and analyze the strengths and weaknesses of representative methods and their potential connections.
3. We summarize the datasets and evaluation metrics commonly used for talking-head generation models. Meanwhile, we highlight the significance of the consumption time to generate videos as a measure of model performance.

The rest of the paper is organized as follows: Section 2 describes the architecture of a multimodal human–computer interaction system, including a voice module, dialogue system, and talking-head generation module; Section 3 surveys two different methods of generating virtual human talking-head in recent five years: pipeline and end-to-end; Section 4 discusses the dataset used to train the virtual human talking-head generation model and the indicators used to evaluate the model performance; we described three potential methods to improve the speed of virtual human talking-head generation model in Section 5 and conclude in Section 6.

2. Human–Computer Interaction System Architecture

Based on artificial intelligence technologies, such as natural language processing, voice, and image processing, the system pursues multimodal interaction with low-latency and high-fidelity anthropomorphic virtual humans. As shown in Figure 3, the system is mainly composed of four modules: (1) the system converts the voice information input by the user into text information through the automatic speech recognition (ASR) module; (2) the dialogue system (DS) takes the text information output by the ASR module as input; (3) the text-to-speech (TTS) module converts the text output by the DS into realistic speech information; (4) the talking-head generation module preprocesses the picture, video, or blendshape as the model input to extract its facial features. Then, the model maps the lower-dimensional voice signal by the TTS module to the higher-dimensional video signals, including the mouth, expression, motion, etc. Finally, the model uses the rendering system to fuse the features and multimodal output video and displays it on the user side.

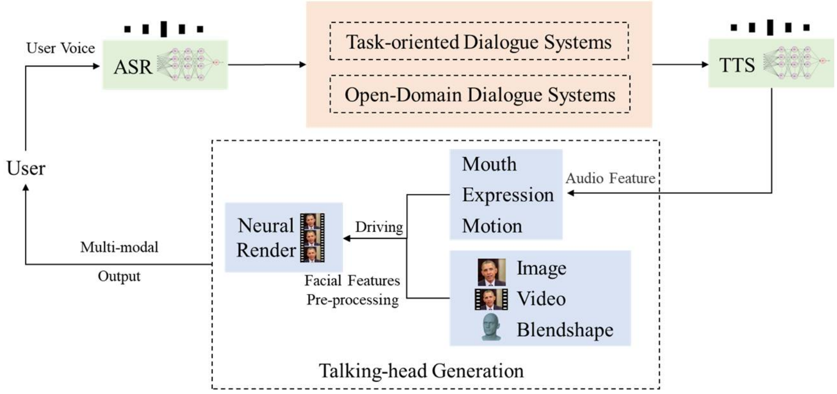

Figure 3. The system architecture of multimodal human–computer interaction.

2.1. Speech Module

The ASR and TTS of the speech module correspond to the human hearing and language function, respectively. After decades of research, speech recognition and text-to-speech synthesis have been widely used in various commercial products. We use the PaddleSpeech [32], a code, open-sourced by Baidu. One model can complete both ASR and TTS tasks, which greatly reduces the complexity of model deployment and enables better collaboration with other modules. In addition, we can also choose API services provided by commercial companies, such as Baidu, Sogou, iFLYTEK, etc.

2.2. Dialogue System Module

Our dialogue system module needs to have the ability to have multiple rounds of dialogue. The system needs to answer domain-specific questions and meet users' needs to chat. As shown in Figure 3, the question is passed to the dialogue module after the user's voice passes through the ASR. The dialogue module must retrieve or generate matching answers from the knowledge base according to the user's question. However, it is impossible to rely entirely on the model to generate answers in a specific domain multi-turn dialogue. In some scenarios, to better consider the context information, the above information will be aggregated to identify the user's intent and return the answer in the way of QA.

2.3. Talking-Head Generation

The facial appearance data of the talking-head generation module mainly comes from real-person photos, videos, or blendshape character model coefficients. Taking video as an example, we first perform video preprocessing on these facial appearance data and then map the audio signal of TTS in Figure 3 to higher-dimensional signals such as human face lip shape, facial expression, and facial action, and finally, use a neural network. The model performs video rendering and outputs multimodal video data.

In human–computer interaction, a timely response can improve user experience. However, the time delay of the whole system is equal to the sum of the time consumed by each data processing module. Among them, the voice module and the dialogue module have been commercialized by a wide range of users, which can meet the real-time requirements of human–computer interaction. At present, it takes a long time for the talking-head generation model to render and output multimodal video. Therefore, we need to improve the data processing efficiency of the talking-head generation model, reduce the rendering time of the multimodal video, and reduce the response time of the human–computer interaction system extension. Although the virtual human has achieved low-latency response in some commercial products such as JD's ViDA-MAN [33], etc., the long production cycle, high cost, and poor portability are also problems that cannot be ignored.

3. Talking-Head Generation

Talking-head video generation, i.e., lip motion sequence generation, aims to synthesize lip motion sequences corresponding to the driving source (a segment of audio or text). Based on synthesizing the lip motion, the video synthesis of the talking head also needs to consider its facial attributes, such as facial expressions and head movements.

In the early talking-head video generation methods, researchers mainly used cross-modal retrieval [3] and HMM-based methods [34] to realize the dynamic mapping of driving sources to lip motion data. However, these methods have relatively high requirements on the application environment of the model, visual phoneme annotation, etc. Thies et al. [3] introduced an image-based lip-motion synthesis method, which generates the real oral cavity by retrieving and selecting the optimal lip shape from offline samples. However, the method is based on text–morpheme–morpheme. The retrieval of the map does not truly take into account the contextual information of the speech. Zhang et al. [30] introduced key pose interpolation and smoothing modules to synthesize pose sequences based on cross-modal retrieval and used a GAN model to generate videos.

Recently, the rapid development of deep learning technology has provided technical support for talking-head video generation and promoted the vigorous development of talking-head video generation methods. Figure 1 shows that the image dimension of the talking head can be divided into 2D-based and 3D-based methods. Figure 2 shows that the talking-head video generation framework based on deep learning can be roughly divided into two types: pipeline and end-to-end. Table 1 summarizes the representative works on talking-head video generation.

(1) **2D-based methods**.

In 2D-based methods, talking-head generation mainly used landmarks, semantic maps, or other image-like representations to solve the problem, which dates back to Bregler et al. 1997 [4]. In talking-head video generation, Chen et al. [17] used landmarks as an intermediate layer for mapping from low-dimensional audio to high-dimensional video and divided the whole method into two stages. Chung et al. [9] used two decoders to decouple the voice and the speaker identity to generate the video without the influence of the speaker identity. Lip synthesis can also use image-to-image translation to generate [35] an extension of this method. Zhou et al. [16] and Song et al. [15] use a combination of separate audio-visual representations and neural networks to optimize the synthesis.

(2) **3D-based methods**.

Early 3D-based approaches pre-built 3D models of specific people and then rendered the models. Compared to 2D methods, this method can have better control over motion. However, the construction cost of such a 3D model is relatively high, and the effect of changing a new identity cannot be guaranteed. In synthesizing Barack Obama's videos, these works [8,11] synthesize realistic speaking facial videos by pre-constructing 3D facial models and learning to map audio sequences to video sequences to drive the model. In addition, there are many generative talking-head models based on 3DMM parameters [10,19,20,23], and models, such as blendshape [19], flame [36], and 3D mesh [37], are used with the audio as model input. Among them, VOCA [16] uses the blendshape of the character's head to create the model. Meshtalk [37] uses the neutral face template mesh as the basis to generate the talking-head video. However, the model with intermediate parameter 3DMM brings a certain loss of information. Moreover, VOCA is an independent 3D talking-head synthesis model that can capture different speech styles, while Meshtalk can parse out the absolute latent space of audio-related and audio-independent facial movements.

Table 1. The main model of talking-head generation in recent years. ID: The model can be divided into three types: identity-dependent (D), identity-independent(I), and hybrid(H). Driving Data: Audio(A), Text(T), and Video(V).

References	Key Idea	Driving Factor	ID D/I	3D Model
Suwajanakorn [8]	Audio-to-mouth editing to video	A	D	3D
Karras [10]	From audio and emotion-state to 3D vertices	A	D	3D
Kumar [11]	Text-to-audio-to-mouth key-points to video	T	D	2D
[9,12]	Joint embedding of audio and identity features	A	I	2D
Kim [13]	DVP: parameter replacement and facial reenactment with cGAN to video	V	I	3D
Vougioukas [14]	Audio-driven GAN	A	I	2D
Zhou [16]	Joint embedding of person-id and word-id features	V or A	I	2D
Chen [17]	From Audio to facial landmarks to video synthesis	A	I	2D
Yu [18]	From text or audio feature to facial landmarks to video synthesis	A and T	I	2D
Cudeiro [19]	VOCA: from audio to FLAME head model with facial motions	A	I	3D
Fried [20]	3D reconstruction and parameter recombination	T	D	3D
Zhou [21]	Audio-driven landmark prediction	A	I	2D
Prajwal [22]	Wav2Lip: audio-driven, based GAN lip-sync discriminator	A	I	2D
Thies [23]	NVP: from the fusion of audio expression feature extraction and intermediate 3D model to video	A	H	3D
Guo [25]	AD-NeRF: Audio-to-video generation via two individual neural radiance fields	A	D	3D
Li [26]	TE: text-driven to video generation combine phoneme alignment, viseme search and parameter blending	T	D	3D
Fan [28]	FaceFormer: Audio-to-3D Mesh to video	A	I	3D
Yang [29]	A unified framework for visual-to-speech recognition and audio-to-video synthesis	A	I	3D
Zhang [30]	Text2Video: GAN+phoneme-pose dictionary	T	I	3D

Most current methods reconstruct 3D models from training videos directly. NVP (neural voice puppetry) has since designed the Audio2ExpressionNet and the 3D model of the independent identity. NeRF (Neural Radiance Fields) [38–41], which simulates implicit representation with MLP, can store 3D spatial coordinates and appearance information and be used for large-resolution scenes. To reduce information loss, AD-NeRF [25] trains two NeRFs for head and drive rendering of talking-head synthesis and obtains good visual effects. Many models require unrestricted universal identity and speech as input in practical application scenarios. Prajwal et al. [22,42] take any unidentified video and arbitrary speech as input to synthesize unrestricted talking-head video.

This section will mainly introduce problem formulation and the framework of talking-head generation with pipeline and end-to-end.

3.1. Problem Formulation

Given an input audio A and a subject reference video V, our talking-head video generation aims to synthesize an action-natural video S synchronized with A. The general steps to generate a neural talking head can be expressed as follows:

$$\begin{gathered} \mathcal{F}_{lips} = G(E(A)) \\ S = R\left(\mathcal{F}_{lips}, V\right) \end{gathered} \tag{1}$$

The explicit features created by the generator G are designated by the term $\mathcal{F}_{lips}$. E denotes the audio feature extraction network, and R the rendering network to convert the synthesized features into the output video.

In the speech feature extraction network, existing methods often use Mel Frequency Cepstral Coefficients (MFCC) to extract audio features. Previous studies have found that identity information coupled with audio features will reduce the accuracy of mapping from speech to lip movements, so it is necessary to extract audio content representations that are independent of identity features:

$$\begin{gathered} E(A) = W_i x + b_i = \widetilde{W_i}\widetilde{x}\ , \\ where\ \widetilde{W_i} = (W_i,\ \ b_i),\ \ \widetilde{x} = (x;1), \\ \widetilde{W_i} = I + \sum_{j=1}^{m} \gimel_j\, \widetilde{W_j} \end{gathered} \tag{2}$$

In Equation (2), x and $E(A)$ denote the raw and the transferred audio feature, respectively, while $\widetilde{W_i} = I + \sum_{j=1}^{m} \gimel_j\, \widetilde{W_j}$ represents the identity information adaptation parameter that is factorized into an identity matrix I plus the weighted sum of m components $\widetilde{W_j}$, and the parameters $\gimel_j$ need to be learned from the input audio feature [43].

In the rendering network from speech to video, the existing models respectively introduce network structures such as U-Net, GAN, Vision Transformer (ViT), and the emerging NeRF.

1. In generating talking-head videos using GAN, wav2lip [22] proposes the expert lip-sync discriminator based on SyncNet, and the formula is as follows:

$$D_{sync} = \frac{v \cdot E(A)}{max(\|v\|_2 \cdot \|E(A)\|_2,\ \mathcal{E})} = \frac{v \cdot f_{lip}}{max\left(\|v\|_2 \cdot \|f_{lip}\|_2,\ \mathcal{E}\right)} \tag{3}$$

In Equation (3), R computes the video embedding v from an image encoder and the audio embedding *E(A)* from an audio encoder. We replace the low-dimensional audio embedding of *E(A)* with the mapped high-dimensional video embedding f_{lip} to facilitate the calculation of the cosine similarity of the synchronous probability between v and f_{lip}.

2. In generating talking-head videos using ViT, FaceFormer [28] propose a novel seq2seq architecture to autoregressively predict facial movements, and the formula is as follows:

$$\hat{v}_t = FaceFormer_{\theta}(\hat{v} < t,\ s_n, E(A)\ ) \tag{4}$$

In Equation (4), θ denotes the model parameters, t is the current time-step in the sequence, $\hat{v}_t$ denotes the decoder autoregressively prediction result conditioned on $E(A)$, and s_n denotes the speaker identities, not from raw data.

$$\begin{aligned} \hat{v}_t &= v_t + PPE(t) \\ where\ PPE_{(t,2i)} &= \sin\Big((t\ mod\ p)/1000^{2i/d}\Big), \\ PPE_{(t,2i+1)} &= \cos\Big((t\ mod\ p)/1000^{2i/d}\Big), \\ v_t &= \begin{cases} (W^v \cdot \hat{y}_{t-1} + b^v) + s_n,\ 1 < t \le T, \\ s_n,\ \ t = 1 \end{cases} \end{aligned} \tag{5}$$

In Equation (5), t denotes the token position, d the model dimension, i the dimension index, W^v the weight, b^v the bias, $\hat{y}_{t-1}$ the prediction from the last time, and PPE is applied to v_t to provide the temporal order information periodically.

3. NeRF is a powerful and elegant 3D scene representation framework. It can encode the scene into a 3D volume space using MLP F_θ and then renders the 3D volume into an image by integrating color and densities along camera rays. The formula is as follows:

$$\begin{gathered} (c,\ \sigma) = F_\theta(p, d) \\ where\ p = (x,\ y,\ z),\ d = (\theta, \phi) \end{gathered} \tag{6}$$

In Equation (6), p denotes the query point in 3D voxel space, d the 2D view direction, c and σ denote RGB color and densities along the dispatched rays, respectively.

In generating talking-head videos using NeRF, the audio-driven talking-head generation formula is as follows:

$$(c,\ \sigma) = F_\theta(p, d, A) \tag{7}$$

In Equation (7), θ and A denote the weight and the audio embedding, respectively.

3.2. Pipeline

The Pipeline methods are mainly divided into two steps: low-dimensional driving source data are mapped to facial parameters; then, GPU rendering, video editing, or GAN is used to convert the learned facial parameters into high-dimensional video output.

According to the data type of the facial parameters, the Pipeline methods can be divided into Landmark-based, Coefficient-based, and Vertex-based methods.

3.2.1. Landmark-Based Method

Face landmarks are widely used in various face analysis tasks, including head video synthesis. In their pioneering work, Suwajanakorn et al. [8] used a single-layer LSTM to map low-dimensional speech data into nonlinear lip key points, followed by face texture synthesis, video retiming, and target video synthesis, in turn. Kumar et al. [11] proposed the architecture of LSTM+UNet and used Pix2Pix instead of the Pipeline-based video synthesis method to improve the model. Meanwhile, the LSTM+UNet architecture has also been widely used in many works [21,44].

Due to the wide range of application scenarios of the synthesized video of the talking head, a method that is not limited by the input voice and identity is needed. Therefore, the works [8,11] that only use the Obama-speaking video as the data cannot meet the business requirements and synthesize another person or voice. Jalalifar et al. [45] introduced the basic conditional generative adversarial network (C-GAN) as a standalone module for the problem of audio-to-video mapping to generate videos given face landmarks. Since the two modules are independent, the model can use any audio as a driving source to synthesize a new video. Chen et al. [17] further considered the correlation between the video frames before and after the synthesis process. They proposed a dynamic pixel-level loss to solve the pixel jitter problem in the target area. However, in the generative adversarial network part of the model, due to the insufficient accuracy of dlib [46] detector lip landmarks, there is an error with the lip landmark data of the dataset, which affects the effect of the model's output video.

In addition to methods for 2D landmarks, the mapping of low-dimensional driving source data to high-dimensional 3D landmarks has also been extensively studied. The speech signal contains not only semantic-level information but also information such as speech, speech style, and emotion. Zhou et al. [21] used neural networks to learn separate speech content and identity features, predicted 3D landmarks with speech content features, and synthesized talking-head videos with a UNet-style generator.

3.2.2. Coefficients-Based Method

2D Coefficient based. The active appearance model (AAM) is one of the most commonly used facial coefficient models and represents the variation in shape, texture, and their correlations. Fan et al. [47] used overlapping triphones as the data input for a two-layer Bi-LSTM model to learn AAM coefficients in the lip region and then mapped the learned data to face images to synthesize talking-head videos. However, AAM coefficients lead to potential errors and limited flexibility when transferring reference faces to new objects.

3D Coefficient based. In addition to 2D facial coefficient models, [48,49] proposed CNN+RNN-based models to map low-dimensional speech signals to blendshape coefficients of 3D faces. Thies et al. [23] proposed a CNN-based Audio2Expression network and a content-aware filtering network, which can map any person's speaking-voice sequence to a 3D blendshape that can represent a specific person's speaking style. Meanwhile, the method of NVP [23] is first to infer emotion from voice, thereby rendering high-quality speaking-heads video.

Many methods just control and generate lip movements and facial expressions, but these methods cannot synthesize full talking-head videos under full 3D head control. Kim et al. [13] introduced the 3D Morphable Model (3DMM, a denser representation of 3D face parameters) [50] to talking-head generation, and the method can completely control the action parameters, such as facial movements, expressions, and eyes, or only adjust the facial expression parameters and keep the others unchanged. The 3DMM coefficients include rigid head pose parameters, facial recognition coefficients, expression coefficients, binocular gaze direction parameters, and spherical harmonic illumination coefficients. Zhang et al. [51] proposed a framework with a style-specific animation generator and flow-guided video generator to synthesize high-visual quality videos. Among them, a style-specific animation generator can successfully separate lip movements from eyebrow and head poses. Since the method does not consider temporal coherence, the lips in the generated talking-head video may be disturbed. Simultaneously, the regularized head-pose and eye-movement parameters limit the motion space of the entire 3D head. Ji et al. [52] proposed an emotional video portrait (EVP) to achieve speech-driven video synthesis that can control the emotions of talking heads and faces.

3.2.3. Vertex-Based Method

3D facial vertices are other commonly used 3D models for talking-head video synthesis. For example, Karras et al. [10] used a deep neural network to learn a nonlinear mapping from input audio to 3D vertex coordinates corresponding to a fixed topology mesh. At the same time, an additional emotional code is designed to learn the corresponding emotional state from the training data to control the facial expressions of the talking head. However, many proposed models are mainly for speaker audio with specific identities. To solve this problem, Cudeiro et al. [19] proposed a VOCA model, which fuses the audio features extracted by DeepSpeech with the feature vectors of different speakers and outputs the displacement data of 3D vertices. The main contribution of VOCA is to solve the coupling problem of facial identity and facial motion, using identity control parameters to change its visual dynamics. Since the method uses a high-definition 4D dataset in the laboratory, it cannot be trained with wild videos. Fan et al. [28] proposed a FaceFormer model based on Transformer, which encodes the context information of long-term audio and predicts a series of animated 3D face meshes by autoregression.

Richard et al. [37] proposed a latent space for facial animation classification based on a cross-modal loss that not only disentangles audio-related and audio-unrelated information, such as facial actions (blinking and movement of the eyebrows). However, some researchers introduced the neural network architecture of the UNet-style decoder with additive skip connections. The method can predict 3D vertex coordinates, disentangle the motion of the lower and upper surface regions, and prevent over-smoothing, synthesizing a more plausible, realistic talking-head video. To guarantee high-fidelity video quality, the model requires a large-scale high-definition 3D training dataset.

3.3. End-to-End

Before 2018, the pipeline methods of talking-head video generation were a major research direction. However, this pipeline-based method has a complex processing flow, expensive and time-consuming facial parameter annotation, and additional auxiliary techniques, such as facial landmark detection and 3D/4D/5D face reconstruction. Therefore, many researchers began to explore the end-to-end talking-head video synthesis method. The end-to-end approach refers to an architecture that generates talking-lip (face) videos directly from the driving source without involving the facial parameters of any intermediate links.

Specch2vid, proposed by Chung et al. [9], is one of the earliest frameworks to explore end-to-end synthetic talking-face videos. As shown in Figure 4, it consists of four modules: an audio encoder, an identity image encoder, a speaker face image decoder, and a deblurring module. The voice encoder is designed to extract speech features from raw audio; the identity image encoder is designed to extract identity features from the input image; the speaker face image decoder takes the speech and identity features as input through a transposed convolution and up-sampling method to perform feature fusion and output the synthesized image. However, in the above process to obtain high-quality images, the model replaces the L2 loss function commonly used in image generation and autoencoders with an L1 loss function. In addition, a CNN-based deblurring module is separately trained to improve the quality of the output images. However, the shortcomings of this model are also obvious: (1) Since Specch2vid does not consider the continuity in the time series, it will produce incoherent video sequences with skipped frames or jitters; (2) The L1 reconstruction loss is performed on the whole face, and it is difficult to infer multiple facial expressions of a person from single audio. Note: the images of the political figure Obama used for academic research in this article are mainly derived from the dataset.

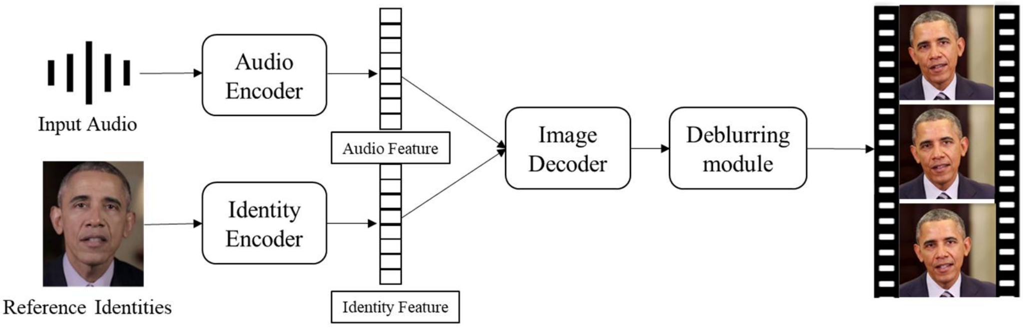

Figure 4. An overview of Specch2vid structure.

To overcome the above limitations of Speech2Vid, many researchers have proposed new solutions [16,53–55] by utilizing generative adversarial training strategies [56]. Taking the audio-driven talking-head video generation model as an example, a piece of audio contains various information such as speech, emotion, and speaking style. Hence, decoupling the complex audio information is a significant problem in the talking-head video generation task. To alleviate this problem, Zhou et al. [16] proposed the detachable Audio-visual System (DAVS). The supervised adversarial training model focused more on extracting speech and identity feature information than previous methods. However, such methods were too reliant on extra Word-ID and Person-ID tags in the training phase. This method ignored the correlation between head pose and audio [57]. Si et al. [53] used knowledge distillation to separate emotional, identity, and speech features from audio input with the help of a pre-trained emotion-recognition teacher network and a pre-trained face-recognition teacher network. Recently, many researchers introduced coded attributes, such as facial expression, head posture, and blink frequency, into the model to generate a more natural talking head. For example, ref [58,59] introduced the emotion encoder into the model, and [60] designed an implicit pose encoding module into the generation pipeline. Ji et al. [61] designed an Audio2Facial-Dynamics module to learn the movement of facial key points and the displacement of implicit emotions from audio. Biswas et al. [62] proposed a speech-driven method for synthesizing speaking faces, which can achieve coherent head movement, accurate lip-shape synchronization, natural blinking, and high-fidelity texture. Waibel et al. [63] proposed an end-to-end neural system for the lip-synchronous translation of videos for videos in another language.

GAN-based methods focus on tailoring more efficient learning objectives for the talking-head video generation model to avoid the disadvantage of using only image reconstruction loss. Prajwal et al. [22,42] introduced a simple audio-visual synchronization discriminator for synthesizing speech and lip-synced talking-head videos. In addition, Chen et al. [12] proposed an audio-visual derivative correlation loss to optimize the consistency of the two modalities in the feature space. They proposed a three-stream GAN discriminator to force generation from the input audio signal Talking Mouth Video. Biswas et al. [62] propose an attention-based GAN network to identify audio features related to head movement and can also learn the important correlation between the prosodic features of speech and lip synchronization, blinking, and head movement.

In addition to the GAN-based end-to-end approach, the researchers were inspired by the neural radiation field (NeRF) [38]. Guo et al. [25] proposed the audio-driven neural radiation field (AD-NeRF) model. AD-NeRF-integrated DeepSpeech Audio features are used as conditional input to learn an implicit neural scene representation function that maps audio features to dynamic neural radiation fields for speaker-face rendering. AD-NeRF can model the head and upper body by learning two separate neural radiation fields and can also manipulate attributes such as action pose and background replacement, but this method cannot generalize the mismatched driving speech and speaker. However, AD-NeRF often suffers from head and torso separation during the rendering stage, resulting in unnature synthesized video. Therefore, Liu et al. [27] proposed a method called semantic-aware speak portrait NeRF (SSP-NeRF), which uses the semantic awareness of speech to address the problem of incongruity between local facial dynamics and global head–torso. Meanwhile, the problem cannot be ignored by the slow rendering speed of NeRF. These methods [41,64–66] improve the rendering speed of NeRF. Different from the fusion strategy of the previous pipeline method, Ye et al. [6] proposed a fully convolutional neural network with a dynamic convolution kernel (DCK) for cross-modal feature fusion and audio-driven face video generation for multimodal tasks and is robust to different identities, head poses, and audio. The real-time performance of the talking-head video generation model is significantly improved due to the simple and efficient network architecture. Yao et al. [67] proposed a novel framework based on the neural radiation field. Among them, lip movement is predicted directly from the input audio to achieve the synchronization of sound and picture. A transformer variational automatic encoder based on Gaussian process sampling is also designed to learn reasonable and natural personalized attributes, such as head posture and blinking.

4. Datasets and Evaluation Metrics

4.1. Datasets

In the era of artificial intelligence, the dataset is an important part of the deep learning model. At the same time, data sets also promote the solution of complex problems in the field of virtual human synthesis. However, in practical applications, there are few high-quality annotation data sets that cannot meet the training needs of the speech synthesis model. Moreover, many institutions/researchers are affected by the deepfake technical ethics issues, which increase the difficulty of obtaining some data sets. For example, only researchers, teachers, and engineers from universities, research institutions, and enterprises are allowed to apply, and students are prohibited from applying. In Table 2, we briefly highlighted the data sets commonly used by most researchers, including statistics and download links.

Table 2. Summary of talking-head video datasets.

Dataset Name	Year	Hours	Image Size FPS	Speaker	Sentence	Head Movement	Envir.
GRID	2006	27.5	720 × 576, 25	33	33 k	N	Lab
LRW	2017	173	256 × 256, 25	1 k+	539 K	N	TV
ObamaSet	2017	14	N/A	1	N/A	Y	TV
VoxCeleb2	2018	2.4 k	N/A, 25	6.1 k+	153.5 K	Y	TV
LRW-1000	2019	57	N/A	2 K+	718 K	Y	TV
VOCASET	2019	N/A	5023 vertices, 60	12	255	Y	Lab
MEAD	2020	39	1920 × 1080, 30	60	20	Y	Lab
HDTF	2021	15.8	N/A	362	10 K	Y	TV

The GRID [68] dataset was recorded in a laboratory setting with 34 volunteers, which is relatively small in a large dataset, but each volunteer spoke 1000 phrases for a total of 34,000 utterance instances. The phrase composition of the dataset conforms to certain rules. Each phrase contains six words, randomly selected from each of the six types of words to form a random phrase. The six categories of words are “command”, “color”, “preposition”, “letter”, “number”, and “adverb”. Dataset is available at https://spandh.dcs.shef.ac.uk//gridcorpus/, accessed on 30 December 2022.

LRW [69], known for various speaking styles and head poses, is an English-speaking video dataset collected from the BBC program with over 1000 speakers. The vocabulary size is 500 words, and each video is 1.16 s long (29 frames), involving the target word and a context. Dataset is available at https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrw1.html, accessed on 30 December 2022.

LRW-1000 [70] is a Mandarin video dataset collected from over 2000 vocabulary subjects. The purpose of the dataset is to cover the natural variation of different speech modalities and imaging conditions to incorporate the challenges encountered in real-world applications. There are large variations in the number of samples in each category, video resolution, lighting conditions, and attributes such as speaker pose, age, gender, and makeup. Note: the official URL (http://vipl.ict.ac.cn/en/view_database.php?id=13, accessed on 30 December 2022.) is no longer available, you can go to the paper page for details about the data and download the agreement file here (https://github.com/VIPL-Audio-Visual-Speech-Understanding/AVSU-VIPL, accessed on 30 December 2022.) if you plan to use this dataset for your research.

ObamaSet [8] is a specific audio-visual dataset that focuses on analyzing the visual speech of former US President Barack Obama. All video samples are collected from his weekly address footage. Unlike previous datasets, it only focuses on Barack Obama and does not provide any human annotations. Dataset is available at https://github.com/supasorn/synthesizing_obama_network_training, accessed on 30 December 2022.

VoxCeleb2 [71] is extracted from YouTube videos, including the video URL and discourse timestamp. At the same time, it is currently the largest public audio-visual data set. Although this dataset is used for speaker recognition tasks, it can also be used to train a talking-head generation model. However, the dataset needs to apply to obtain the download permission to prevent misuse of the dataset. The URL for the permit application is https://www.robots.ox.ac.uk/~vgg/data/voxceleb/, accessed on 30 December 2022. Because the dataset is huge, it requires 300 G+ storage space and supporting download tools. The download method is available at https://github.com/walkoncross/voxceleb2-download, accessed on 30 December 2022.

VOCASET [19] is a 4D-face dataset with approximately 29 min of 4D face scans, synchronized audio from 12-bit speakers (six women and six men), and recorded 4D-face scans at 60 fps. As a representative high-quality 4D face-to-face audio-visual dataset, Vocaset greatly facilitates the research of 3D VSG. Dataset is available at https://voca.is.tue.mpg.de/, accessed on 30 December 2022.

MEAD [44], the Multi-View Emotional Audio-Visual Dataset, is a large-scale, high-quality emotional audio-video dataset. Unlike previous datasets, it focuses on facial generation for natural emotional speech and takes into account multiple emotional states (eight different emotions on three intensity levels). Dataset is available at https://wywu.github.io/projects/MEAD/MEAD.html, accessed on 30 December 2022.

HDTF [51], a large in-the-wild high-resolution audio-visual dataset, stands for the High-definition Talking-Face Dataset. The HDTF dataset consists of approximately 362 different videos of 15.8 h. The resolution of the original video is 720 P or 1080 P. Each cropped video is resized to 512 × 512. Dataset is available at https://github.com/MRzzm/HDTF, accessed on 30 December 2022.

4.2. Evaluation Metrics

The evaluation task of talking-head video generation is an open problem that requires the evaluation of generation results from both objective and subjective aspects. Chen et al. [72] reviewed several state-of-the-art talking-head generation methods. They designed a unified benchmark based on their properties. Subjective evaluation is often used to compare the generated content's visual quality and sensory effects, such as whether lip-sync audio is in sync with the picture. Due to the high cost of subjective factors in the evaluation process, many researchers have attempted to quantitatively evaluate generated content using objective metrics [22,28,29,52]. These metrics can be classified into three types: visual quality, audio-visual semantic consistency, and real time based on quantitative model performance evaluations.

Visual Quality. Reconstruction error measures (e.g., mean squared error) are a natural way to evaluate the quality of generated video frames. However, the reconstruction error only focuses on the pixel alignment, ignoring the global visual quality. Therefore, existing works typically employ the peak signal-to-noise ratio (PSNR), structural similarity index metric (SSIM) [29,73], and learned perceptual image patch similarity (LPIPS) [74] to evaluate the global vision of generated image quality. Since metrics, such as PSNR and SSIM, cannot explain human perception well, and LPIPS is closer to human perception in visual similarity judgment, it is recommended to use LPIPS to evaluate the visual quality of generated content quantitatively. More recently, Prajwal et al. [22] introduced the Fréchet inception distance (FID) [75] to measure the distance between synthetic and real data distributions, as FID is more consistent with human perception assessments.

Audio-visual semantic consistency. The semantic consistency of the generated video and the driving source mainly includes audio-visual synchronization and speech consistency. For audio-visual synchronization, the landmark distance (LMD) [12] computes the Euclidean distance of lip region landmarks between the synthetic video frame and the ground truth frame. Another synchronization evaluation metric uses SyncNet [7] to predict the offset of generated frames and ground truth. For phonetic coherence, Chen et al. [74] proposed a synchronization evaluation metric, the Lip-Reading Similarity Distance (LRSD), which can evaluate semantically synchronized lip movements.

Real-time performance. The length of time to generate the talking-head video is an important indicator for existing models. In the practical application of human–computer interaction, people are very sensitive to factors such as waiting time and video quality. Therefore, the model should generate the video as quickly as possible without sacrificing visual quality or the coherence of audio-visual semantics. NVIDIA [10] uses a deep neural network to map low-dimensional speech waveform data to high-dimensional facial 3D vertex coordinates and uses traditional motion capture technology to obtain high-quality video animation data to train the model.

Delayed Talking-Head Synthesis Model. Ye et al. [6] presented a novel, fully convolutional network with DCKs for the multimodal task of audio-driven talking-face video generation. Due to the simple yet effective network architecture and the video pre-processing, there is a significant improvement in the real-time performance of talking-head generation. Lu et al. [76] present a live system that generates personalized photorealistic talking-head animation only driven by audio signals at over 30 fps. However, many studies ignore real-time performance, and we should consider it as a critical evaluation metric in the future.

Human–computer interaction is a method for the future development of virtual humans. Unlike one-way information output, digital human needs to have multimodal information such as natural language, facial expression, and natural human-like gestures. Meanwhile, it also needs to be able to feedback on high-quality video in a short time after a given voice request [33,77].

However, in generating high-quality and low-latency digital human video, many researchers do not take real time as the evaluation index of the model. Hence, many models generate videos too slowly to cover the application requirements. For example, to generate 256 × 256 resolution facial video without background, ATVGnet [17] takes 1.07 s, You Said That [35] takes 14.13 s, X2Face [78] takes 15.10 s, DAVS [16] takes 17.05 s, and 1280 × 720 resolution video with background takes longer. Although it only takes 3.83 s for Wav2lip [22] to synthesize a video with a background, the definition of the lower part of the face is lower than that of other areas [6].

Many studies have attempted to establish a new evaluation benchmark and proposed more than a dozen evaluation metrics for virtual human video generation. Therefore, the existing evaluation metrics for virtual human video generation are not uniform. In addition to objective indicators, there are also subjective indicators, such as user research.

5. Future Directions

Dataset construction and methods for learning with fewer samples. A high-quality dataset is beneficial for the model to generate realistic, vivid, and human-like videos of talking heads. Existing open-source datasets are mainly composed of wild videos, and some are used for visual speech recognition tasks. In addition, a limitation of current methods is that deep-learning-based talking-head video generation methods mainly depend on labeled data. Recently, some work has begun to explore other effective learning paradigms, such as knowledge distillation and few-shot learning, to study the value of talking-head video generation tasks. At the same time, some researchers have begun to build high-quality visual-speech datasets with hidden features such as semantic and emotional annotations.

The realistic talking-head video generation with spontaneous movement. Humans are sensitive to any motion changes in synthetic videos, and they unconsciously pay attention to lip, eye, eyebrow, and spontaneous talking-head movements. Lip movement with audiovisual consistency is an indispensable part of the talking-head video generation, and implicit features, such as eye and head movements and emotional features, can rarely be inferred from audio. Recently, based on the study of lip movement, many works have begun to explore the application of implicit features, such as eye blinking and head pose, in the generation of talking-head videos. Introducing these implicit features in the research can make the video more realistic. Especially in the human–computer dialogue system, the speech synthesized by the TTS module is not as rich as the information contained in the wild audio.

Talking-head video generation detection. The development of talking-head video generation technology has greatly threatened social development. Misusing talking-head video generation technology and pictures or videos of people may reduce the production cost of false information, facilitate its dissemination, and cause severe ethical and legal issues, especially for celebrities or politicians. Talking-head video generation, fake video recognition, and detection are natural symbiotic tasks. At the same time, the natural and realistic content output by the talking-head video generation model has brought great difficulties and challenges to related forensic work, which has attracted the attention of many researchers. However, many existing methods focus on improving the model's performance, thus ignoring the problem of poor model interpretability. In addition, most of the methods are only optimized on a fixed data set, and the effect on wild data is negative. Interpretable and robust video generation detection of talking heads is important in accelerating technology development and preventing technology abuse.

Multi-person talking-head video generation. In the speech recognition task, the ASR model can recognize the number of speakers according to the difference in the voiceprint of the input voice and divide the speaker and the speech content. Multi-person talking-head video generation, mapping a piece of speech with different voiceprints to facial key point information of different talking heads, is a challenging task. It can be applied to many scenarios in life, such as the news connection when broadcasting news. However, it may not be optimal to transfer the single talking-head video generation method to multiple talking-head video generation tasks. Recently, the task has begun to attract the attention of researchers. Considering that the real-time problem of the talking-head generation model based on deep learning cannot be solved, there is still a lot of research space in this area. Below we provide some ideas and discuss potential approaches to address the poor real-time performance of talking-head generation models.

CG-based talking-head generation method. With the introduction of the metaverse concept, computational graphics (CG) companies used to create virtual characters in games, movies, and other scenes have launched virtual human generation programs. Currently, programs that can be used to create virtual human talking heads, including Audio2Face (https://www.nvidia.cn/omniverse/apps/audio2face/, accessed on 30 December 2022.) in Nvidia Omniverse, Meta-Human Creator (https://www.unrealengine.com/zh-CN/metahuman, accessed on 30 December 2022.) in Epic's Unreal Engine, digital humans (https://unity.com/demos/enemies#digital-humans, accessed on 30 December 2022.) in Unity 3D, and 3D Engine (https://developer.huawei.com/consumer/cn/hms/3d-engine/, accessed on 30 December 2022.) in Huawei's HMS Core. Now, the real-time rendering of the virtual human talking-head based on the CG program has a way of combining Audio2Face and Meta-Human Creator to output the rendered video. Meanwhile, some researchers have learned speech, lip, and expression parameters by letting the model learn. It combines the action generation algorithm of the virtual human with the 3D Engine and outputs the rendered cartoon image video in real time, such as Huawei's sign language digitizers and pose-guided generation based on 3D human meshes [79]. For the depth generation of the talking head, combining the deep learning method with the CG program ensures the real-time performance of video rendering. Although it has tremendous application potential, the high expense is a disadvantage with the output of the virtual human talking-head video. Speech-to-animation (S2A) technology is a method that automatically estimates synchronized facial animation parameters from given speech and generates the final animated avatar with these predicted parameters based on a rendering engine such as Unreal Engine 4 (UE4). Based on S2A, Chen et al. [31] combined the MOE transformer to model the context, which improved the inference speed of the model.

A method based on NeRF (Neural Radiance Fields) rendering. In the field of computer vision, the use of deep neural networks to encode objects and scenes is a new research direction. NeRF is an implicit neural representation that can render sharp photos of any viewing angle from images from multiple angles. Among them, AD-NeRF introduces NeRF to talking-head video generation. Although the native NeRF algorithm's slow rendering speed prevents it from generating talking-head videos in real time, many researchers have proposed numerous methods to increase NeRF's rendering speed [64–66,80]. For example, DONeRF Can render 20 frames per second on a single GPU, and Plenoctrees [66] is over 3000 times faster than traditional NeRFs.

A method for fusing speech recognition (ASR) and computer vision. With the continuous increase of input speech data, Streaming ASR outputs the text results of speech recognition in real time. Among them, the development of streaming decoders CTC [81], RNN-T [82], and LAC [83] has promoted the rapid development of Streaming ASR. In contrast, in the computer vision field of deep generation, no model can output talking-head video in real time. Therefore, in the real-time talking-head video generation study, the streaming decoder in the ASR field can be introduced into the talking-head video generation model to reduce the real-time rate of video generation. Among them, the real-time rate (RTF) is the ratio between the model processing time and the audio. For example, it takes 6 s to process 3 s of audio, RTF = 6 s/3 s = 2. Since the model is modeling historical input, the historical input will continue to grow over time, doubling the model's computational load, and the RTF will also increase accordingly. If RTF > 1.0, the model is too late to process the audio buffer. Therefore, it is possible to achieve real-time streaming output by reducing the RTF of the video generated by the talking head to less than 1.0.

6. Conclusions

This paper presents a systematic framework for multimodal human–computer interaction, which provides a new idea for applying talking-head generation models. It reviews talking-head generation models based on deep learning, including datasets, evaluation protocols, representative methods, etc. We analyze the strengths and weaknesses of representative methods and their potential connections. Thanks to the amazing development of deep learning, we have witnessed the rapid development of talking-head video models, from generating low-resolution and coarse images to high-resolution, detailed, and realistic images. However, talking-head video generation methods' real-time performance still needs improvement.

It is impossible to dismiss the possibility that malevolent activities, such as fraud, defamation, and malicious dissemination, may be carried out using the virtual human voice head synthesis technology. We vehemently oppose misusing this technology. This essay primarily reviews the study done in this area to further technological advancement and improve people's quality of life.

Author Contributions: Conceptualization, J.C. and W.S.; writing–original draft preparation, R.Z. and W.S.; writing–review and editing, J.C. and L.S.; visualization, R.Z. and Q.H.; supervision, L.S. and J.L.; All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key Research and Development Program of China (No. 2020AAA0108700, No. 2022YFC3302103), the Fundamental Research Funds for the Central Universities (CUC220C004), and the State Key Laboratory of Media Convergence Production Technology and Systems (SKLMCPTS2020012).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Garrido, P.; Valgaerts, L.; Sarmadi, H.; Steiner, I.; Varanasi, K.; Perez, P.; Theobalt, C. Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. In *Computer Graphics Forum*; Wiley Online Library: Hoboken, NJ, USA, 2015; Volume 34, pp. 193–204.
2. Garrido, P.; Valgaerts, L.; Rehmsen, O.; Thormahlen, T.; Perez, P.; Theobalt, C. Automatic face reenactment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 23–28 June 2014; pp. 4217–4224.
3. Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; Nießner, M. Face2face: Real-time face capture and reenactment of rgb videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2387–2395.
4. Bregler, C.; Covell, M.; Slaney, M. Video rewrite: Driving visual speech with audio. In Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, Los Angeles, CA, USA, 3–8 August 1997; pp. 353–360.
5. Xie, L.; Liu, Z.Q. Realistic mouth-synching for speech-driven talking face using articulatory modelling. *IEEE Trans. Multimed.* **2007**, *9*, 500–510.
6. Ye, Z.; Xia, M.; Yi, R.; Zhang, J.; Lai, Y.K.; Huang, X.; Zhang, G.; Liu, Y.J. Audio-driven talking face video generation with dynamic convolution kernels. *IEEE Trans. Multimed.* **2022**. [CrossRef]
7. Chung, J.S.; Zisserman, A. Out of time: Automated lip sync in the wild. In *Asian Conference on Computer Vision*; Springer: Cham, Switzerland, 2017; pp. 251–263.
8. Suwajanakorn, S.; Seitz, S.M.; Kemelmacher-Shlizerman, I. Synthesizing obama: Learning lip sync from audio. *ACM Trans. Graph. (ToG)* **2017**, *36*, 1–13. [CrossRef]
9. Chung, J.S.; Jamaludin, A.; Zisserman, A. You said that? *arXiv* **2017**, arXiv:1705.02966.
10. Karras, T.; Aila, T.; Laine, S.; Herva, A.; Lehtinen, J. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Trans. Graph. (ToG)* **2017**, *36*, 1–12. [CrossRef]
11. Kumar, R.; Sotelo, J.; Kumar, K.; de Brébisson, A.; Bengio, Y. Obamanet: Photo-realistic lip-sync from text. *arXiv* **2017**, arXiv:1801.01442.
12. Chen, L.; Li, Z.; Maddox, R.K.; Duan, Z.; Xu, C. Lip movements generation at a glance. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 520–535.
13. Kim, H.; Garrido, P.; Tewari, A.; Xu, W.; Thies, J.; Niessner, M.; Pérez, P.; Richardt, C.; Zollhöfer, M.; Theobalt, C. Deep video portraits. *ACM Trans. Graph. (ToG)* **2018**, *37*(4), 1–14. [CrossRef]
14. Vougioukas, K.; Petridis, S.; Pantic, M. End-to-end speech-driven facial animation with temporal gans. *arXiv* **2018**, arXiv:1805.09313.
15. Song, Y.; Zhu, J.; Li, D.; Wang, X.; Qi, H. Talking face generation by conditional recurrent adversarial network. *arXiv* **2018**, arXiv:1804.04786.
16. Zhou, H.; Liu, Y.; Liu, Z.; Luo, P.; Wang, X. Talking face generation by adversarially disentangled audio-visual representation. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 8–12 October 2019; 33, pp. 9299–9306.
17. Chen, L.; Maddox, R.K.; Duan, Z.; Xu, C. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In Proceedings of the IEEE/CVF Conference on Cmputer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 7832–7841.
18. Yu, L.; Yu, J.; Ling, Q. Mining audio, text and visual information for talking face generation. In Proceedings of the 2019 IEEE International Conference on Data Mining (ICDM), Beijing, China, 8–11 November 2019; IEEE: Manhattan, NY, USA, 2019; pp. 787–795.
19. Cudeiro, D.; Bolkart, T.; Laidlaw, C.; Ranjan, A.; Black, M.J. Capture, learning, and synthesis of 3D speaking styles. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 10101–10111.
20. Fried, O.; Tewari, A.; Zollhöfer, M.; Finkelstein, A.; Shechtman, E.; Goldman, D.B.; Genova, K.; Jin, Z.; Theobalt, C.; Agrawala, M. Text-based editing of talking-head video. *ACM Trans. Graph. (ToG)* **2019**, *38*, 1–14. [CrossRef]
21. Zhou, Y.; Han, X.; Shechtman, E.; Echevarria, J.; Kalogerakis, E.; Li, D. Makelttalk: Speaker-aware talking-head animation. *ACM Trans. Graph. (ToG)* **2020**, *39*, 1–15. [CrossRef]
22. Prajwal, K.R.; Mukhopadhyay, R.; Namboodiri, V.P.; Jawahar, C.V. A lip sync expert is all you need for speech to lip generation in the wild. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 484–492.
23. Thies, J.; Elgharib, M.; Tewari, A.; Theobalt, C.; Nießner, M. Neural voice puppetry: Audio-driven facial reenactment. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 716–731.
24. Chen, W.; Tan, X.; Xia, Y.; Qin, T.; Wang, Y.; Liu, T.Y. DualLip: A system for joint lip reading and generation. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1985–1993.
25. Guo, Y.; Chen, K.; Liang, S.; Liu, Y.J.; Bao, H.; Zhang, J. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021.
26. Li, L.; Wang, S.; Zhang, Z.; Ding, Y.; Zheng, Y.; Yu, X.; Fan, C. Write-a-speaker: Text-based emotional and rhythmic talking-head generation. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, Canada, 2–8 February 2021; pp. 1911–1920.

27. Liu, X.; Xu, Y.; Wu, Q.; Zhou, H.; Wu, W.; Zhou, B. Semantic-aware implicit neural audio-driven video portrait generation. *arXiv* **2022**, arXiv:2201.07786.
28. Fan, Y.; Lin, Z.; Saito, J.; Wang, W.; Komura, T. FaceFormer: Speech-Driven 3D Facial Animation with Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 18770–18780.
29. Yang, C.C.; Fan, W.C.; Yang, C.F.; Wang, Y.C.F. Cross-Modal Mutual Learning for Audio-Visual Speech Recognition and Manipulation. In Proceedings of the 36th AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 22 February–1 March 2022; Volume 22.
30. Zhang, S.; Yuan, J.; Liao, M.; Zhang, L. Text2video: Text-Driven Talking-Head Video Synthesis with Personalized Phoneme-Pose Dictionary. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Manhattan, NY, USA, 2022; pp. 2659–2663.
31. Chen, L.; Wu, Z.; Ling, J.; Li, R.; Tan, X.; Zhao, S. Transformer-S2A: Robust and Efficient Speech-to-Animation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Manhattan, NY, USA, 2022; pp. 7247–7251.
32. Zhang, H.; Yuan, T.; Chen, J.; Li, X.; Zheng, R.; Huang, Y.; Chen, X.; Gong, E.; Chen, Z.; Hu, X.; et al. PaddleSpeech: An Easy-to-Use All-in-One Speech Toolkit. *arXiv* **2022**, arXiv:2205.12007.
33. Shen, T.; Zuo, J.; Shi, F.; Zhang, J.; Jiang, L.; Chen, M.; Zhang, Z.; Zhang, W.; He, X.; Mei, T. ViDA-MAN: Visual Dialog with Digital Humans. In Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021; pp. 2789–2791.
34. Sheng, C.; Kuang, G.; Bai, L.; Hou, C.; Guo, Y.; Xu, X.; Pietikäinen, M.; Liu, L. Deep Learning for Visual Speech Analysis: A Survey. *arXiv* **2022**, arXiv:2205.10839.
35. Jamaludin, A.; Chung, J.S.; Zisserman, A. You said that? Synthesising talking faces from audio. *Int. J. Comput. Vis.* **2019**, *127*, 1767–1779. [CrossRef]
36. Li, T.; Bolkart, T.; Black, M.J.; Li, H.; Romero, J. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.* **2017**, *36*, 194-1. [CrossRef]
37. Richard, A.; Zollhöfer, M.; Wen, Y.; De la Torre, F.; Sheikh, Y. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 1173–1182.
38. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* **2021**, *65*, 99–106. [CrossRef]
39. Garbin, S.J.; Kowalski, M.; Johnson, M.; Shotton, J.; Valentin, J. Fastnerf: High-fidelity neural rendering at 200fps. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 14346–14355.
40. Müller, T.; Evans, A.; Schied, C.; Keller, A. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv* **2022**, arXiv:2201.05989. [CrossRef]
41. Li, R.; Tancik, M.; Kanazawa, A. NerfAcc: A General NeRF Acceleration Toolbo. *arXiv* **2022**, arXiv:2210.04847.
42. KR, P.; Mukhopadhyay, R.; Philip, J.; Jha, A.; Namboodiri, V.; Jawahar, C.V. Towards automatic face-to-face translation. In Proceedings of the 27th ACM international conference on multimedia, Nice, France, 21–25 October 2019; pp. 1428–1436.
43. Song, L.; Wu, W.; Qian, C.; He, R.; Loy, C.C. Everybody's talkin': Let me talk as you want. *IEEE Trans. Inf. Forensics Secur.* **2022**, *17*, 585–598. [CrossRef]
44. Wang, K.; Wu, Q.; Song, L.; Yang, Z.; Wu, W.; Qian, C.; He, R.; Qiao, Y.; Loy, C.C. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 700–717.
45. Jalalifar, S.A.; Hasani, H.; Aghajan, H. Speech-driven facial reenactment using conditional generative adversarial networks. *arXiv* **2018**, arXiv:1803.07461.
46. King, D.E. Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.* **2009**, *10*, 1755–1758.
47. Fan, B.; Wang, L.; Soong, F.K.; Xie, L. Photo-real talking head with deep bidirectional LSTM. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Manhattan, NY, USA, 2015; pp. 4884–4888.
48. Pham, H.X.; Cheung, S.; Pavlovic, V. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 80–88.
49. Tzirakis, P.; Papaioannou, A.; Lattas, A.; Tarasiou, M.; Schuller, B.; Zafeiriou, S. Synthesising 3D facial motion from "in-the-wild" speech. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*; IEEE: Manhattan, NY, USA, 2020; pp. 265–272.
50. Deng, Y.; Yang, J.; Xu, S.; Chen, D.; Jia, Y.; Tong, X. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 15–20 June 2019.
51. Zhang, Z.; Li, L.; Ding, Y.; Fan, C. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021. pp. 3661–3670.

52. Ji, X.; Zhou, H.; Wang, K.; Wu, W.; Loy, C.C.; Cao, X.; Xu, F. Audio-driven emotional video portraits. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14080–14089.
53. Si, S.; Wang, J.; Qu, X.; Cheng, N.; Wei, W.; Zhu, X.; Xiao, J. Speech2video: Cross-modal distillation for speech to video generation. *arXiv* **2021**, arXiv:2107.04806.
54. Sun, Y.; Zhou, H.; Liu, Z.; Koike, H. Speech2Talking-Face: Inferring and Driving a Face with Synchronized Audio-Visual Representation. *IJCAI* **2021**, *2*, 4.
55. Vougioukas, K.; Petridis, S.; Pantic, M. End-to-End Speech-Driven Realistic Facial Animation with Temporal GANs. In *CVPR Workshops*; CVF: New York, NY, USA, 2019; pp. 37–40.
56. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
57. Zhang, C.; Zhao, Y.; Huang, Y.; Zeng, M.; Ni, S.; Budagavi, M.; Guo, X. Facial: Synthesizing dynamic talking face with implicit attribute learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3867–3876.
58. Sadoughi, N.; Busso, C. Speech-driven expressive talking lips with conditional sequential generative adversarial networks. *IEEE Trans. Affect. Comput.* **2019**, *12*, 1031–1044. [CrossRef]
59. Eskimez, S.E.; Zhang, Y.; Duan, Z. Speech driven talking face generation from a single image and an emotion condition. *IEEE Trans. Multimed.* **2022**, *24*, 3480–3490. [CrossRef]
60. Zhou, H.; Sun, Y.; Wu, W.; Loy, C.C.; Wang, X.; Liu, Z. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4176–4186.
61. Ji, X.; Zhou, H.; Wang, K.; Wu, Q.; Wu, W.; Xu, F.; Cao, X. EAMM: One-Shot Emotional Talking Face via Audio-Based Emotion-Aware Motion Model. *arXiv* **2022**, arXiv:2205.15278.
62. Biswas, S.; Sinha, S.; Das, D.; Bhowmick, B. Realistic talking face animation with speech-induced head motion. In Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing, Jodhpur, India, 19–22 December 2021; pp. 1–9.
63. Waibel, A.; Behr, M.; Eyiokur, F.I.; Yaman, D.; Nguyen, T.N.; Mullov, C.; Demirtas, M.A.; Kantarcı, A.; Constantin, H.; Ekenel, H.K. Face-Dubbing++: Lip-Synchronous, Voice Preserving Translation of Videos. *arXiv* **2022**, arXiv:2206.04523.
64. Hong, Y.; Peng, B.; Xiao, H.; Liu, L.; Zhang, J. Headnerf: A real-time nerf-based parametric head model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 20374–20384.
65. Neff, T.; Stadlbauer, P.; Parger, M.; Kurz, A.; Mueller, J.H.; Chaitanya, C.R.A.; Kaplanyan, A.; Steinberger, M. DONeRF: Towards Real-Time Rendering of Compact Neural Radiance Fields using Depth Oracle Networks. In *Computer Graphics Forum*; Wiley-Blackwell: Hoboken, NJ, USA, 2021; Volume 40, pp. 45–59.
66. Yu, A.; Li, R.; Tancik, M.; Li, H.; Ng, R.; Kanazawa, A. Plenoctrees for real-time rendering of neural radiance fields. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 5752–5761.
67. Yao, S.; Zhong, R.; Yan, Y.; Zhai, G.; Yang, X. DFA-NeRF: Personalized Talking Head Generation via Disentangled Face Attributes Neural Rendering. *arXiv* **2022**, arXiv:2201.00791.
68. Cooke, M.; Barker, J.; Cunningham, S.; Shao, X. An audio-visual corpus for speech perception and automatic speech recognition. *J. Acoust. Soc. Am.* **2006**, *120*, 2421–2424. [CrossRef]
69. Chung, J.S.; Zisserman, A. Lip reading in the wild. In Asian conference on computer vision. In *Asian Conference on Computer Vision*; Springer: Cham, Switzerland, 2017; pp. 87–103.
70. Yang, S.; Zhang, Y.; Feng, D.; Yang, M.; Wang, C.; Xiao, J.; Long, K.; Shan, S.; Chen, X. LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*; IEEE: Manhattan, NY, USA, 2019; pp. 1–8.
71. Chung, J.S.; Nagrani, A.; Zisserman, A. Voxceleb2: Deep speaker recognition. *arXiv* **2018**, arXiv:1806.05622.
72. Chen, L.; Cui, G.; Kou, Z.; Zheng, H.; Xu, C. What comprises a good talking-head video generation? A survey and benchmar. *arXiv* **2020**, arXiv:2005.03201.
73. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef]
74. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 586–595.
75. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6626–6637.
76. Lu, Y.; Chai, J.; Cao, X. Live speech portraits: Real-time photorealistic talking-head animation. *ACM Trans. Graph. (TOG)* **2021**, *40*, 1–17. [CrossRef]
77. Zhen, R.; Song, W.; Cao, J. Research on the Application of Virtual Human Synthesis Technology in Human-Computer Interaction. In *2022 IEEE/ACIS 22nd International Conference on Computer and Information Science (ICIS)*; IEEE: Manhattan, NY, USA, 2022; pp. 199–204.

78. Wiles, O.; Koepke, A.; Zisserman, A. X2face: A network for controlling face generation using images, audio, and pose codes. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–17 September 2018; pp. 670–686.
79. Liu, L.; Xu, W.; Zollhoefer, M.; Kim, H.; Bernard, F.; Habermann, M.; Wang, W.; Theobalt, C. Neural animation and reenactment of human actor videos. *arXiv* **2018**, arXiv:1809.03658.
80. Martin-Brualla, R.; Radwan, N.; Sajjadi, M.S.; Barron, J.T.; Dosovitskiy, A.; Duckworth, D. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7210–7219.
81. Graves, A.; Fernández, S.; Gomez, F.; Schmidhuber, J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 369–376.
82. Graves, A. Sequence transduction with recurrent neural networks. *arXiv* **2012**, arXiv:1211.3711.
83. Chan, W.; Jaitly, N.; Le, Q.; Vinyals, O. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Manhattan, NY, USA, 2016; pp. 4960–4964.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.