information

Review

Audio-Driven Facial Animation with Deep Learning: A Survey

Diqiong Jiang [1,*], Jian Chang [1], Lihua You [1], Shaojun Bian [2], Robert Kosk [1] and Greg Maguire [3]

[1] National Centre for Computer Animation, Bournemouth University, Poole BH12 5BB, UK; jchang@bournemouth.ac.uk (J.C.); lyou@bournemouth.ac.uk (L.Y.); rkosk@bournemouth.ac.uk (R.K.)
[2] School of Creative and Digital Industries, Buckinghamshire New University, High Wycombe HP11 2JZ, UK; shaojun.bian@bnu.ac.uk
[3] Belfast School of Art, Ulster University, Belfast BT15 1ED, UK; g.maguire@ulster.ac.uk
* Correspondence: djiang@bournemouth.ac.uk

Abstract: Audio-driven facial animation is a rapidly evolving field that aims to generate realistic facial expressions and lip movements synchronized with a given audio input. This survey provides a comprehensive review of deep learning techniques applied to audio-driven facial animation, with a focus on both audio-driven facial image animation and audio-driven facial mesh animation. These approaches employ deep learning to map audio inputs directly onto 3D facial meshes or 2D images, enabling the creation of highly realistic and synchronized animations. This survey also explores evaluation metrics, available datasets, and the challenges that remain, such as disentangling lip synchronization and emotions, generalization across speakers, and dataset limitations. Lastly, we discuss future directions, including multi-modal integration, personalized models, and facial attribute modification in animations, all of which are critical for the continued development and application of this technology.

Keywords: deep learning; audio processing; talking head; face generation

Citation: Jiang, D.; Chang, J.; You, L.; Bian, S.; Kosk, R.; Maguire, G. Audio-Driven Facial Animation with Deep Learning: A Survey. *Information* **2024**, *15*, 675. https://doi.org/10.3390/info15110675

Academic Editors: Nikolaos Mitianoudis and Ilias Theodorakopoulos

Received: 30 September 2024
Revised: 20 October 2024
Accepted: 22 October 2024
Published: 28 October 2024

1. Introduction

Human speech is one of the most complex and expressive forms of communication, involving a combination of phonetic sounds, intonation, rhythm, and emotion. It conveys not only linguistic content but also the speaker's emotional state, intent, and identity. Human speech is inherently bimodal, incorporating both auditory and visual components [1]. Auditory speech comprises the sound waves produced during speaking, which can be captured and analyzed to convey linguistic content, emotional nuances, and emphasis through variations in tone and pitch. Visual speech refers to the visible movements made by the speaker during speech production. They include movements of the lips, tongue, jaw, cheeks, and other facial muscles that accompany the vocalization of sounds.

Audio-driven facial animation is a technique designed to generate realistic facial movements and expressions from an audio input, typically speech. At its core, it maps auditory speech to corresponding visual speech patterns. The goal of this technique is to automate the creation of facial animations that synchronize with the spoken content and reflect the emotional tone conveyed by the audio. By enabling the creation of highly realistic digital characters, avatars, and virtual humans, audio-driven facial animation enhances user immersion across various fields, including virtual reality (VR)/augmented reality (AR) [2], gaming [3] and human–computer interaction [2].

The importance of audio-driven facial animation lies in its ability to significantly reduce the labor-intensive and expensive process of manual animation, which traditionally requires skilled animators or complex motion capture systems. This is particularly valuable in industries where real-time interactions with digital characters are becoming more prevalent.

As discussed above, audio-driven facial animation involves mapping audio features to corresponding facial movements. This process can be broken down into two key components: audio feature extraction and facial animation generation based on features. Extracting meaningful features from speech [4–13], such as phonetic content and emotional cues, is fundamental for applications such as speech recognition, emotion analysis, and audio-driven facial animation. Traditional techniques like MFCCs (Mel-Frequency Cepstral Coefficients) [4] and LPC (Linear Predictive Coding) [5] focus on analyzing the frequency and power spectrum of speech. Modern methods employ deep neural networks to extract richer audio features. Pre-trained models like Wav2Vec 2.0 [13], Whisper [10], or VALL-E 2 [12] can capture both low-level phonetic details and high-level emotional or prosodic patterns from the speech. This allows for a more dynamic and nuanced representation of speech audio, including rhythm, pitch, and tone, which are crucial for generating realistic facial animation.

Once audio features are extracted, the next step is to generate the corresponding facial movements, including lip sync, jaw movements, eyebrow raises, and other expressions. The challenge at this step lies in translating these audio-derived features into 2D images or 3D meshes. Recent deep learning approaches, such as Generative Adversarial Networks (GANs) [14], Variational Autoencoders (VAEs) [15], diffusion models [16], and Transformers [17], have significantly advanced the generation of fluid, realistic, and emotionally rich facial animations. In terms of 3D mesh generation, techniques such as 3D Morphable Models (3DMMs) [18], Convolutional Neural Networks (CNNs), Graph Neural Networks (GNNs) [19], and Neural Radiance Fields (NeRFs) [20] play a crucial role. While the 3DMM provides a parametric space that ensures the controllability of facial deformations, CNNs, GNNs, and NeRFs capture complex visual details required for lifelike facial animations.

1.1. The Development of Audio-Driven Animation

The development of audio-driven facial animation is graphically shown in Figure 1. In the early stages of audio-driven facial animation, the approach was based on linguistics [21]. It focused on the visual counterparts of phonemes, termed visemes [22], and integrated complex coarticulation rules. Given the many-to-many mapping between phonemes and visemes, later research [23] used hidden Markov models (HMMs) based on facial dynamics observed in videos. Subsequent works improved trajectory sampling methods [24,25] and replaced HMMs with Gaussian process latent variable models [26,27], hidden semi-Markov models [28], or recurrent networks [29].

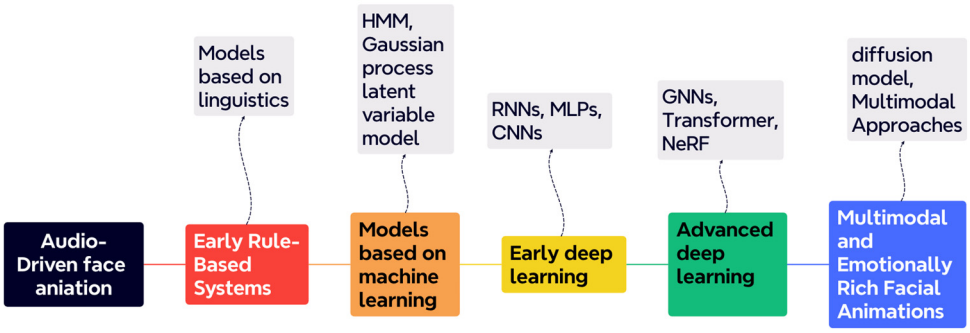

Figure 1. The development of audio-driven animation.

In the early era of deep learning, end-to-end network structures were typically used for facial regression, with the network architecture utilizing Recurrent Neural Networks (RNNs) [30] or fully connected layers [31]. For speech feature extraction, MFCCs (Mel-Frequency Cepstral Coefficients) and LPC (Linear Predictive Coding) were commonly employed. This was primarily due to the limited adoption of powerful audio processing and advanced frameworks like Transformers and GANs during that time.

As speech feature processing became more sophisticated and the capabilities of GANs were increasingly explored, the field has entered the advanced deep learning era. During this phase, the fidelity, resolution, and robustness of face generation based on speech have been significant improvements. The transition from simpler models to the adoption of GANs and advanced deep learning techniques greatly enhanced the generalization ability of these systems, allowing for more realistic and detailed facial animations driven by speech inputs.

Currently, researchers are focusing on enriching input data and improving algorithms to enhance expressive power. By integrating audio, video, and textual information, models can more accurately capture and express emotions and micro-expressions. Some algorithms also support secondary editing of the generated talking heads. Additionally, leveraging deep learning technologies such as CNNs and RNNs improves the detail and realism of facial animations. These advancements enable facial animations in applications like virtual assistants, gaming, and digital avatars to not only accurately reflect spoken content but also convey deeper emotional nuances, thereby enhancing user immersion and interaction experiences.

1.2. The Scope of This Survey

The main goal of this survey is to provide a comprehensive review of deep learning-based methods and datasets used in audio-driven facial animation. By examining both audio feature extraction and subsequent facial animation generation, this survey aims to highlight the strengths, limitations, and future directions of current methodologies in the field. The scope of this survey is shown in Figure 2 and briefly introduced below. Detailed reviews will be given in the following sections.

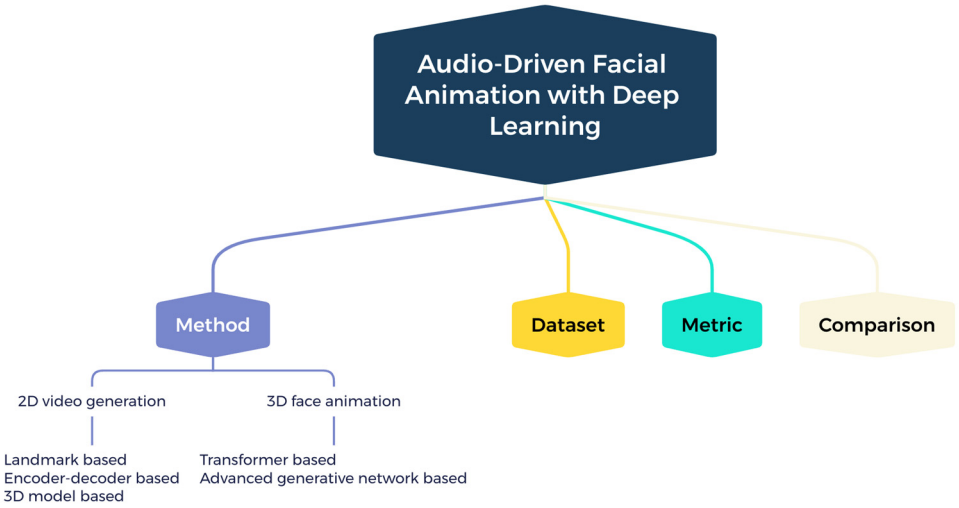

Figure 2. Scope of this survey.

Methods: This survey explores the methods used in both 2D video generation and 3D facial animation, classifies 2D video generation methods into landmark-based, encoder–decoder, and 3D model-based ones, and categorizes 3D face animation approaches into transformer-based and advanced generative network-based (diffusion model and NeRF) ones, with a focus on their application in audio-driven animation. It provides a thorough review of how deep learning techniques are employed to enhance the synchronization of facial animations with audio input, outlining advancements and challenges in both areas. Furthermore, the survey emphasizes the integration of emotional expression in these animations.

Datasets: the survey offers an in-depth analysis of datasets used for training and evaluating audio-driven facial animation models, with a focus on dataset diversity and emotional content.

Metrics and comparison: Metrics are essential for evaluating and comparing the performance of different methods in audio-driven facial animation. This survey outlines the key evaluation metrics currently in use and offers a comparison of leading deep learning techniques for both 2D video generation and 3D facial animation. The emphasis will be on evaluating the effectiveness of these techniques in producing realistic and emotionally expressive animations driven by audio input.

1.3. Differences with Existing Surveys

This survey distinguishes itself from prior works by offering a comprehensive and up-to-date analysis of the integration of emotion into audio-driven facial animation, particularly emphasizing advancements in audio-driven 3D facial animation. It explores the application of modern deep learning techniques, providing an in-depth examination of how these innovations are advancing audio-driven facial animation. Furthermore, it highlights the emerging trends that are pushing the boundaries of what is possible in facial animation.

Kammoun et al. [32] conducted a review of Generative Adversarial Networks (GANs) in the context of facial generation, covering their architectures, applications, semantics, and evaluation methodologies. Siddharth et al. [33] investigated text-guided image generation employing diffusion models, highlighting significant advancements in the synthesis of images based on textual descriptions. In contrast, this survey offers a broader perspective by covering all deep learning techniques relevant to audio-driven facial animation.

Liu et al. [34] provided a comprehensive review of audio-driven talking head synthesis. Tolosana et al. [35] and Mirsky et al. [36] examined audio-driven facial animation through the lens of deepfake technology, while Zhen et al. [37] approached the subject from a human–computer interaction perspective. However, their studies primarily focused on 2D video-based approaches and did not offer an analysis of 3D facial animation.

Sha et al. [38] and Gowda et al. [39] provided extensive reviews on talking head synthesis. Meng et al. [40] advanced the discourse by offering a detailed investigation of editing techniques and cutting-edge methodologies for incorporating diffusion models into talking head synthesis.

This survey emphasizes the generation of facial expressions and 3D face animation. We rigorously assess the application of cutting-edge technologies such as GANs and diffusion models, evaluating their current capabilities and outlining their potential future applications. Additionally, this work highlights the existing challenges within the field and proposes clear directions for future research trajectories.

2. Two-Dimensional Video and Three-Dimensional Facial Animation Generation Methods

2.1. Pipeline for Audio-Driven Facial Animation

Audio-driven facial animation aims to generate facial animations from audio input, with the goal of producing realistic and emotionally expressive facial movements that synchronize with spoken content. The pipeline outlines the problem formulation and task description for audio-driven facial animation, including the key challenges and objectives involved in the process.

The pipeline for audio-driven facial animation consists of three primary stages: audio feature extraction, talking head generation, and talking head editing. The first two stages are the core modules of audio-driven facial animation. The talking head editing stage serves as an optional extension module, offering additional opportunities for refinement and customization. Each stage is essential for ensuring that the final output is both visually coherent and dynamically expressive.

Figure 3 illustrates the framework for deep learning-based generation of audio-driven 2D video and 3D facial animation. In this framework, the speech encoder corresponds to the audio feature extraction module, the generative network aligns with the talking head generation module, and the controller encoder and feature fusion represent the talking head editing module. The following is an introduction to the functions of these three modules:

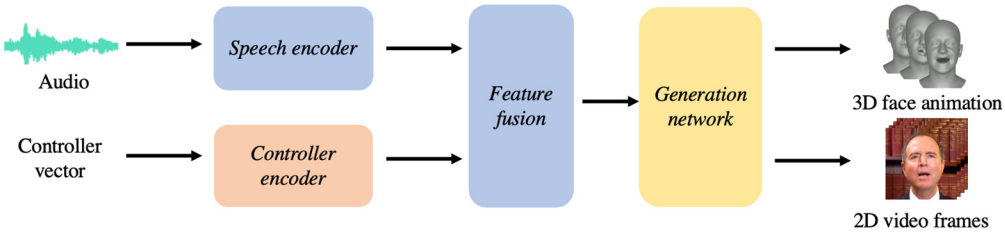

Figure 3. Graphical illustration of deep learning-based generation of audio-driven 2D video and 3D facial animation.

Audio feature extraction: the primary objective of audio feature extraction is to derive meaningful and informative features from the audio signal that correlate with facial movements and expressions.

Talking head generation: this stage generates facial animations that accurately represent the audio features, including lip movements, facial expressions, and synchronization with the speech.

Talking head editing: this stage improves the quality and customization of the generated facial animations, ensuring they meet specific esthetic and functional requirements.

2.2. Two-Dimensional Video Generation

Two-dimensional video generation from audio involves creating animated facial movements in a two-dimensional space. This section describes three primary methods used in this process: landmark-based methods, encoder—decoder methods, and 3D model-based methods.

2.2.1. Landmark-Based Methods

Landmark-based methods focus on tracking and manipulating key facial points (landmarks) to generate facial animations.

As illustrated in the Figure 4, landmark-based methods are typically divided into two modules: motion generation and texture generation. The first module, motion generation, converts the input audio into audio features, which are then used to deform the landmarks. The second module, texture generation, utilizes an image-to-image framework that takes the generated landmarks and facial images as input to produce audio-driven facial animation. The design of these two modules is central to these methods. By leveraging the position and movement of facial landmarks, these approaches effectively synchronize facial movement with audio input. This separation of lip sync and facial representation allows for improved handling of various facial characteristics, addressing the limitations of other methods that struggle to generate realistic animations from speech on unknown faces due to their poor generalization capabilities.

Suwajanakorn et al. [30] and Jalalifar et al. [41] introduce a method for synthesizing video from audio, focusing on the mouth region, but it lacks the ability to generalize to unseen individuals. Additionally, the approaches only regress key points around the mouth, even though speakers exhibit significant facial movement when talking. In contrast, subsequent works [42–45] extend this by regressing landmarks for the entire face using audio input. Their approach enhances the model's ability to generalize across different speakers and corresponding audio inputs.

Recent works have improved both network architecture and editability. For example, EchoMimic [46] not only generates portrait videos from either audio or facial landmarks individually but also allows control over facial landmarks to achieve diverse expressions and head movements. Additionally, Tan et al. [47] and Zhong et al. [48] replace the traditional image-to-image network structure with a diffusion model, resulting in more realistic generated animation.

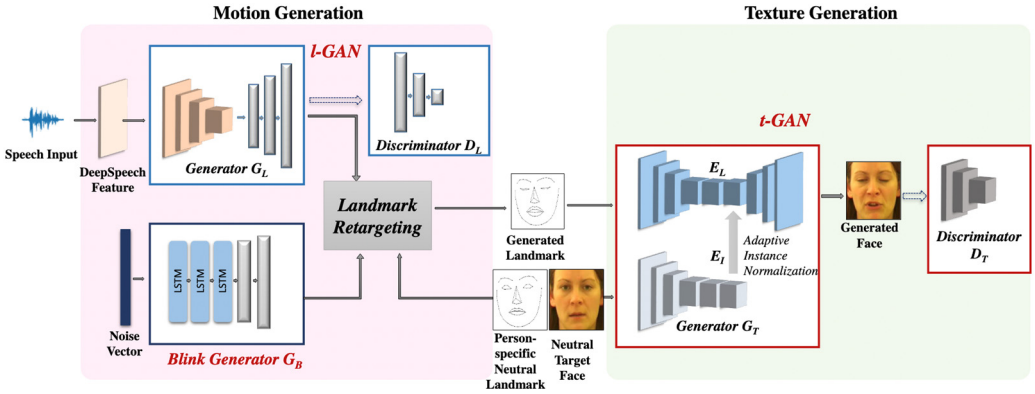

Figure 4. Graphical illustration of landmark-based methods.

2.2.2. Encoder–Decoder Methods

The encoder–decoder method can be divided into the following modules: audio encoding, identity encoding, audio–identity fusion, and face generation. The most notable difference from previous landmark-based methods is the introduction of the audio–identity fusion module, which replaces the use of landmarks. Recent methods have added an attribute encoding module to control additional facial animation properties, such as expression and pose, while the audio–identity fusion module has been enhanced to include feature vectors for other attributes.

The audio–identity fusion in early methods [49–51] simply concatenated features of the audio and identity encodings. In this architecture, the audio–identity fusion module plays a crucial role, combining audio information (such as tone and rhythm) with identity features (such as facial shape and expression) to generate dynamic facial animations that match the speaker's identity. Although the simple feature combination approach can produce initial results, it lacks expressiveness and detail in handling complex scenarios. Therefore, later research introduced more sophisticated fusion mechanisms.

X2Face [52] learns to map pixels from the target face to the generated frame. Chen et al. [53] fuse audio- and identity-based features using duplication and concatenation. They duplicate the identity feature along the frequency dimension at each time step. Some works [54–57] utilize RNNs to incorporate both image and audio features within the recurrent unit, effectively capturing temporal dependencies.

Zhu et al. [58] also present a method that connects the features of audio encoding and identity encoding, enhancing the correlation between these features through Attentional Audio–Visual Coherence Learning. Similarly, some works [59–62] train a powerful lip sync discrimination model to obtain a pre-trained lip sync expert, strengthening feature correlation. Their method enhances the generalization capability of audio-driven face animation, allowing for the generation of animations for arbitrary identities.

Recent works [63–66] regress motion vectors from audio features and use these motion vectors to control the mouth shapes and other attributes of the input video's face by warping the feature layers. For example, AniTalker [65] employs a multi-layer Conformer to regress motion vectors from audio features. Some approaches [67,68] directly embed the audio feature layers and identity features into the feature layers of the face generation network for fusion. Compared to earlier methods, these approaches allow pixel-level control, resulting in capturing more detailed facial movements.

Attribute Editing: Some papers [69–71] focus on attribute editing. For example, the Expression-Tailored Generative Adversarial Network (ET-GAN) [67] aims to generate expression-enriched talking face videos of arbitrary identities. The ET-GAN focuses on enhancing facial animations with expression variations through attribute editing. Pose-Controllable Audio–visual System (PC-AVS) [70] aims to achieve free pose control while driving static photos to speak with audio. Zhao et al. [63] and Mittal et al. [72] disentangle audio–visual representation, separating subject-related and speech-related information.

2.2.3. Three-Dimensional Model-Based Methods

Three-dimensional model-based methods [73–80] for 2D video generation involve creating and manipulating detailed 3D facial models, mapping audio features to these models, and projecting the resulting animations onto 2D planes. By leveraging 3D Morphable Models and advanced rendering techniques, these methods produce high-quality and realistic facial animations synchronized with audio. Despite their advantages in detail and flexibility, they face challenges related to computational complexity and real-time performance.

In 3D model-based methods, 3D facial models are used to replace facial landmarks and feature vectors as inputs to the generation network. These 3D models can be broadly categorized into two types: 3D Morphable Models (3DMMs) and non-3D Morphable Models. Compared to other methods, such as landmark-based and encoder–decoder approaches, 3D model-based methods inherently encode richer semantic information about facial geometry and expressions. This allows for more precise control over facial attributes, such as identity, expression, and pose, enabling more detailed and accurate facial animation generation.

Yi et al. [75] propose a method that directly extracts facial expression and pose parameters from audio. By replacing the 3D Morphable Model (3DMM) parameters of the target image, they generate facial expressions and poses corresponding to speech. Zhang et al. [76] and Zhang et al. [77] expand the parametric model of the 3D face by introducing more detailed controls, particularly on eye movements (e.g., blinking) and lip synchronization, thereby enhancing the naturalness and realism of the generated facial animations. Additionally, other works [75,79,80] utilize facial keypoints as control mechanisms, allowing for more fine-grained adjustments and control over facial expression details such as those involving the lips, eyes, and eyebrows. These methods have significantly improved the precision and detail representation in audio-driven 3D face generation tasks.

2.3. Three-Dimensional Face Animation

Early works [31,75,81–94] primarily utilized Recurrent Neural Networks (RNNs) or fully connected layers to regress the 3D Morphable Model parameters or vertex coordinates of the face. These methods take audio or other signals as input, predicting the 3D Morphable Model parameters frame by frame or directly regressing the 3D coordinates of each vertex to generate dynamic facial expressions synchronized with speech. Although these approaches rely on the network to capture the relationship between facial movements and audio signals over time, they may have limitations in handling complex expressions and faces of different identities. Nevertheless, these early methods laid the foundation for subsequent research in 3D face generation, driving the development of more structured and modular generation techniques in later studies. Recent works in audio-driven 3D face animation primarily utilize Transformer and diffusion models. In the following subsections, we will explore these two aspects in detail.

2.3.1. Transformer-Based Methods

Transformer-based methods [95–112] for audio-driven 3D face animation utilize self-attention mechanisms, sequence-to-sequence architectures, and multimodal integration to handle complex relationships and generate high-quality animations. These methods effectively capture and translate audio features into dynamic and realistic facial movements. However, they also face challenges concerning computational resources, data requirements, and model complexity.

FaceFormer [96] encodes long-term audio context and the history of face motions to autoregressively predict a sequence of animated 3D face meshes. It achieves highly realistic and temporally stable animation of the whole face including both the upper face and the lower face. Imitator [95] enhances the generation of facial animations by integrating identity-specific information. EmoTalk effectively disentangles emotions from spoken content by introducing the Emotion Disentangling Encoder (EDE) and an emotion-guided feature fusion decoder. While EmoTalk [100] allows for control over the intensity of

emotional expressions, it lacks the ability to differentiate between specific emotion types. In contrast, EMOTE [101] systematically integrates the effects of both emotion and speech on the resulting 3D animation through a novel emotion-content disentanglement mechanism, enabling more nuanced emotional representations.

2.3.2. Advanced Generative Network-Based Methods

Diffusion model-based or NeRF-based methods [113–123] for audio-driven 3D face animation offer a promising approach to generating high-quality and realistic facial animations from audio inputs. By leveraging the iterative denoising process of diffusion models, these methods can capture complex data distributions and produce detailed animations. However, they also face challenges related to computational demands, data requirements, and model complexity.

Most diffusion model-based methods or NeRF-based methods are also based on Transformer-based methods, but they replace the generation network with a diffusion model. Compared to Transformer-based methods that generate 3DMM parameters, diffusion models offer stronger facial expression capabilities and finer detail generation.

3. Comparison

In the field of audio-driven face animation, various methods have been developed to create realistic and expressive facial animations from audio input. As discussed above, these methods can be broadly categorized into five categories: landmark-based methods, encoder–decoder methods, 3D model-based methods, transformer-based methods, and advanced generative network-based methods. Each approach has its strengths, limitations, and unique characteristics, which are essential to consider when evaluating their effectiveness. This section provides a comparative overview of these methods.

Among the above five categories, landmark-based approaches are simpler and less computationally demanding but may lack detail and expressiveness. Encoder–decoder methods offer flexibility and detail but require substantial data and computational resources. Three-dimensional model-based methods excel in realism and customizability but involve complex integration and high data demands. Transformer-based methods provide strong performance and handle long-range dependencies effectively, though they are computationally intensive and complex. Advanced generative network-based methods stand out for their ability to generate high-quality animations but also face challenges related to training complexity and data requirements. The choice of method depends on the specific requirements of the application, including the desired level of detail, computational resources, and available data.

In this section, we will first introduce the datasets, then provide an overview of the evaluation metrics, and finally present the results from some of the current state-of-the-art methods.

3.1. Datasets

Various datasets used for deep learning-based 2D video and 3D facial animation generation are summarized in Table 1. Below, we briefly introduce all of them.

Table 1. Summary of datasets.

Dataset Name	Year	Subjects	Utterance	Environment	Language	Emotions	Emotion Level	Views	Facial Mesh	Link
GRID [124]	2006	54	5400	Lab	English	Neutral	-	2 views	No	[125]
CREMA-D [126]	2014	91	7442	Lab	English	6	3	front	No	[126]
BIWI [127]	2014	14	1109	Lab	English	2	-	-	Yes	[128]
TCD-TIMIT [129]	2015	62	6913	Lab	English	Neutral	-	2 views	No	[130]
MODALITY [131–133]	2015	35	5880	Lab	English	Neutral	-	-	No	[134]
MSP-IMPROV [135]	2016	12	8438	Lab	English	4	-	front	No	[135]
LRW [136]	2016	-	∼539 K	Wild	English	-	-	-	No	[137]
LRS [138]	2017	-	∼118 k	Wild	English	-	-	-	No	[139]
MV-LRS [140]	2018	-	∼500 k	Wild	English	-	-	-	No	[139]

Table 1. *Cont.*

Dataset Name	Year	Subjects	Utterance	Environment	Language	Emotions	Emotion Level	Views	Facial Mesh	Link
LRS2-BBC [141]	2018	∼62.8 k	∼144.5 k	Wild	English	-	-	-	No	[142]
LRS3-TED [143]	2018	9.5 k+	∼165 k+	Wild	English	-	-	-	No	[144]
VoxCeleb [145–147]	2018	7 k+	1 M+	Wild	English	-	-	-	No	[148]
RAVDESS [149]	2018	24	7356	Lab	English	8	2	front	No	[150]
MELD [151]	2018	-	13 k	Wild	English	6	-	-	No	[152]
VOCASET [86]	2019	12	480	Lab	English	-	-	-	Yes	[153]
HDTF [77]	2020	300+	10 k+	Wild	English	-	-	-	No	[154]
MEAD [155]	2020	60	281.4 k	Lab	English	8	3	7	No	[155]
CelebV-HQ [156]	2022	15,653	35,666	Wild	English	8	-	-	No	[156]
Multiface [157]	2022	13	299 k	Lab	English	118	18	150	Yes	[158]
MMFace4D [159]	2023	431	35,904	Lab	Chinese	7	-	Front	Yes	[160]
MultiTalk [161]	2024	-	294 k	Wild	20 Languages	-	-	-	No	[162]

3.1.1. Audio–Visual Speech Data for Face Animation

These datasets are primarily designed for synchronizing facial animations with speech and can be used to generate dynamic 3D facial movements from audio input.

- GRID: provides sentence-level audio–visual data ideal for animating lip movements in speech-driven face animation.
- TCD-TIMIT: contains synchronized audio–visual recordings for audio-driven speech animation tasks.
- VOCASET: specifically built for 3D facial mesh generation from speech audio, making it highly suitable for audio-driven face animation.
- LRS: a sentence-level continuous speech dataset for speech-driven facial animation, particularly for lip movement.
- LRS2-BBC: offers continuous speech data from BBC shows, useful for high-quality lip sync and facial motion generation.
- LRS3-TED: provides diverse speech data from TED talks, valuable for training models on varied speakers and expressions in audio-driven face animation.
- MultiTalk: a multilingual audiovisual dataset designed to enhance 3D talking head generation across multiple languages.

3.1.2. Emotional Expression and Speech Synthesis

These datasets focus on emotion-driven face animation, where the facial expressions are influenced by both speech content and emotional cues.

- CREMA-D: a multimodal emotional dataset useful for generating facial animations that incorporate emotional expressions along with speech.
- MSP-IMPROV: ideal for creating expressive facial animations that respond to both emotional and speech input.
- RAVDESS: combines emotional speech and facial expressions, facilitating models that generate emotional facial animations from audio.
- MEAD: a large-scale dataset of emotional talking faces, perfect for generating facial expressions and lip syncing with emotional variations in speech.

3.1.3. High-Resolution and Detailed Facial Data for 3D Animation

These datasets are focused on high-resolution 3D facial data, essential for generating detailed and realistic face animations from audio signals.

- HDTF: high-definition 3D facial sequences useful for creating fine-grained facial animations driven by speech.
- Multiface: captures facial landmarks and expressions, valuable for generating precise facial animations synchronized with audio.
- MMFace4D: a 4D facial dataset that can be used for generating dynamic 3D facial expressions driven by speech audio.

3.1.4. Multimodal Data with Speech for Face Animation

This dataset offers multimodal data, including both audio and visual signals, for complex animation tasks that require synchronizing facial movements with speech and other modalities.

- MODALIT: a multimodal interaction dataset that can be adapted for audio-driven facial animation, particularly for tasks involving synchronized speech and expressions.

3.1.5. General and Speaker Recognition Datasets with Potential for Face Animation

While not explicitly designed for facial animation, this dataset contains rich audiovisual data that can be repurposed for tasks like speaker-dependent facial animations.

- VoxCeleb: a large-scale audiovisual dataset that can be adapted for generating personalized facial animations from speaker-specific audio.

Laboratory Dataset vs. Wild Dataset

Laboratory datasets, such as CREMA-D and RAVDESS, typically consist of thousands to tens of thousands of samples collected in controlled environments with consistent lighting and camera angles. This setup enables precise analyses of facial expressions and lip synchronization, with datasets like MEAD providing detailed labels for emotions and viewing angles, making them better suited for specific tasks. In contrast, wild datasets like VoxCeleb contain over a million samples, offering diversity and realism but introducing variability in data quality, which can complicate model training and evaluation.

Challenges and Considerations

Data Quality: high-quality and diverse datasets are essential for developing robust models. Poor data quality or limited diversity can lead to overfitting and reduced generalization.

Ethical and Privacy Concerns: when using personal data, it is crucial to address ethical and privacy issues, ensuring that data collection and usage comply with relevant regulations and respect individuals' rights.

Dataset Size and Diversity: larger and more diverse datasets generally lead to better model performance, but they also require more resources to collect, manage, and process.

Synthetic Data: in cases where real data are scarce, synthetic datasets generated using 3D modeling or simulation techniques can be used to augment training data.

Datasets are foundational to the development of audio-driven face animation technologies, providing the necessary data for training and evaluating models. The choice of dataset—whether audio–video combined, 3D facial animations, or emotion-specific—affects the performance and applicability of the resulting models. Addressing challenges related to data quality, ethical considerations, and dataset diversity is crucial for advancing the field and ensuring that models are both effective and ethically sound.

3.2. Evaluation of Audio-Driven Facial Animation Methods

Metrics are essential for evaluating the performance and quality of audio-driven face animation systems. They help quantify how well a model generates realistic and expressive facial animations from audio input. Below is a comprehensive overview of the key metrics used in this field:

(A) Quantitative Metrics

The performance of deep learning-based 2D video and 3D facial animation systems is typically evaluated using a range of quantitative metrics, as summarized in Table 2. Each metric focuses on different aspects of the generated animation's accuracy, realism, or perceptual quality. MSE and PSNR assess pixel wise error and Signal-To-Noise Ratio, with lower MSE and higher PSNR values indicating higher quality outputs. SSIM measures structural similarity between images. Perceptual metrics, such as LPIPS, FID, and CPBD, evaluate how human viewers perceive the similarity of generated animations to ground-truth images and the sharpness and clarity of frames.

Table 2. Summary of metrics.

Metric	Compare Level	Focus on	Typical Range/Values
MSE	Pixel wise	Pixel wise error	0 (perfect) to ∞
PSNR	Pixel wise	Signal-to-Noise	20 dB (poor) to 40+ dB (excellent)
SSIM [163]	Perception	Structural details	−1 (poor) to 1 (perfect)
CPBD [164]	Perception	Sharpness and clarity	0 (excellent) to 1 (poor)
LPIPS [165]	Perception	Perceptual similarity	0 (perfect) to ∞
FID [166]	Perception	Distribution similarity	0 (perfect) to ∞
LMD [53]	Pixel wise	Landmark position error	Varies based on dataset, ideally < 5 pixels
LRA [57]	Perception	Lip synchronization	0 (poor) to 1 (perfect)
WER [167]	Perception	Word errors	0 (perfect) to 1 (poor)
EAR [168]	Pixel wise	Openness of the eyes	0 (closed) to 1 (fully open)
ESD [169]	Perception	Emotion similarity	0 (different) to 1 (same)
LVE [89]	Vertex wise	Lip vertices error	Varies based on dataset, ideally < 2 mm
FDD [99]	Perception	Facial dynamics	0 (no motion) to ∞ (extreme motion)
BA [170]	Temporal wise	Temporal difference	0 (poor) to 1 (perfect)
LSE-D [59]	Temporal wise	Lip synchronization	0 (perfect) to ∞, ideally < 2
LSE-C [59]	Temporal wise	Lip synchronization	Varies based on dataset, ideally > 8

Temporal metrics like LSE-D and LSE-C focus on lip synchronization accuracy, while LRA and EAR measure aspects like synchronization of facial movements and eye blink detection. Other metrics like LVE capture vertex wise lip error, crucial for 3D mesh accuracy. By combining these metrics, researchers are able to conduct a more comprehensive evaluation of animation fidelity across various facets such as motion, appearance, synchronization, and emotion conveyance. Below, we briefly introduce all of them.

3.2.1. Pixel Wise Metrics (Direct Comparison of Pixel Values)

These metrics measure errors or differences at the pixel level, directly comparing the generated image to a reference.

- MSE (Mean Squared Error): Measures the average squared difference between the predicted and ground truth facial animations. Lower MSE values indicate better accuracy in reproducing facial movements.
- PSNR (Peak Signal-to-Noise Ratio): Evaluates the quality of generated facial animations by comparing them to reference animations. Higher PSNR values indicate better visual fidelity and less distortion.
- LMD (Landmark Distance Error): LMD quantifies the accuracy of lip movement generation by calculating the distance between predicted and actual landmark positions on the lips during animation. This metric is vital for assessing the fidelity of facial animation systems that generate lip movements, ensuring that they closely match the intended expressions.
- EAR (Eye Aspect Ratio): EAR is a quantitative measure used to assess the openness of the eyes by analyzing the geometric relationships between specific facial landmarks around the eyes. It is primarily utilized in computer vision and facial recognition applications to detect eye blinks and monitor eye movement.

3.2.2. Perception-Based Metrics (Visual Quality and Realism)

These metrics evaluate the perceptual quality, realism, and visual similarity of generated outputs, focusing on how humans would perceive the result.

- SSIM (Structural Similarity Index): Assesses the similarity between generated and real facial animations by comparing structural details such as luminance, contrast, and texture. SSIM provides a more perceptually relevant measure of quality than MSE or PSNR.
- CPBD (Cumulative Probability of Blur Detection): Cumulative Probability of Blur Detection (CPBD) is an effective metric for assessing the sharpness and clarity of

images. By leveraging a probabilistic model that evaluates edge sharpness, CPBD provides a comprehensive measure of image quality that aligns well with human visual perception.

- LPIPS (Learned Perceptual Image Patch Similarity): LPIPS (Learned Perceptual Image Patch Similarity) is a metric designed to evaluate the perceptual similarity between images, focusing on how humans perceive differences in visual content. Unlike traditional metrics such as MSE or PSNR, which rely solely on pixel wise comparisons, LPIPS leverages deep learning to assess image quality in a manner that aligns more closely with human perception.
- FID (Fréchet Inception Distance): FID measures the distance between the distributions of generated images from a Generative Adversarial Network (GAN) and real images, based on feature representations extracted from a pre-trained Inception network. It is commonly used to evaluate the quality of GANs, particularly under a two-time-scale update rule, as it helps determine convergence towards a local Nash equilibrium in the training process.
- LRA (Lip-Reading Accuracy): Conventional metrics such as PSNR, SSIM, and LMD are inadequate for accurately evaluating the correctness of generated lip movements. To enhance the assessment of lip synchronization, LRA (Lip-Reading Accuracy) is analyzed through a cutting-edge deep lip-reading model trained on real speech videos. This method has demonstrated effectiveness in providing a more precise evaluation of lip synchronization quality.
- WER (Word Error Rate): WER is calculated by comparing the predicted words generated from the audio input against a reference transcription of the spoken content. Specifically, WER quantifies the number of errors by assessing the minimum number of word insertions, substitutions, and deletions required to align the predicted transcription with the ground truth.
- ESD (Emotion Similarity Distance): ESD is a metric designed to quantify the similarity of emotional features extracted from video data. ESD is grounded in the concept of cosine similarity, which measures the cosine of the angle between two non-zero vectors in a high-dimensional space.
- FDD (Upper-Face Dynamics Deviation): FDD measures the variation in facial dynamics (the changes in motion over time) of the upper face by comparing the generated motion to the ground truth motion. Upper-face movements are less directly tied to speech, meaning they exhibit more subtle and complex dynamics. These dynamics are influenced by emotion, intention, and personal speaking style. Therefore, FDD helps assess whether the generated facial motions capture this complexity by comparing how well the motion variations match the ground truth.

3.2.3. Vertex Wise Metrics (Geometric Comparisons at Mesh Level)

- LVE (Lip Vertex Error): LVE refers to the difference between the actual and predicted positions of vertices on the lips in a 3D facial model during animation or synchronization tasks.

3.2.4. Temporal Wise Metrics (Time-Based Evaluation)

- BA (Beat Align Score): a metric commonly used in the context of evaluating the alignment between audio signals (such as speech or music) and their corresponding visual representations.
- LSE-D (Lip Sync Error—Distance): LSE-D evaluates lip synchronization by measuring the distance-based error between predicted lip movements and the ground truth over time. It focuses on the degree of mismatch in lip movement positions relative to the audio, assessing how well the generated animation aligns with the timing of speech.
- LSE-C (Lip Sync Error—Confidence): LSE-C assesses lip synchronization but focuses on the confidence of the synchronization rather than just positional differences. It eval-

uates how confidently the system can predict lip movements based on the input audio, aiming to capture the reliability of the synchronization across the animation sequence.

(B) Qualitative Metrics

In audio-driven face animation, qualitative evaluation plays a significant role in assessing the overall performance of the generated animations, focusing on aspects such as realism, naturalness, and synchronization. Unlike quantitative metrics that provide numerical evaluations, qualitative testing relies on subjective human judgment and expert evaluations. Below are the common qualitative evaluation methods used in audio-driven face animation:

(I) Visual Realism and Naturalness: evaluates how visually realistic and natural the generated facial animations appear to human observers.

- Human Subjective Evaluation: Human evaluators are asked to rate the realism, naturalness, and smoothness of the generated animations. This is often carried out through surveys, where participants watch animated clips and rate them based on various criteria (e.g., 1 to 5 scale).
- Expert Evaluation: Professionals in animation, gaming, or film industries may be asked to evaluate the quality of the facial animations based on their experience. This often focuses on the accuracy of facial expressions, lip sync, and overall animation quality.
- Comparative Realism: evaluators compare the generated face animations with real video recordings to assess how closely the animated faces resemble human behavior.

(II) Lip Synchronization Accuracy: evaluates how well the lip movements of the animated face are synchronized with the audio input.

- Visual Lip Sync Test: Evaluators watch the animated face and determine whether the lip movements are in sync with the speech audio. The key criteria include whether the lip movements match the speech phonemes and whether the transitions between visemes are smooth.
- A/B Comparison: human evaluators are shown side-by-side comparisons of the generated animation and the ground truth (real video) and asked which one has better lip sync accuracy or if they are indistinguishable.

(III) Expression Realism and Emotional Consistency: focuses on how well the animated face conveys emotions and expressions that are consistent with the content of the audio.

- Expression Realism Rating: Human evaluators rate how realistic and appropriate the facial expressions are based on the context of the spoken words. They may be asked whether the expressions match the emotional tone of the speech (e.g., happiness, sadness, surprise).
- Emotional Consistency Test: Evaluators assess whether the generated facial expressions are consistent with the emotion implied by the audio (e.g., if happy speech leads to a ng face). They can also evaluate how smoothly emotional transitions occur during the animation.
- Contextual Appropriateness: evaluators judge if the facial expressions are contextually appropriate, i.e., whether the animation expresses the right emotion or expression for the specific **dialogue** or speech content.

(IV) Temporal Smoothness and Continuity: assesses whether the generated animations are temporally coherent and visually smooth over time.

- Smoothness Evaluation: Evaluators focus on how smoothly the facial movements transition from one frame to the next, particularly in areas like the mouth, eyes, and eyebrows. Jerky or unnatural transitions can lead to lower ratings.
- Temporal Coherence: human evaluators examine the overall temporal continuity of the animation, checking for any glitches, jitter, or sudden changes that disrupt the natural flow of expressions and movements.

- Emotion Transition Smoothness: when emotions change throughout the speech (e.g., neutral to happy), evaluators assess how naturally the facial expressions transition from one emotion to another.

(V) Overall Perceptual Quality: combines several aspects (lip sync, realism, expressions) to give a general assessment of the perceived quality of the animation.

Perceptual Survey: a broad survey where participants evaluate various aspects of the face animation—such as naturalness, expression, lip sync, and overall realism—and provide holistic feedback on the quality of the animation.

- Immersiveness and Engagement: Evaluators rate how engaging and immersive the animation feels. In applications like gaming or virtual assistants, a higher sense of immersion means the animated faces feel more believable and human-like.
- Consistency with Personality or Identity: Evaluators assess whether the generated face animations are consistent with the identity of the character or speaker. This is particularly important for applications where maintaining a character's distinct personality through facial expressions is crucial.

(VI) User Experience and Acceptance Tests: these tests focus on how end-users perceive the system in real-world applications, especially in interactive environments like virtual assistants or video games.

- User Interaction Feedback: end-users interact with the system and provide qualitative feedback on how well the animated face corresponds to the speech, its responsiveness, and whether they find the system engaging and easy to use.
- Task-Based Evaluation: in applications like virtual tour guides or digital assistants, users are asked to complete tasks while interacting with the animated face and then rate their overall experience, focusing on the responsiveness and believability of the facial animations.
- Naturalness in Social Interaction: Evaluators are asked how natural the face animations are during conversations or social interactions. This is especially relevant in digital human or virtual assistant applications, where natural interaction is crucial.

3.3. Results

As illustrated in the Table 3, FaceDiffuser outperforms all other methods in both LVE (Lowest Vertex Error) and FDD (Facial Dynamics Deviation), achieving the lowest values of 4.2985 mm and 3.9101 $\times$ 10^{-5} m, respectively. Following closely are CodeTalker and FaceFormer, which also demonstrate strong results in generating accurate and dynamic facial animations.

Table 3. Objective results computed over the BIWI dataset.

Method	LVE (mm)	FDD ($\times 10^{-5}$ m)
VOCA [86]	6.7155	7.5320
MeshTalk [89]	5.9181	5.1025
FaceFormer [96]	4.9847	5.0972
CodeTalker [99]	4.7914	4.1170
FaceDiffuser [115]	4.2985	3.9101

When evaluating methods for generating facial animations, the Lowest Vertex Error (LVE) is a key metric. Ideally, the LVE should be less than 2 mm, which indicates that the generated animations have very little difference from real facial movements. Values below this threshold suggest excellent lip synchronization, effectively aligning with the audio input. Within an acceptable range, the LVE should be below 5 mm, and values within this range are generally sufficient to ensure good lip synchronization performance.

Although FaceDiffuser has an LVE of 4.2985 mm, which is slightly above the acceptable upper limit, it still demonstrates strong generative capabilities, indicating its advantage in

producing dynamic facial animations. However, to further enhance its lip synchronization performance, it is recommended to optimize the algorithm to lower the LVE value to below the ideal threshold of 2 mm. This would not only improve the realism of the generated animations but also enhance the user's immersive experience.

Ideally, FDD should be less than 2×10^{-5} m, but values below 5×10^{-5} m can still be acceptable. In contrast, VOCA shows the highest LVE at 6.7155 mm, reflecting a less precise alignment of lip movements with audio input compared to the other methods. This highlights the strengths of FaceDiffuser, CodeTalker, and FaceFormer in maintaining better lip synchronization in their generated animations.

3.3.1. Lip Synchronization

As shown in Table 4, wav2Lip achieves the highest scores for lip synchronization using LSE-C (10.08/8.13), demonstrating its excellence in generating highly synchronized lip movements with audio input. A score of LSE-C greater than 8 is considered good, indicating Wav2Lip's strong performance. In contrast, MakeItTalk has the lowest lip sync scores, reflecting weaker synchronization compared to other methods.

Table 4. Comparison with the state-of-the-art methods on HDTF and VoxCeleb dataset. The data presented in the table are in the order of HDTF/VoxCeleb.

Method	Lip Synchronization	Motion Diversity		Image Quality		
	LSE-C	Diversity	Beat Align	FID	PSNR	SSIM
Wav2Lip [59]	10.08/8.13	-	-	22.67/23.85	32.33/35.19	0.740/0.653
MakeItTalk [45]	4.89/2.96	0.238/0.260	0.221/0.252	28.96/31.77	17.95/21.08	0.623/0.529
SadTalker [74]	6.11/4.51	0.275/0.319	0.296/0.328	23.76/24.19	35.78/37.90	0.746/0.690
DiffTalk [62]	6.06/4.38	0.235/0.258	0.226/0.253	23.99/24.06	36.51/36.17	0.721/0.686
DreamTalk [78]	6.93/4.76	0.236/0.257	0.213/0.249	24.30/23.61	32.82/33.16	0.738/0.692

3.3.2. Motion Diversity

As shown in Table 4, SadTalker leads in motion diversity with scores of 0.275/0.319, indicating its ability to generate more varied facial movements. MakeItTalk and DreamTalk follow closely, while Dif-fTalk shows the lowest motion diversity, suggesting more conservative or uniform facial motions. A Beat Align score above 0.3 is considered good, highlighting SadTalker's superior capability in this aspect.

3.3.3. Image Quality

As shown in Table 4, SadTalker excels in image quality as well, achieving high PSNR (35.78/37.90) and SSIM (0.746/0.690) scores, which indicate better structural similarity and sharpness of the generated images. Although Wav2Lip has relatively high PSNR values, it has slightly lower FID and SSIM scores than SadTalker, reflecting good but less detailed visual quality. Conversely, MakeItTalk records the lowest PSNR (17.95/21.08) and SSIM (0.623/0.529) values, indicating weaker image quality overall. For reference, a PSNR above 40 and an SSIM above 0.9 are considered good, underscoring SadTalker's strong performance in image generation.

Wav2Lip excels in lip synchronization, making it suitable for tasks that require precise audio–visual alignment. SadTalker stands out for motion diversity and image quality, providing more varied and visually appealing animations. DiffTalk performs well across all categories but is slightly conservative in terms of motion diversity.

However, it is important to note that there are currently no absolute metrics to define ranges for bad, acceptable, good, and very good performance, as these values can vary depending on the dataset used.

3.3.4. Impact of Language on Animation Quality

As shown in Table 5, differences in phonemes and pronunciation between languages may affect the accuracy of facial animations. For example, the pronunciation features of Italian and Greek differ significantly from those of English and French, potentially leading to discrepancies in the quality and expressiveness of the generated animations.

Table 5. Quantitative comparison to existing methods. We compare existing methods on the test split of the MultiTalk dataset on 4 languages: English (En), Italian (It), French (Fr), and Greek (El).

diffMethod	LVE (mm)			
	En	It	Fr	El
VOCA [86]	1.95	2.78	1.93	2.18
FaceFormer [96]	1.82	2.56	1.78	1.99
CodeTalker [99]	1.98	2.56	1.99	2.09
SelfTalk [110]	1.99	2.59	1.98	2.11
MultiTalk [161]	1.16	1.06	1.39	1.26

This indicates that training models on specific language datasets may influence their generalization capabilities in other languages.

4. Conclusions and Future Directions

Audio-driven face animation methods have advanced significantly, leveraging a variety of techniques to produce realistic and expressive facial animations from audio inputs. Landmark-based methods offer a straightforward approach but may lack detail and expressiveness. Encoder–decoder methods provide flexibility and detail but require substantial data and computational resources. Three-dimensional model-based methods excel in realism and customizability but involve complex integration. Transformer-based methods handle long-range dependencies effectively, while advanced generative network-based methods achieve high-quality outputs but face challenges in training complexity and data requirements.

These diverse methods have applications across various fields, including entertainment, virtual and augmented reality, digital assistants, education, social media, and accessibility. For example, landmark-based methods can enhance user experiences in mobile applications by providing engaging facial animations that respond dynamically to audio. In contrast, more resource-intensive methods can be leveraged in enterprise applications, such as film production or sophisticated animation projects, where higher accuracy and complexity are crucial.

The future of audio-driven facial animation research is set for remarkable progress, emphasizing realism, emotional depth, and user engagement. Key areas of focus include the integration of nonverbal elements, such as paralanguage and silence, which are vital for conveying complex emotions and enhancing character authenticity. Advances in synchronization and temporal consistency will likely improve lip sync accuracy and expression transitions, aided by innovative sequence-to-sequence models and diverse linguistic datasets to tackle cross-lingual challenges. Moreover, the trend toward personalization will enable users to create unique avatars reflecting their individual expressions, enhancing immersion. By addressing these aspects, future research can significantly boost the expressiveness and realism of digital humans, expanding their applications in interactive media like games and virtual assistants.

4.1. Nonverbal Elements and Silence

"Nonverbal elements", such as "paralanguage", including laughter, sighs, and other sounds, play a crucial role in conveying emotional depth and enhancing the realism of animated characters. Research indicates that these elements can significantly improve the emotional expressiveness and interactive quality of virtual characters. Additionally, the importance of how moments of silence convey complex emotions deserves attention, as

this aspect remains underexplored yet is essential for creating more nuanced animations. Furthermore, establishing standardized evaluation metrics that encompass both verbal and nonverbal cues will be vital for advancing the field and facilitating comparisons between different methods. By focusing on these areas, future research can greatly enhance the expressiveness and authenticity of digital humans, ultimately benefiting their application in interactive media such as games and virtual assistants.

4.2. Synchronization and Temporal Consistency

Further advancements are expected in improving synchronization and temporal dynamics, particularly in achieving more accurate lip sync and natural transitions between facial expressions. Researchers will explore enhanced sequence-to-sequence models and temporal consistency techniques to meet these goals. Supporting cross-lingual and multilingual animation is also a critical future direction, with models trained on diverse linguistic datasets to account for differences in phonetic and prosodic features across languages.

4.3. Generalization and Customization

The future of facial animation will increasingly involve personalization and adaptability. Models will need to be capable of adjusting to individual facial characteristics and expressions, creating personalized avatars. Additionally, providing users with tools to customize expression intensity and style will enable more interactive and immersive experiences.

Author Contributions: Investigation, D.J. and L.Y.; writing—original draft preparation, D.J.; writing—review and editing, D.J., L.Y., R.K., and S.B.; project administration, G.M.; funding acquisition, J.C. All authors have read and agreed to the published version of the manuscript.

Funding: The survey was supported by funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 900025.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

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

References

1. Chen, T. Audiovisual Speech Processing. *IEEE Signal Process. Mag.* **2001**, *18*, 9–21. [CrossRef]
2. Seymour, M.; Evans, C.; Libreri, K. Meet Mike: Epic avatars. In *ACM SIGGRAPH 2017 VR Village*; ACM: Los Angeles, CA, USA, 2017; pp. 1–2.
3. Charalambous, C.; Yumak, Z.; Van Der Stappen, A.F. Audio-driven Emotional Speech Animation for Interactive Virtual Characters. *Comput. Animat. Virtual* **2019**, *30*, e1892. [CrossRef]
4. Xu, M.; Duan, L.Y.; Cai, J.; Chia, L.T.; Xu, C.; Tian, Q. HMM-Based Audio Keyword Generation. In *Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 566–574. [CrossRef]
5. Deng, L.; O'Shaughnessy, D. *Speech Processing: A Dynamic and Optimization-Oriented Approach*; Signal Processing and Communications; Marcel Dekker: New York, NY, USA, 2003.
6. Gulati, A.; Qin, J.; Chiu, C.-C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. Conformer: Convolution-Augmented Transformer for Speech Recognition. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020; ISCA: Shanghai, China, 2020; pp. 5036–5040. [CrossRef]
7. Han, W.; Zhang, Z.; Zhang, Y.; Yu, J.; Chiu, C.-C.; Qin, J.; Gulati, A.; Pang, R.; Wu, Y. ContextNet: Improving Convolutional Neural Networks for Automatic Speech Recognition with Global Context. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020; ISCA: Shanghai, China, 2020; pp. 3610–3614. [CrossRef]
8. Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T.-Y. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. *arXiv* **2022**, arXiv:2006.04558.
9. Hsu, W.-N.; Bolte, B.; Tsai, Y.-H.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3451–3460. [CrossRef]
10. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision. In Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; pp. 28492–28518.
11. Chan, W.; Park, D.; Lee, C.; Zhang, Y.; Le, Q.; Norouzi, M. SpeechStew: Simply Mix All Available Speech Recognition Data to Train One Large Neural Network. *arXiv* **2021**, arXiv:2104.02133.

12. Chen, S.; Liu, S.; Zhou, L.; Liu, Y.; Tan, X.; Li, J.; Zhao, S.; Qian, Y.; Wei, F. VALL-E 2: Neural Codec Language Models Are Human Parity Zero-Shot Text to Speech Synthesizers. *arXiv* **2024**, arXiv:2406.05370.
13. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. Neural Inf. Process. Syst.* **2023**, *33*, 12449–12460.
14. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*. Available online: https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf (accessed on 21 October 2024).
15. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2022**, arXiv:1312.6114.
16. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
17. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
18. Blanz, V.; Vetter, T. A Morphable Model for the Synthesis of 3D Faces. In Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques—SIGGRAPH '99, Los Angeles, CA, USA, 8–13 August 1999; pp. 187–194. [CrossRef]
19. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
20. 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* **2022**, *65*, 99–106. [CrossRef]
21. Magnenat Thalmann, N.; Thalmann, D. *Models and Techniques in Computer Animation*; Springer: Tokyo, Japan, 2013.
22. Fisher, C.G. Confusions Among Visually Perceived Consonants. *J. Speech Hear. Res.* **1968**, *11*, 796–804. [CrossRef] [PubMed]
23. Brand, M. Voice Puppetry. In Proceedings of the 26th Annual Conference on Computer graphics and Interactive Techniques—SIGGRAPH '99, Los Angeles, CA, USA, 8–13 August 1999; pp. 21–28. [CrossRef]
24. Anderson, R.; Stenger, B.; Wan, V.; Cipolla, R. Expressive Visual Text-to-Speech Using Active Appearance Models. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; IEEE: Portland, OR, USA, 2013; pp. 3382–3389. [CrossRef]
25. Wang, L.; Soong, F.K. HMM Trajectory-Guided Sample Selection for Photo-Realistic Talking Head. *Multimed. Tools Appl.* **2015**, *74*, 9849–9869. [CrossRef]
26. Deena, S.; Galata, A. Speech-Driven Facial Animation Using a Shared Gaussian Process Latent Variable Model. In *Advances in Visual Computing*; Springer: Berlin/Heidelberg, Germany, 2009; Volume 5875, pp. 89–100. [CrossRef]
27. Deena, S.; Hou, S.; Galata, A. Visual Speech Synthesis Using a Variable-Order Switching Shared Gaussian Process Dynamical Model. *IEEE Trans. Multimed.* **2013**, *15*, 1755–1768. [CrossRef]
28. Schabus, D.; Pucher, M.; Hofer, G. Joint Audiovisual Hidden Semi-Markov Model-Based Speech Synthesis. *IEEE J. Sel. Top. Signal Process.* **2014**, *8*, 336–347. [CrossRef]
29. Fan, B.; Xie, L.; Yang, S.; Wang, L.; Soong, F.K. A Deep Bidirectional LSTM Approach for Video-Realistic Talking Head. *Multimed. Tools Appl.* **2016**, *75*, 5287–5309. [CrossRef]
30. Suwajanakorn, S.; Seitz, S.M.; Kemelmacher-Shlizerman, I. Synthesizing Obama: Learning Lip Sync from Audio. *ACM Trans. Graph.* **2017**, *36*, 1–13. [CrossRef]
31. 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.* **2017**, *36*, 1–12. [CrossRef]
32. Kammoun, A.; Slama, R.; Tabia, H.; Ouni, T.; Abid, M. Generative Adversarial Networks for Face Generation: A Survey. *ACM Comput. Surv.* **2023**, *55*, 1–37. [CrossRef]
33. Kandwal, S.; Nehra, V. A Survey of Text-to-Image Diffusion Models in Generative AI. In Proceedings of the 2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 18–19 January 2024; IEEE: Noida, India, 2024; pp. 73–78. [CrossRef]
34. Liu, S. Audio-Driven Talking Face Generation: A Review. *J. Audio Eng. Soc.* **2023**, *71*, 408–419. [CrossRef]
35. Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A.; Ortega-Garcia, J. Deepfakes and beyond: A Survey of Face Manipulation and Fake Detection. *Inf. Fusion* **2020**, *64*, 131–148. [CrossRef]
36. Mirsky, Y.; Lee, W. The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.* **2022**, *54*, 1–41. [CrossRef]
37. 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. [CrossRef]
38. Sha, T.; Zhang, W.; Shen, T.; Li, Z.; Mei, T. Deep Person Generation: A Survey from the Perspective of Face, Pose and Cloth Synthesis. *ACM Comput. Surv.* **2023**, *55*, 1–37. [CrossRef]
39. Gowda, S.N.; Pandey, D.; Gowda, S.N. From Pixels to Portraits: A Comprehensive Survey of Talking Head Generation Techniques and Applications. *arXiv* **2023**, arXiv:2308.16041.
40. Meng, M.; Zhao, Y.; Zhang, B.; Zhu, Y.; Shi, W.; Wen, M.; Fan, Z. A Comprehensive Taxonomy and Analysis of Talking Head Synthesis: Techniques for Portrait Generation, Driving Mechanisms, and Editing. *arXiv* **2024**, arXiv:2406.10553.
41. Jalalifar, S.A.; Hasani, H.; Aghajan, H. Speech-Driven Facial Reenactment Using Conditional Generative Adversarial Networks. *arXiv* **2018**, arXiv:1803.07461.

42. Chen, L.; Maddox, R.K.; Duan, Z.; Xu, C. Hierarchical Cross-Modal Talking Face Generation with Dynamic Pixel-Wise Loss. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; IEEE: Long Beach, CA, USA, 2019; pp. 7824–7833. [CrossRef]
43. Das, D.; Biswas, S.; Sinha, S.; Bhowmick, B. Speech-Driven Facial Animation Using Cascaded GANs for Learning of Motion and Texture. In Proceedings of the Computer Vision—ECCV 2020, Glasgow, UK, 23–28 August 2020; Volume 12375, pp. 408–424. [CrossRef]
44. Lu, Y.; Chai, J.; Cao, X. Live Speech Portraits: Real-Time Photorealistic Talking-Head Animation. *ACM Trans. Graph.* **2021**, *40*, 1–17. [CrossRef]
45. Zhou, Y.; Han, X.; Shechtman, E.; Echevarria, J.; Kalogerakis, E.; Li, D. MakeltTalk: Speaker-Aware Talking-Head Animation. *ACM Trans. Graph.* **2020**, *39*, 1–15. [CrossRef]
46. Chen, Z.; Cao, J.; Chen, Z.; Li, Y.; Ma, C. EchoMimic: Lifelike Audio-Driven Portrait Animations through Editable Landmark Conditions. *arXiv* **2024**, arXiv:2407.08136.
47. Tan, J.; Cheng, X.; Xiong, L.; Zhu, L.; Li, X.; Wu, X.; Gong, K.; Li, M.; Cai, Y. Landmark-Guided Diffusion Model for High-Fidelity and Temporally Coherent Talking Head Generation. *arXiv* **2024**, arXiv:2408.01732.
48. Zhong, W.; Lin, J.; Chen, P.; Lin, L.; Li, G. High-Fidelity and Lip-Synced Talking Face Synthesis via Landmark-Based Diffusion Model. *arXiv* **2024**, arXiv:2408.05416.
49. Jamaludin, A.; Chung, J.S.; Zisserman, A. You Said That?: Synthesising Talking Faces from Audio. *Int. J. Comput. Vis.* **2019**, *127*, 1767–1779. [CrossRef]
50. Chung, J.S.; Jamaludin, A.; Zisserman, A. You Said That? *arXiv* **2017**, arXiv:1705.02966.
51. Vougioukas, K.; Petridis, S.; Pantic, M. Realistic Speech-Driven Facial Animation with GANs. *Int. J. Comput. Vis.* **2020**, *128*, 1398–1413. [CrossRef]
52. Wiles, O.; Koepke, A.S.; Zisserman, A. X2Face: A Network for Controlling Face Generation Using Images, Audio, and Pose Codes. In Proceedings of the Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018; Volume 11217, pp. 690–706. [CrossRef]
53. Chen, L.; Li, Z.; Maddox, R.K.; Duan, Z.; Xu, C. Lip Movements Generation at a Glance. In Proceedings of the Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018; Volume 11211, pp. 538–553. [CrossRef]
54. Fan, B.; Wang, L.; Soong, F.K.; Xie, L. Photo-Real Talking Head with Deep Bidirectional LSTM. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015; IEEE: South Brisbane, QLD, Australia, 2015; pp. 4884–4888. [CrossRef]
55. Pham, H.X.; Cheung, S.; Pavlovic, V. Speech-Driven 3D Facial Animation with Implicit Emotional Awareness: A Deep Learning Approach. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; IEEE: Honolulu, HI, USA, 2017; pp. 2328–2336. [CrossRef]
56. Vougioukas, K.; Petridis, S.; Pantic, M. End-to-End Speech-Driven Realistic Facial Animation with Temporal GANs. In Proceedings of the CVPR Workshops, Long Beach, CA, USA, 16–20 June 2019; Volume 887, pp. 37–40.
57. Song, Y.; Zhu, J.; Li, D.; Wang, A.; Qi, H. Talking Face Generation by Conditional Recurrent Adversarial Network. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019; International Joint Conferences on Artificial Intelligence Organization: Macao, China, 2019; pp. 919–925. [CrossRef]
58. Zhu, H.; Huang, H.; Li, Y.; Zheng, A.; He, R. Arbitrary Talking Face Generation via Attentional Audio-Visual Coherence Learning. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, Yokohama, Japan, 11–17 July 2020; International Joint Conferences on Artificial Intelligence Organization: Yokohama, Japan, 2020; pp. 2362–2368. [CrossRef]
59. 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; ACM: Seattle WA USA, 2020; pp. 484–492. [CrossRef]
60. Kumar, N.; Goel, S.; Narang, A.; Hasan, M. Robust One Shot Audio to Video Generation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; IEEE: Seattle, WA, USA, 2020; pp. 3334–3343. [CrossRef]
61. Yaman, D.; Eyiokur, F.I.; Bärmann, L.; Aktı, S.; Ekenel, H.K.; Waibel, A. Audio-Visual Speech Representation Expert for Enhanced Talking Face Video Generation and Evaluation. *arXiv* **2024**, arXiv:2405.04327.
62. Shen, S.; Zhao, W.; Meng, Z.; Li, W.; Zhu, Z.; Zhou, J.; Lu, J. DiffTalk: Crafting Diffusion Models for Generalized Audio-Driven Portraits Animation. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; IEEE: Vancouver, BC, Canada, 2023; pp. 1982–1991. [CrossRef]
63. Zhao, D.; Shi, J.; Li, W.; Wang, S.; Xu, S.; Pan, Z. Controllable Talking Face Generation by Implicit Facial Keypoints Editing. *arXiv* **2024**, arXiv:2406.02880.
64. Yin, F.; Zhang, Y.; Cun, X.; Cao, M.; Fan, Y.; Wang, X.; Bai, Q.; Wu, B.; Wang, J.; Yang, Y. StyleHEAT: One-Shot High-Resolution Editable Talking Face Generation via Pre-Trained StyleGAN. In Proceedings of the Computer Vision—ECCV 2022, Tel Aviv, Israel, 23–27 October 2022; Volume 13677, pp. 85–101. [CrossRef]
65. Liu, T.; Chen, F.; Fan, S.; Du, C.; Chen, Q.; Chen, X.; Yu, K. AniTalker: Animate Vivid and Diverse Talking Faces through Identity-Decoupled Facial Motion Encoding. *arXiv* **2024**, arXiv:2405.03121.

66. Wang, S.; Li, L.; Ding, Y.; Yu, X. One-Shot Talking Face Generation from Single-Speaker Audio-Visual Correlation Learning. *AAAI* **2022**, *36*, 2531–2539. [CrossRef]
67. Yao, Z.; Cheng, X.; Huang, Z. FD2Talk: Towards Generalized Talking Head Generation with Facial Decoupled Diffusion Model. *arXiv* **2024**, arXiv:2408.09384.
68. Lin, G.; Jiang, J.; Liang, C.; Zhong, T.; Yang, J.; Zheng, Y. CyberHost: Taming Audio-Driven Avatar Diffusion Model with Region Codebook Attention. *arXiv* **2024**, arXiv:2409.01876.
69. Zeng, D.; Liu, H.; Lin, H.; Ge, S. Talking Face Generation with Expression-Tailored Generative Adversarial Network. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; ACM: Seattle WA USA, 2020; pp. 1716–1724. [CrossRef]
70. 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 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; IEEE: Nashville, TN, USA, 2021; pp. 4174–4184. [CrossRef]
71. 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]
72. Mittal, G.; Wang, B. Animating Face Using Disentangled Audio Representations. In Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA, 1–5 March 2020; IEEE: Snowmass Village, CO, USA, 2020; pp. 3279–3287. [CrossRef]
73. Ji, X.; Zhou, H.; Wang, K.; Wu, W.; Loy, C.C.; Cao, X.; Xu, F. Audio-Driven Emotional Video Portraits. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; IEEE: Nashville, TN, USA, 2021; pp. 14075–14084. [CrossRef]
74. Zhang, W.; Cun, X.; Wang, X.; Zhang, Y.; Shen, X.; Guo, Y.; Shan, Y.; Wang, F. SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation. *arXiv* **2023**, arXiv:2211.12194.
75. Yi, R.; Ye, Z.; Zhang, J.; Bao, H.; Liu, Y.-J. Audio-Driven Talking Face Video Generation with Learning-Based Personalized Head Pose. *arXiv* **2020**, arXiv:2002.10137.
76. 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 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; IEEE: Montreal, QC, Canada, 2021; pp. 3847–3856. [CrossRef]
77. 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 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; IEEE: Nashville, TN, USA, 2021; pp. 3660–3669. [CrossRef]
78. Ma, Y.; Zhang, S.; Wang, J.; Wang, X.; Zhang, Y.; Deng, Z. DreamTalk: When Emotional Talking Head Generation Meets Diffusion Probabilistic Models. *arXiv* **2023**, arXiv:2312.09767.
79. Lahiri, A.; Kwatra, V.; Frueh, C.; Lewis, J.; Bregler, C. LipSync3D: Data-Efficient Learning of Personalized 3D Talking Faces from Video Using Pose and Lighting Normalization. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; IEEE: Nashville, TN, USA, 2021; pp. 2754–2763. [CrossRef]
80. Liang, J.; Lu, F. Emotional Conversation: Empowering Talking Faces with Cohesive Expression, Gaze and Pose Generation. *arXiv* **2024**, arXiv:2406.07895.
81. Zhou, H.; Liu, Y.; Liu, Z.; Luo, P.; Wang, X. Talking Face Generation by Adversarially Disentangled Audio-Visual Representation. *AAAI* **2019**, *33*, 9299–9306. [CrossRef]
82. Taylor, S.; Kim, T.; Yue, Y.; Mahler, M.; Krahe, J.; Rodriguez, A.G.; Hodgins, J.; Matthews, I. A Deep Learning Approach for Generalized Speech Animation. *ACM Trans. Graph.* **2017**, *36*, 1–11. [CrossRef]
83. Pham, H.X.; Wang, Y.; Pavlovic, V. End-to-End Learning for 3D Facial Animation from Speech. In Proceedings of the 20th ACM International Conference on Multimodal Interaction, Boulder, CO, USA, 16–20 October 2018; ACM: Boulder, CO, USA, 2018; pp. 361–365. [CrossRef]
84. Zhou, Y.; Xu, Z.; Landreth, C.; Kalogerakis, E.; Maji, S.; Singh, K. Visemenet: Audio-Driven Animator-Centric Speech Animation. *ACM Trans. Graph.* **2018**, *37*, 1–10. [CrossRef]
85. Sadoughi, N.; Busso, C. Speech-Driven Expressive Talking Lips with Conditional Sequential Generative Adversarial Networks. *IEEE Trans. Affect. Comput.* **2021**, *12*, 1031–1044. [CrossRef]
86. Cudeiro, D.; Bolkart, T.; Laidlaw, C.; Ranjan, A.; Black, M.J. Capture, Learning, and Synthesis of 3D Speaking Styles. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; IEEE: Long Beach, CA, USA, 2019; pp. 10093–10103. [CrossRef]
87. Richard, A.; Lea, C.; Ma, S.; Gall, J.; La Torre, F.D.; Sheikh, Y. Audio- and Gaze-Driven Facial Animation of Codec Avatars. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; IEEE: Waikoloa, HI, USA, 2021; pp. 41–50. [CrossRef]
88. Song, L.; Wu, W.; Qian, C.; He, R.; Loy, C.C. Everybody's Talkin': Let Me Talk as You Want. *IEEE Trans. Inform. Forensic Secur.* **2022**, *17*, 585–598. [CrossRef]
89. Richard, A.; Zollhofer, M.; Wen, Y.; De La Torre, F.; Sheikh, Y. MeshTalk: 3D Face Animation from Speech Using Cross-Modality Disentanglement. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; IEEE: Montreal, QC, Canada, 2021; pp. 1153–1162. [CrossRef]

90. Fan, Y.; Lin, Z.; Saito, J.; Wang, W.; Komura, T. Joint Audio-Text Model for Expressive Speech-Driven 3D Facial Animation. *Proc. ACM Comput. Graph. Interact. Tech.* **2022**, *5*, 1–15. [CrossRef]
91. Abdelaziz, A.H.; Theobald, B.-J.; Dixon, P.; Knothe, R.; Apostoloff, N.; Kajareker, S. Modality Dropout for Improved Performance-Driven Talking Faces. *arXiv* **2020**, arXiv:2005.13616.
92. Chen, L.; Cui, G.; Liu, C.; Li, Z.; Kou, Z.; Xu, Y.; Xu, C. Talking-Head Generation with Rhythmic Head Motion. In Proceedings of the Computer Vision—ECCV 2020, Glasgow, UK, 23–28 August 2020; Volume 12354, pp. 35–51. [CrossRef]
93. Huang, D.-Y.; Chandra, E.; Yang, X.; Zhou, Y.; Ming, H.; Lin, W.; Dong, M.; Li, H. Visual Speech Emotion Conversion Using Deep Learning for 3D Talking Head. In Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and First Multi-Modal Affective Computing of Large-Scale Multimedia Data, Seoul, Republic of Korea, 26 October 2018; ACM: Seoul, Republic of Korea, 2018; pp. 7–13. [CrossRef]
94. Wang, Q.; Fan, Z.; Xia, S. 3D-TalkEmo: Learning to Synthesize 3D Emotional Talking Head. *arXiv* **2021**, arXiv:2104.12051.
95. Thambiraja, B.; Habibie, I.; Aliakbarian, S.; Cosker, D.; Theobalt, C.; Thies, J. Imitator: Personalized Speech-Driven 3D Facial Animation. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–3 October 2023; IEEE: Paris, France, 2023; pp. 20564–20574. [CrossRef]
96. Fan, Y.; Lin, Z.; Saito, J.; Wang, W.; Komura, T. FaceFormer: Speech-Driven 3D Facial Animation with Transformers. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; IEEE: New Orleans, LA, USA, 2022; pp. 18749–18758. [CrossRef]
97. Lu, L.; Zhang, T.; Liu, Y.; Chu, X.; Li, Y. Audio-Driven 3D Facial Animation from In-the-Wild Videos. *arXiv* **2023**, arXiv:2306.11541.
98. Chai, Y.; Shao, T.; Weng, Y.; Zhou, K. Personalized Audio-Driven 3D Facial Animation via Style-Content Disentanglement. *IEEE Trans. Visual. Comput. Graphics* **2024**, *30*, 1803–1820. [CrossRef]
99. Xing, J.; Xia, M.; Zhang, Y.; Cun, X.; Wang, J.; Wong, T.-T. CodeTalker: Speech-Driven 3D Facial Animation with Discrete Motion Prior. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; IEEE: Vancouver, BC, Canada, 2023; pp. 12780–12790. [CrossRef]
100. Peng, Z.; Wu, H.; Song, Z.; Xu, H.; Zhu, X.; He, J.; Liu, H.; Fan, Z. EmoTalk: Speech-Driven Emotional Disentanglement for 3D Face Animation. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–3 October 2023; IEEE: Paris, France, 2023; pp. 20630–20640. [CrossRef]
101. Daněček, R.; Chhatre, K.; Tripathi, S.; Wen, Y.; Black, M.J.; Bolkart, T. Emotional Speech-Driven Animation with Content-Emotion Disentanglement. In Proceedings of the SIGGRAPH Asia 2023 Conference Papers, Sydney, NSW, Australia, 12–15 December 2023; pp. 1–13. [CrossRef]
102. Han, T.; Gui, S.; Huang, Y.; Li, B.; Liu, L.; Zhou, B.; Jiang, N.; Lu, Q.; Zhi, R.; Liang, Y.; et al. PMMTalk: Speech-Driven 3D Facial Animation from Complementary Pseudo Multi-Modal Features. *arXiv* **2023**, arXiv:2312.02781.
103. Sun, M.; Xu, C.; Jiang, X.; Liu, Y.; Sun, B.; Huang, R. Beyond Talking—Generating Holistic 3D Human Dyadic Motion for Communication. *arXiv* **2024**, arXiv:2403.19467.
104. He, S.; He, H.; Yang, S.; Wu, X.; Xia, P.; Yin, B.; Liu, C.; Dai, L.; Xu, C. Speech4Mesh: Speech-Assisted Monocular 3D Facial Reconstruction for Speech-Driven 3D Facial Animation. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–3 October 2023; IEEE: Paris, France, 2023; pp. 14146–14156. [CrossRef]
105. Liang, X.; Zhuang, W.; Wang, T.; Geng, G.; Geng, G.; Xia, H.; Xia, S. CSTalk: Correlation Supervised Speech-Driven 3D Emotional Facial Animation Generation. *arXiv* **2024**, arXiv:2404.18604.
106. Lin, Y.; Peng, L.; Hu, J.; Li, X.; Kang, W.; Lei, S.; Wu, X.; Xu, H. EmoFace: Emotion-Content Disentangled Speech-Driven 3D Talking Face with Mesh Attention. *arXiv* **2024**, arXiv:2408.11518.
107. Jafari, F.; Berretti, S.; Basu, A. JambaTalk: Speech-Driven 3D Talking Head Generation Based on Hybrid Transformer-Mamba Model. *arXiv* **2024**, arXiv:2408.01627.
108. Zhuang, Y.; Cheng, B.; Cheng, Y.; Jin, Y.; Liu, R.; Li, C.; Cheng, X.; Liao, J.; Lin, J. Learn2Talk: 3D Talking Face Learns from 2D Talking Face. *arXiv* **2024**, arXiv:2404.12888. [CrossRef] [PubMed]
109. Ji, X.; Lin, C.; Ding, Z.; Tai, Y.; Zhu, J.; Hu, X.; Luo, D.; Ge, Y.; Wang, C. RealTalk: Real-Time and Realistic Audio-Driven Face Generation with 3D Facial Prior-Guided Identity Alignment Network. *arXiv* **2024**, arXiv:2406.18284.
110. Peng, Z.; Luo, Y.; Shi, Y.; Xu, H.; Zhu, X.; Liu, H.; He, J.; Fan, Z. SelfTalk: A Self-Supervised Commutative Training Diagram to Comprehend 3D Talking Faces. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; ACM: Ottawa, ON, Canada, 2023; pp. 5292–5301. [CrossRef]
111. Fan, X.; Li, J.; Lin, Z.; Xiao, W.; Yang, L. UniTalker: Scaling up Audio-Driven 3D Facial Animation through A Unified Model. *arXiv* **2024**, arXiv:2408.00762.
112. Chu, Z.; Guo, K.; Xing, X.; Lan, Y.; Cai, B.; Xu, X. CorrTalk: Correlation Between Hierarchical Speech and Facial Activity Variances for 3D Animation. *arXiv* **2023**, arXiv:2310.11295. [CrossRef]
113. Thambiraja, B.; Aliakbarian, S.; Cosker, D.; Thies, J. 3DiFACE: Diffusion-Based Speech-Driven 3D Facial Animation and Editing. *arXiv* **2023**, arXiv:2312.00870.
114. Xu, Z.; Zhang, J.; Liew, J.H.; Zhang, W.; Bai, S.; Feng, J.; Shou, M.Z. PV3D: A 3D Generative Model for Portrait Video Generation. *arXiv* **2022**, arXiv:2212.06384.

115. Stan, S.; Haque, K.I.; Yumak, Z. FaceDiffuser: Speech-Driven 3D Facial Animation Synthesis Using Diffusion. In Proceedings of the ACM SIGGRAPH Conference on Motion, Interaction and Games, Rennes, France, 15–17 November 2023; ACM: Rennes, France, 2023; pp. 1–11. [CrossRef]
116. Chen, P.; Wei, X.; Lu, M.; Zhu, Y.; Yao, N.; Xiao, X.; Chen, H. DiffusionTalker: Personalization and Acceleration for Speech-Driven 3D Face Diffuser. *arXiv* **2023**, arXiv:2311.16565.
117. Papantoniou, F.P.; Filntisis, P.P.; Maragos, P.; Roussos, A. Neural Emotion Director: Speech-Preserving Semantic Control of Facial Expressions in "in-the-Wild" Videos. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; IEEE: New Orleans, LA, USA, 2022; pp. 18759–18768. [CrossRef]
118. Ma, Z.; Zhu, X.; Qi, G.; Qian, C.; Zhang, Z.; Lei, Z. DiffSpeaker: Speech-Driven 3D Facial Animation with Diffusion Transformer. *arXiv* **2024**, arXiv:2402.05712.
119. Aneja, S.; Thies, J.; Dai, A.; Nießner, M. FaceTalk: Audio-Driven Motion Diffusion for Neural Parametric Head Models. *arXiv* **2024**, arXiv:2312.08459.
120. Lin, Y.; Fan, Z.; Xiong, L.; Peng, L.; Li, X.; Kang, W.; Wu, X.; Lei, S.; Xu, H. GLDiTalker: Speech-Driven 3D Facial Animation with Graph Latent Diffusion Transformer. *arXiv* **2024**, arXiv:2408.01826.
121. Xu, Z.; Gong, S.; Tang, J.; Liang, L.; Huang, Y.; Li, H.; Huang, S. KMTalk: Speech-Driven 3D Facial Animation with Key Motion Embedding. *arXiv* **2024**, arXiv:2409.01113.
122. Zhao, Q.; Long, P.; Zhang, Q.; Qin, D.; Liang, H.; Zhang, L.; Zhang, Y.; Yu, J.; Xu, L. Media2Face: Co-Speech Facial Animation Generation with Multi-Modality Guidance. In Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24, Denver, CO, USA, 27 July–1 August 2024; ACM: Denver, CO, USA, 2024; pp. 1–13. [CrossRef]
123. Kim, G.; Seo, K.; Cha, S.; Noh, J. NeRFFaceSpeech: One-Shot Audio-Driven 3D Talking Head Synthesis via Generative Prior. *arXiv* **2024**, arXiv:2405.05749.
124. Alghamdi, N.; Maddock, S.; Marxer, R.; Barker, J.; Brown, G.J. A Corpus of Audio-Visual Lombard Speech with Frontal and Profile Views. *J. Acoust. Soc. Am.* **2018**, *143*, EL523–EL529. [CrossRef] [PubMed]
125. The Sheffield Audio-Visual Lombard Grid Corpus. Available online: https://spandh.dcs.shef.ac.uk/avlombard/ (accessed on 21 October 2024).
126. Cao, H.; Cooper, D.G.; Keutmann, M.K.; Gur, R.C.; Nenkova, A.; Verma, R. CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset. *IEEE Trans. Affect. Comput.* **2014**, *5*, 377–390. Available online: https://github.com/CheyneyComputerScience/CREMA-D (accessed on 21 October 2024). [CrossRef]
127. Fanelli, G.; Gall, J.; Romsdorfer, H.; Weise, T.; Van Gool, L. A 3-D Audio-Visual Corpus of Affective Communication. *IEEE Trans. Multimed.* **2010**, *12*, 591–598. [CrossRef]
128. 3-D Audio-Visual Corpus EULA. Available online: https://data.vision.ee.ethz.ch/cvl/datasets/B3DAC2/CorpusEULA.pdf (accessed on 21 October 2024).
129. Harte, N.; Gillen, E. TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech. *IEEE Trans. Multimed.* **2015**, *17*, 603–615. [CrossRef]
130. TCD-TIMIT Corpus. Available online: https://sigmedia.tcd.ie (accessed on 21 October 2024).
131. Czyzewski, A.; Kostek, B.; Bratoszewski, P.; Kotus, J.; Szykulski, M. An Audio-Visual Corpus for Multimodal Automatic Speech Recognition. *J. Intell. Inf. Syst.* **2017**, *49*, 167–192. [CrossRef]
132. Jachimski, D.; Czyzewski, A.; Ciszewski, T. A Comparative Study of English Viseme Recognition Methods and Algorithms. *Multimed Tools Appl.* **2018**, *77*, 16495–16532. [CrossRef]
133. Kawaler, M.; Czyżewski, A. Database of Speech and Facial Expressions Recorded with Optimized Face Motion Capture Settings. *J. Intell. Inf. Syst.* **2019**, *53*, 381–404. [CrossRef]
134. MODALITY Corpus. Available online: http://www.modality-corpus.org/ (accessed on 21 October 2024).
135. Busso, C.; Parthasarathy, S.; Burmania, A.; AbdelWahab, M.; Sadoughi, N.; Provost, E.M. MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception. *IEEE Trans. Affect. Comput.* **2017**, *8*, 67–80. Available online: https://ecs.utdallas.edu/research/researchlabs/msp-lab/MSP-Improv.html (accessed on 21 October 2024). [CrossRef]
136. Chung, J.S.; Zisserman, A. Lip Reading in the Wild. In *Computer Vision—ACCV 2016*; Lai, S.-H., Lepetit, V., Nishino, K., Sato, Y., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2017; Volume 10112, pp. 87–103. [CrossRef]
137. Lip Reading in the Wild dataset. Available online: https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrw1.html (accessed on 21 October 2024).
138. Chung, J.S.; Senior, A.; Vinyals, O.; Zisserman, A. Lip Reading Sentences in the Wild. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Honolulu, HI, USA, 2017; pp. 3444–3453. [CrossRef]
139. Lip Reading Sentences Dataset. Available online: https://www.robots.ox.ac.uk/~vgg/data/lip_reading/ (accessed on 21 October 2024).
140. Son, J.S.; Zisserman, A. Lip Reading in Profile. In Proceedings of the British Machine Vision Conference 2017, London, UK, 4–7 September 2017; British Machine Vision Association: London, UK, 2017; p. 155. [CrossRef]

141. Afouras, T.; Chung, J.S.; Senior, A.; Vinyals, O.; Zisserman, A. Deep Audio-Visual Speech Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 8717–8727. [CrossRef] [PubMed]
142. The Oxford-BBC Lip Reading Sentences 2 (LRS2) Dataset. Available online: https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs2.html (accessed on 21 October 2024).
143. Afouras, T.; Chung, J.S.; Zisserman, A. LRS3-TED: A Large-Scale Dataset for Visual Speech Recognition. *arXiv* **2018**, arXiv:1809.00496.
144. Lip Reading Sentences 3. Available online: https://mmai.io/datasets/lip_reading/ (accessed on 21 October 2024).
145. Nagrani, A.; Chung, J.S.; Zisserman, A. VoxCeleb: A Large-Scale Speaker Identification Dataset. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; ISCA: Shanghai, China, 2017; pp. 2616–2620. [CrossRef]
146. Nagrani, A.; Chung, J.S.; Xie, W.; Zisserman, A. Voxceleb: Large-Scale Speaker Verification in the Wild. *Comput. Speech Lang.* **2020**, *60*, 101027. [CrossRef]
147. Chung, J.S.; Nagrani, A.; Zisserman, A. VoxCeleb2: Deep Speaker Recognition. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; ISCA: Shanghai, China, 2018; pp. 1086–1090. [CrossRef]
148. VoxCeleb Dataset. Available online: https://www.robots.ox.ac.uk/~vgg/data/voxceleb/ (accessed on 21 October 2024).
149. Livingstone, S.R.; Russo, F.A. 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* **2018**, *13*, e0196391. [CrossRef]
150. RAVDESS dataset. Available online: https://zenodo.org/records/1188976#.YFZuJ0j7SL8 (accessed on 21 October 2024).
151. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; Association for Computational Linguistics: Florence, Italy, 2019; pp. 527–536. [CrossRef]
152. MELD: Multimodal EmotionLines Dataset. Available online: https://affective-meld.github.io/ (accessed on 21 October 2024).
153. Vocaset Project. Available online: https://voca.is.tue.mpg.de (accessed on 21 October 2024).
154. HDTF Dataset. Available online: https://github.com/MRzzm/HDTF (accessed on 21 October 2024).
155. 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 Proceedings of the Computer Vision—ECCV 2020, Glasgow, UK, 23–28 August 2020. Available online: https://wywu.github.io/projects/MEAD/MEAD.html (accessed on 21 October 2024).
156. Zhu, H.; Wu, W.; Zhu, W.; Jiang, L.; Tang, S.; Zhang, L.; Liu, Z.; Loy, C.C. CelebV-HQ: A Large-Scale Video Facial Attributes Dataset. In Proceedings of the Computer Vision—ECCV 2022, Tel Aviv, Israel, 23–27 October 2022; Volume 13667, pp. 650–667. Available online: https://celebv-hq.github.io (accessed on 21 October 2024). [CrossRef]
157. Wuu, C.; Zheng, N.; Ardisson, S.; Bali, R.; Belko, D.; Brockmeyer, E.; Evans, L.; Godisart, T.; Ha, H.; Huang, X.; et al. Multiface: A Dataset for Neural Face Rendering. *arXiv* **2022**, arXiv:2207.11243.
158. Multiface Dataset. Available online: https://github.com/facebookresearch/multiface (accessed on 21 October 2024).
159. Wu, H.; Jia, J.; Xing, J.; Xu, H.; Wang, X.; Wang, J. MMFace4D: A Large-Scale Multi-Modal 4D Face Dataset for Audio-Driven 3D Face Animation. *arXiv* **2023**, arXiv:2303.09797.
160. MMFace4D Dataset. Available online: https://wuhaozhe.github.io/mmface4d (accessed on 21 October 2024).
161. Sung-Bin, K.; Chae-Yeon, L.; Son, G.; Hyun-Bin, O.; Ju, J.; Nam, S.; Oh, T.-H. MultiTalk: Enhancing 3D Talking Head Generation Across Languages with Multilingual Video Dataset. *arXiv* **2024**, arXiv:2406.14272.
162. MultiTalk Dataset. Available online: https://arxiv.org/pdf/2406.14272 (accessed on 21 October 2024).
163. 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]
164. Narvekar, N.D.; Karam, L.J. A No-Reference Image Blur Metric Based on the Cumulative Probability of Blur Detection (CPBD). *IEEE Trans. Image Process.* **2011**, *20*, 2678–2683. [CrossRef] [PubMed]
165. 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 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; IEEE: Salt Lake City, UT, USA, 2018; pp. 586–595. [CrossRef]
166. 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. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
167. Assael, Y.M.; Shillingford, B.; Whiteson, S.; de Freitas, N. LipNet: End-to-End Sentence-Level Lipreading. *arXiv* **2016**, arXiv:1611.01599.
168. Soukupova, T.; Cech, J. Eye blink detection using facial landmarks. In Proceedings of the 21st Computer Vision Winter Workshop, Laško, Slovenia, 3–5 February 2016; Volume 2.

169. Chen, L.; Cui, G.; Kou, Z.; Zheng, H.; Xu, C. What Comprises a Good Talking-Head Video Generation?: A Survey and Benchmark. *arXiv* **2020**, arXiv:2005.03201.
170. Siyao, L.; Yu, W.; Gu, T.; Lin, C.; Wang, Q.; Qian, C.; Loy, C.C.; Liu, Z. Bailando: 3D Dance Generation by Actor-Critic GPT with Choreographic Memory. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; IEEE: New Orleans, LA, USA, 2022; pp. 11040–11049. [CrossRef]