A Speech-Driven Talking Head based on a Two-Stage Generative Framework

Brayan Bernardo de Souza, Paula Dornhofer Paro Costa
Dept. of Computer Engineering and Automation (DCA)
School of Electrical and Computer Engineering
Universidade Estadual de Campinas (UNICAMP)
paulad@unicamp.br

Abstract

Speech-driven facial animation, a technique employing speech signals as input, aims to generate realistic and expressive talking head animations. Despite advancements in talking head synthesis methods, challenges persist in terms of achieving precise control, robust generalization, and adaptability to various scenarios and speaker characteristics. Additionally, the majority of existing approaches are primarily tailored for a restricted range of languages, with English being the predominant focus. This work introduces a novel two-stage framework for Brazilian Portuguese talking head generation, combining the strengths of Transformers and Generative Adversarial Networks (GANs). In the first stage, the transformer-based model extracts rich contextual information from the audio speech input, generating facial landmarks. In the second stage, we employ a GAN-based framework to translate the facial representations into photorealistic video frames. This framework separates the modeling of dynamic shape variations from the realistic appearance, partially addressing the challenge of generalization. Moreover, it becomes possible to assign multiple appearances to the same speaker by adjusting the trained weights of the second stage. Objective metrics were used to evaluate the synthesized facial speech, showing that it closely matches the ground-truth landmarks.

Speech synthesis - Audio driven - Talking head generation

1 Introduction

Expressive facial animation synthesis models, or talking heads, characterize a key technology for constructing embodied social interactive agents capable of enabling collaborative interaction and attributing trustworthiness to AI systems (Mattheyses and Verhelst, 2015).

In this context, deep learning generative modeling techniques have been successful in leveraging virtual talking heads capable of inspiring more natural and empathetic interaction through the synthesis of highly realistic and expressive facial animations by extracting the underlying patterns and features from large datasets of human faces (Sheng et al., 2022). However, state-of-the-art talking head synthesis approaches still grapple with limitations in controllability and generalization. While animation fidelity has improved, tailoring facial expressions and nuances to convey specific emotions or intentions remains challenging. Additionally, models often struggle to adapt to unseen scenarios or variations in speaker appearance and voice, hindering their real-world applicability. (Chen et al., 2020).

Moreover, most of the existing facial animation systems are designed for English or a few other languages, such as Chinese/Mandarin (Tao and Tan, 2004; Li et al., 2021; Lu et al., 2021), French (Dahmani et al., 2019) and German (Thies et al., 2020). This currently limits the applicability and accessibility of facial animation systems for speakers of other languages, especially those with different phonetic and prosodic features. Despite the hypothesis that models trained on large volumes of data in English could be satisfactorily adapted or fine-tuned for other languages, no studies address this issue in more depth, including perceptual assessments. The hypothesis that existing models trained on primarily English data might misinterpret lip movements and expressions for other languages, potentially leading to cultural misunderstandings, persists.

In this work, we present a videorealistic, speech-driven, image-based, Brazilian Portuguese talking head that was built from the training of a novel two-stage framework. The first stage of our framework consists of a *FaceFormer* model, initially proposed by Fan et al. (2022) to convert audio into 3D meshes that we adapted to generate 2D landmarks. The second stage of our framework adopts *vid2vid* model to synthesize photorealistic frames

of animation (Wang et al., 2018). By adopting this new arrangement that separates the modeling of dynamic variations of shape driven by speech (*FaceFormer*) from the modeling of the dynamic variations of appearance driven by shape (*vid2vid*), our framework addresses, albeit partially, the problem of generalization. With the appropriate design, it is possible, for example, to attribute multiple appearances for the same speaker simply by changing the trained weights of the second stage. It is also possible to make the same face talks in multiple languages, changing the trained weights of the first stage. Additionally, the facial landmarks, as first stage output, enhance interpretability, as they directly correspond to visible facial features, enabling intuitive understanding and manipulation.

To the best of our knowledge, our work builds the first neural deep learning-driven talking head for Brazilian Portuguese. In the following sections, we discuss related works and describe our methodology. As a work in progress, the present work does not include results from perceptual evaluation assessment, but objective metrics and links to synthetic videos are shared in Section 4.

2 Related Works

In recent years, there has been growing interest in using deep neural networks to effectively connect auditory and image-based signals. Many works try to generate speech-driven talking heads by directly mapping the speech to the talking head in an end-to-end style (Jamaludin et al., 2019; Zhou et al., 2019). On the other hand, other works utilize intermediate facial parameters to bridge the gap between audio and image (Suwajanakorn et al., 2017; Jalalifar et al., 2018). These facial parameters can be 3D meshes or landmarks. While 3D meshes provide detailed and volumetric representation, they require more computational resources and specialized equipment such as 3D scanners or depth sensors, making data collection more complex and time-consuming. Alternatively, landmarks are lightweight and can be easily obtained from 2D images or videos, making them widely accessible and applicable in various scenarios (Zhen et al., 2023). This work obtains inspiration from two-stage approaches that use landmarks as intermediate facial parameters.

The pioneering work by Suwajanakorn et al. (2017) utilized a time-delay e Long Short Term Memory (LSTM) to map standard Mel-frequency Cepstral Coefficients (MFCCs) representations of speech audio to lip shapes, aligning them with a specific set of 18 lip landmark points. From the lip landmark, a statistical three-step pipeline is employed to render realistic speech texture. Jalalifar et al. (2018) improved the quality of the output image with a simpler pipeline by introducing a Conditional GAN as the second stage (Goodfellow et al., 2020; Mirza and Osindero, 2014). To address the pixel jittering issue, Chen et al. (2019) enhanced the second stage with a novel proposed dynamically adjustable pixel-wise loss with an attention mechanism and a regression discriminator based on perceptual loss (Johnson et al., 2016). Additionally, the intermediate landmarks map 68 facial points, adding more face detail points such as the eyes, nose, and jaw.

Many works also focus on improving speech representation by adopting deep learning-based Automatic Speech Recognition (ASR), instead of only relying on hand-crafted features such as MFCC, to ensure robustness due to the different audio sources, accents, and noise. Sinha et al. (2020) and Das et al. (2020) utilize DeepSpeech, which uses recurrent neural network layers to model the temporal dependencies in the audio signal. Zhou et al. (2020) employed AutoVC, a voice conversion neural network, to learn disentangled speech content and identity features (Qian et al., 2019). Autoregressive Predictive Coding (APC) adopted by Lu et al. (2021), offers a powerful framework for learning speech representations in an unsupervised manner (Chung and Glass, 2020). It is worth mentioning *FaceFormer*, although it is a work focusing on 3D Meshes, it uses *wav2vec 2.0*, a Transformer-encoder-based network that employs unsupervised pre-training with contrastive learning to learn robust speech representations (Baevski et al., 2020).

The predominant choice of LSTM models for synthesizing facial landmarks from speech features has shifted to a variety of advanced deep learning techniques. Contemporary methodologies, including GANs, Convolutional Neural Networks (CNNs), Temporal Convolutional Network (TCNs) and Transformer-based models, have demonstrated significant efficacy in capturing intricate relationships between input features and corresponding facial landmarks (Sinha et al., 2020; Das et al., 2020; Yu et al., 2022).

GANs are commonly employed in the second stage to render landmarks into highly realistic images. The evolution of generator models in this

context has progressed from simple CNNs to more sophisticated architectures. Sinha et al. (2020) included attention mechanisms to focus on specific areas of the face for better detail generation. Lu et al. (2021) and Zheng et al. (2021) incorporate U-Net structures, an architecture known for its effectiveness in image segmentation tasks. Yu et al. (2022), inspired by Wang et al. (2018), utilized optical flow, which captures the motion between consecutive frames of a video. Zhong et al. (2023) employ SPADE layers to modulate the synthesis process with semantic information of the scene (Ronneberger et al., 2015; Park et al., 2019; Ilg et al., 2017).

In this study, we employ *FaceFormer* as the initial stage for its robustness in extracting speech features using *wav2vec 2.0* and its ability in managing long-range dependencies through attention mechanisms. For the second step, taking inspiration from (Yu et al., 2022) and (Wang et al., 2018), we use a GAN framework integrated with optical flow to facilitate the translation of landmarks into realistic images while maintaining temporal consistency.

3 Methodology

3.1 Dataset

The proposed method is trained on a subset of neutral speech videos from CH-Unicamp, a Brazilian Portuguese dataset featuring expressive speech (Costa, 2015). The aim is to first validate the methodology on neutral videos, which are simpler, before enhancing it to include emotional conditioning, thereby enabling use of the entire expressive dataset. These video clips were recorded under controlled conditions to facilitate synchronized audio and video capture. An actress performed various scripts, depicting everyday dialogues and encompassing all phonemes of the Brazilian Portuguese language.

The training dataset contains 124 video clips, while the valid and test dataset contains 13 video clips each. The total duration of all videos is approximately 15 minutes, averaging around 7 seconds per clip. The video and audio were recorded using an HD 1920×1080 pixels, NTSC 29.97 FPS digital video camera.

3.2 Data Preprocessing

Initially, frames were extracted from all videos at 30 frames per second and then subjected to center cropping and downsampling, resulting in a resolution of 256x256 pixels. This reduction was necessary due to the computational demands of training the second stage model, the *vid2vid* model. Subsequently, the *facealign* method was applied to each frame to extract 68 facial keypoints (Bulat and Tzimiropoulos, 2017). Additionally, the audio was extracted from the videos and downsampled to 16kHz to ensure compatibility with *wav2vec 2.0*, which is employed as an audio encoder (Baevski et al., 2020).

3.3 Architecture

As illustrated in Figure 1, the framework consists of two main components. The first is an audio-to-face representation, for which we adapted the output of the *FaceFormer* model implementation (Fan et al., 2022). The second component is a neural renderer, the *vid2vid* model implementation, which converts face representations into realistic speech frames (Wang et al., 2018).

The *FaceFormer* model utilizes a transformer encoder-decoder architecture to process raw audio data and produce a sequence of animated 3D face meshes (Vaswani et al., 2017; Fan et al., 2022). In our modification of the model, we altered the motion encoder dimensions to allow *FaceFormer* to produce 2D landmarks with dimensions of 68x2. This generation is dependent on the contextual information from the audio and the sequence of previously predicted facial landmarks.

The *FaceFormer* encoder utilizes a *wav2vec 2.0* model adapted to synchronize audio features with the predicted frames (Baevski et al., 2020). The *wav2vec 2.0* consists of three primary components: an audio feature extractor, a multi-layer transformer encoder, and a quantization module. The audio feature extractor employs a series of TCNs to transform raw waveform input into feature vectors. The transformer encoder, comprising a stack of self-attention and feed-forward layers, further refines the audio feature vectors into contextualized speech representations. The quantization module then discretizes the output from the TCNs into a finite set of speech units. To mitigate the differences in frequencies between audio (e.g., 16kHz) and video (e.g., 30 FPS) data, linear interpolation is implemented on the TCN output, resampling the audio features to match the video frequency.

The *FaceFormer* decoder includes three main components: a periodic positional encoding (PPE), a biased causal multi-head (MH) self-attention designed for generalizing to longer input sequences,

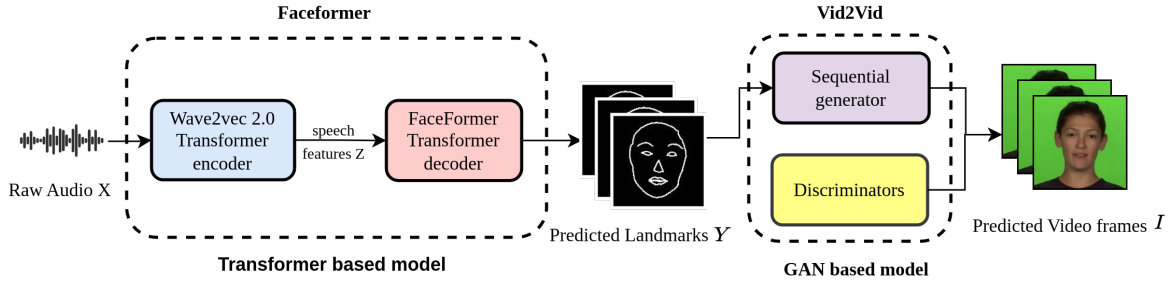

Figure 1: Framework Overview. The framework contains two models: i) *Faceformer*, a transformed-based model for audio-to-face representation; ii) *vid2vid*, a GAN-based model for final photorealistic frame construction.

and a biased cross-modal MH attention to synchronize audio-motion features. These modules are influenced by Attention with Linear Biases (ALiBi), which adapts the traditional Transformer decoder to enhance generalization capabilities (Press et al., 2021).

The *vid2vid* model is a GAN-based framework composed of multiple generators and discriminators, designed to convert a sequence of source video frames into a target sequence (Wang et al., 2018). The generator operates in a coarse-to-fine manner, progressively refining the generation process through hierarchical stages and incorporating optical flow networks to predict subsequent frames (Ilg et al., 2017). To combat the mode collapse issue prevalent in GAN training, two discriminators, Conditional Image Discriminator (CID) and Conditional Video Discriminator (CVD), are utilized (Ghosh et al., 2018; Tulyakov et al., 2018). CID aims to ensure each generated frame closely resembles the corresponding actual frame, while CVD focuses on maintaining the temporal dynamics of consecutive frames, considering the optical flow. This configuration allows the discriminators to assess both the individual frame quality and the coherent flow of the entire video sequence, identifying and penalizing any unnatural or abrupt variations.

3.4 Training

The models were trained separately, using the Adam optimizer (Kingma and Ba, 2014) with a fixed learning rate of 10^{-4} for *FaceFormer* and 2.10^{-4} for *vid2vid*. Both models were trained with a batch size of 1. The experiment was conducted on a Linux server equipped with an Nvidia V100 GPU, eight processor cores, and 32 GB of RAM. The *FaceFormer* model was trained 2560 epochs for approximately one week, with the encoder parameters fixed on the pre-trained *wav2vec 2.0* weights (Grosman, 2021). Meanwhile, *vid2vid* was trained for 120 epochs on both realistic and 2D-facial landmarks video frames, requiring about two weeks to complete.

4 Evaluation and Results

Examples of animations synthesized using our method can be seen at br-bernardo90.github.io/bpsdth.

Well-established methods in the field of computer vision were employed to evaluate the quality of the synthesized animation frames. These include the Structural Similarity Index (SSIM) (Wang et al., 2004), Frechet Inception Distance (FID) (Heusel et al., 2017), and Learned Perceptual Image Patch (LPIPS) (Zhang et al., 2018). FID relies on a pre-trained Inception network to extract and compare feature embeddings from both real and generated images. A lower FID score indicates higher image quality. SSIM provides a comprehensive analysis of two images by assessing their luminance similarity, contrast similarity, and structural similarity within their local neighborhoods. SSIM generates a score ranging from 0 to 1, with 1 denoting perfect similarity. LPIPS is an objective metric for quantifying the perceptual similarity between two images. It is designed to assess how similar two images appear in terms of human perception, with a higher score indicating greater dissimilarity and a lower score indicating higher similarity.

As a first approach to evaluating the proposed framework, we focused on studying the 2D landmark representation synthesized by the adapted *FaceFormer* architecture. To conduct the experiments, we fixed the model checkpoints of the second stage (*vid2vid*), and we varied its inputs (landmarks) to assessing if *FaceFormer* training is capable of learning efficient shape representations

of facial dynamics driven by audio. Finally, we completely removed the first stage of our pipeline and compared previous results with synthesized animation frames driven by 2D landmarks obtained from ground truth videos.

In the experiment, a k-fold cross-validation approach was adopted, with $k = 4$ and each subset comprising 13 test samples. This method partitioned the data into 'k' subsets, systematically using one subset for testing and the remaining data for training in each iteration. The choice of k-fold cross-validation was especially pertinent given the small dataset size, as it allowed for a more robust and thorough evaluation of the model's performance and generalizability across various data subsets. The Table 1 showcases the aggregate results from the k-fold cross-validation iterations, specifically capturing the mean (μ) and standard deviation (σ) of the objective evaluation metrics across different epochs. For each epoch, the mean score is computed from all 13 test samples within a single iteration. Subsequently, the means and standard deviations of these scores are calculated across all iterations for each epoch. This process offers a comprehensive view of the model's performance at various stages of training

Ep	FID ↓		LPIPS ↓		SSIM ↑	
	μ	σ	μ	σ	μ	σ
160	31.1	1.6	0.0576	0.0005	0.317	0.002
320	28.5	0.73	0.0567	0.0004	0.320	0.002
640	27.3	0.35	0.0561	0.0004	0.324	0.002
1280	26.8	0.12	0.0554	0.0003	0.328	0.001
2560	26.6	0.09	0.0552	0.0001	0.330	0.001
GT	**25.4**	0.07	**0.0450**	0.0001	**0.390**	0.001

Table 1: Objective scores were computed using synthesized and ground-truth 2D landmarks as input to the second stage of our pipeline. The arrows up indicate that higher is better, while the arrows down indicate that lower is better. We see that *FaceFormer* training successfully learns facial shape dynamics. With 2560 training epochs, we get landmark representations that result in scores close to those obtained by ground-truth representations. "Ep" stands for epochs. "GT" stands for Ground Truth.

The initial rows of Table 1 display a consistent decrease in FID and LPIPS scores over epochs, signifying an enhancement in image quality. Also, it demonstrates a corresponding increase in SSIM score over the epochs, further confirming improved image quality. These metrics collectively exhibit a positive trend, implying potential for even better results with extended training.

The final row of Table 1 presents the scores obtained when ground-truth landmarks are input to the second stage. Although the use of ground truth yields better photorealism in animations, the scores are comparatively close to those obtained using the fully synthetic pipeline.

5 Conclusion

To the best of our knowledge, our work builds the first neural deep learning-driven talking head for Brazilian Portuguese. We also present a novel two-stage arrangement adapted from existing models capable of delivering photorealistic animations, with an intermediate facial landmark representation that attributes interpretability and generalization aspects to the framework.

Among the limitations of our work, we emphasize that our models were trained with neutral speech only. The next steps include enhancing the framework to incorporate emotion conditioning.

Also, while recognizing the valuable insights offered by objective metrics like SSIM, LPIPS, and FID in quantifying visual fidelity, we readily acknowledge their limitations in comprehensively evaluating the quality of synthesized talking heads. These metrics excel at capturing pixel-level similarity, but the human perception of facial animation extends far beyond mere visual sharpness. Videorealism, for instance, encompasses subtleties in lighting, skin texture, and hair dynamics that defy reduction to single numerical scores. Similarly, cultural nuances in the expression through facial movements cannot be captured by objective metrics alone. Therefore, we plan to complement objective metrics with subjective evaluation by human observers.

Acknowledgements

This project was supported by the Ministry of Science, Technology, and Innovation of Brazil, with resources granted by Federal Law 8.248 of October 23, 1991, under the PPI-Softex. The project was coordinated by Softex and published as Intelligent agents for mobile platforms based on Cognitive Architecture technology [01245.013778/2020-21]. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior – Brasil (CAPES) – Finance Code 001. The authors are also with the Artificial Intelligence Lab., Recod.ai, Institute of Computing, UNICAMP.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Adrian Bulat and Georgios Tzimiropoulos. 2017. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE.

L. Chen, R. K. Maddox, Z. Duan, and C. Xu. 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7824–7833, Los Alamitos, CA, USA. IEEE Computer Society.

Lele Chen, Guofeng Cui, Ziyi Kou, Haitian Zheng, and Chenliang Xu. 2020. What comprises a good talking-head video generation? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

Yu-An Chung and James Glass. 2020. Generative pre-training for speech with autoregressive predictive coding. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3497–3501.

Paula Dornhofer Paro Costa. 2015. *Two-Dimensional Expressive Speech Animation*. Ph.D. thesis, Universidade Estadual de Campinas.

Sara Dahmani, Vincent Colotte, Valérian Girard, and Slim Ouni. 2019. Conditional variational auto-encoder for text-driven expressive audiovisual speech synthesis. In *INTERSPEECH 2019-20th Annual Conference of the International Speech Communication Association*.

Dipanjan Das, Sandika Biswas, Sanjana Sinha, and Brojeshwar Bhowmick. 2020. Speech-driven facial animation using cascaded gans for learning of motion and texture. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, page 408–424, Berlin, Heidelberg. Springer-Verlag.

Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. 2022. Faceformer: Speech-driven 3d facial animation with transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18749–18758. IEEE Computer Society.

Arnab Ghosh, Viveka Kulharia, Vinay P Namboodiri, Philip HS Torr, and Puneet K Dokania. 2018. Multi-agent diverse generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8513–8521.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

Jonatas Grosman. 2021. Fine-tuned XLSR-53 large model for speech recognition in Portuguese. `https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-portuguese`.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. 2017. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470.

Seyed Ali Jalalifar, Hosein Hasani, and Hamid Aghajan. 2018. Speech-driven facial reenactment using conditional generative adversarial networks.

Amir Jamaludin, Joon Son Chung, and Andrew Zisserman. 2019. You said that?: Synthesising talking faces from audio. *International Journal of Computer Vision*, 127(11–12):1767–1779.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision – ECCV 2016*, pages 694–711, Cham. Springer International Publishing.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lincheng Li, Suzhen Wang, Zhimeng Zhang, Yu Ding, Yixing Zheng, Xin Yu, and Changjie Fan. 2021. Write-a-speaker: Text-based emotional and rhythmic talking-head generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1911–1920.

Yuanxun Lu, Jinxiang Chai, and Xun Cao. 2021. Live speech portraits: real-time photorealistic talking-head animation. *ACM Transactions on Graphics (TOG)*, 40(6):1–17.

Wesley Mattheyses and Werner Verhelst. 2015. Audiovisual speech synthesis: An overview of the state-of-the-art. *Speech Communication*, 66:182–217.

Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346.

Ofir Press, Noah Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. 2019. AUTOVC: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*, pages 5210–5219. PMLR.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer International Publishing.

Changchong Sheng, Gangyao Kuang, Liang Bai, Chenping Hou, Yulan Guo, Xin Xu, Matti Pietikäinen, and Li Liu. 2022. Deep learning for visual speech analysis: A survey.

Sanjana Sinha, Sandika Biswas, and Brojeshwar Bhowmick. 2020. Identity-preserving realistic talking face generation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10.

Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4):1–13.

Jianhua Tao and Tieniu Tan. 2004. Emotional chinese talking head system. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 273–280.

Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. 2020. Neural voice puppetry: Audio-driven facial reenactment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 716–731. Springer.

Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. 2018. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. Video-to-video synthesis. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 1152–1164.

Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. 2004. Image quality assessment: From error visibility to structural similarity. *Trans. Img. Proc.*, 13(4):600–612.

Lingyun Yu, Hongtao Xie, and Yongdong Zhang. 2022. Multimodal learning for temporally coherent talking face generation with articulator synergy. *IEEE Transactions on Multimedia*, 24:2950–2962.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595.

Rui Zhen, Wenchao Song, Qiang He, Juan Cao, Lei Shi, and Jia Luo. 2023. Human-computer interaction system: A survey of talking-head generation. *Electronics*, 12(1):218.

Aihua Zheng, Feixia Zhu, Hao Zhu, Mandi Luo, and Ran He. 2021. Talking face generation via learning semantic and temporal synchronous landmarks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3682–3689.

Weizhi Zhong, Chaowei Fang, Yinqi Cai, Pengxu Wei, Gangming Zhao, Liang Lin, and Guanbin Li. 2023. Identity-preserving talking face generation with landmark and appearance priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. 2019. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9299–9306.

Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. 2020. Makelttalk: Speaker-aware talking-head animation. *ACM Trans. Graph.*, 39(6).