UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO

Brayan Bernardo de Souza

Speech-Driven 2D Facial Animation Based on a Two-Stage Generative Framework

Animação Facial 2D Guiada pela Fala Baseada em um Framework Generativo de Dois Estágios

Campinas
2024

Brayan Bernardo de Souza

Speech-Driven 2D Facial Animation Based on a Two-Stage Generative Framework

Animação Facial 2D Guiada pela Fala Baseada em um Framework Generativo de Dois Estágios

Dissertation presented to the School of Electrical and Computer Engineering of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Electrical Engineering, in the area of Computer Engineering.

Dissertação apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica, na área de Engenharia de Computação.

Orientadora: Profa. Dra. Paula Dornhofer Paro Costa

Este exemplar corresponde à versão final da dissertação defendida pelo aluno Brayan Bernardo de Souza, orientado pela Profa. Dra. Paula Dornhofer Paro Costa.

Campinas
2024

Ficha catalográfica
Universidade Estadual de Campinas (UNICAMP)
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva - CRB 8/5974

So89s

Souza, Brayan Bernardo de, 1990-
Speech-driven 2D facial animation based on a two-stage generative framework / Brayan Bernardo de Souza. – Campinas, SP : [s.n.], 2024.

Orientador: Paula Dornhofer Paro Costa.
Dissertação (mestrado) – Universidade Estadual de Campinas (UNICAMP), Faculdade de Engenharia Elétrica e de Computação.

1. Inteligência artificial. 2. Aprendizado de máquina. 3. Animação por computador. 4. Avatares. I. Costa, Paula Dornhofer Paro, 1978-. II. Universidade Estadual de Campinas (UNICAMP). Faculdade de Engenharia Elétrica e de Computação. III. Título.

Informações Complementares

Título em outro idioma: Animação facial 2D guiada pela fala baseada em um framework generativo de dois estágios
Palavras-chave em inglês:
Artificial intelligence
Machine learning
Computer animation
Avatars
Área de concentração: Engenharia de Computação
Titulação: Mestre em Engenharia Elétrica
Banca examinadora:
Paula Dornhofer Paro Costa [Orientador]
José Mario De Martino
Hélio Pedrini
Data de defesa: 24-05-2024
Programa de Pós-Graduação: Engenharia Elétrica

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/ 0009-0000-5401-59
- Currículo Lattes do autor: https://lattes.cnpq.br/9801217370172754

COMISSÃO JULGADORA - DISSERTAÇÃO DE MESTRADO

Candidato: Brayan Bernardo de Souza RA: 180116

Data da Defesa: 24 de maio de 2024

Título da Tese: "Animação facial 2D guiada pela fala baseada em um framework generativo de dois estágios"

Prof. Dr. Paula Dornhofer Paro Costa (Presidente)
Prof. Dr. José Mario De Martino
Prof. Dr. Hélio Pedrini

A ata de defesa, com as respectivas assinaturas dos membros da Comissão Julgadora, encontra-se no SIGA (Sistema de Fluxo de Dissertação/Tese) e na Secretaria de Pós-Graduação da Faculdade de Engenharia Elétrica e de Computação.

A minha amada esposa, Suh, e meu querido filho, Ian. Amo vocês.

A minha mãe, Claudia, e minhas irmãs, Sah e Ana, que tanto amo.

Agradecimentos

Agradeço profundamente a Professora Paula Dornhofer Paro Costa, minha orientadora, por me dar esta oportunidade e me guiar com muita paciência, carinho e inteligência por toda esta jornada até aqui. Agradeço a todos meus colegas de grupo de pesquisa, que com suas críticas e sugestões, me ajudaram a construir esta pesquisa. Por fim, agradeço aos professores da banca, Professores José Mario e Hélio Pedrini, por terem se prontificado a participarem da minha qualificação e agora da minha defesa. Obrigado pelo precioso tempo de vocês, todos as críticas e comentários foram muito bem vindos.

Agradeço a meu amigo, Vagner Inácio, famoso Condô! Por me acompanhar durante boa parte dessa saga, pelo apoio, pelos infinitos debates pós aula e claro, pelos socorros quando a Mimosa me deixava na mão no meio da rodovia! Agradeço também meu amigo Juliano Peres, o ilustre Torrinha! Por sempre me ajudar como pôde, pela camaradagem e por todo seu tempo me apoiando com as apresentações deste trabalho. Agradeço também a todos meus amigos que mesmo que não tenham participado de alguma forma mais direta, saibam que todos nossos encontros e conversas, mesmo que não tão freqüentes como gostaria, sempre me revigoraram me dando mais forças para continuar. Em especial, agradeço meus amigos de longa data, Douglas, Bruno, Jow, Ray, Viny, Osama, Luiz Otávio, Daniel, Ivanzão, Alce, por tudo até aqui, obrigado!

Agradeço imensamente a minha família, minha querida mãe, Claudia, minhas incríveis irmãs Sah e Ana, meus cunhados parceiros, João e Pedro, por todo o carinho, apoio nos momentos mais difíceis, por sempre me encorajarem a continuar, pela comida, festas e por todo o amor. Obrigado mais uma vez a minha mãe por sempre nos ajudar nos momentos mais atarefados, por ser essa vó fabulosa que todos amamos.

Finalmente, agradeço de todo meu coração a minha amada esposa Suelen Camilo, minha companheira de vida. Por todo apoio incondicional, por acreditar até quando eu duvidava, por me encantar todo dia com sua alegria e paixão pela vida. Passamos por momentos desafiadores, mas com muito amor e paciência conseguimos superar juntos. Obrigado por ser essa companheira e mãe incrível. E claro, agradeço ao grande fruto do nosso amor, nosso pequeno Ian, que enche meu coração de amor todos os dias com sua doçura e entusiasmo, e que mesmo quando não estamos juntos, me faz dar risada sozinho com mais freqüência do que poderia imaginar algum dia. Obrigado meus amores!

Este projeto foi apoiado pelo Ministério da Ciência, Tecnologia e Inovação, com recursos da Lei nº 8.248, de 23 de outubro de 1991, no âmbito do PPI-Softex, coordenado pela Softex e publicado como Agentes inteligentes para plataformas móveis baseados em tecnologia de Arquitetura Cognitiva (processo 01245.013778/2020-21). O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001. Os autores também agradecem ao Laboratório de Inteligência Artificial, Recod.ai, Instituto de Computação, UNICAMP, pela disponibilização dos recursos computacionais e todo o apoio da equipe.

It is human to feel imperfect. That was what you wanted, above all else: to be human. And now that is what you are. The imperfections - the weaknesses - the imprecisions - they are the very things which define humans as human. And which drive them to transcend their own failings.

— *Isaac Asimov*

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 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 is designed to be language-agnostic. The proof-of-concept model was trained using a Brazilian Portuguese audiovisual dataset, illustrating its initial application. The work is based on the hypothesis that similar effectiveness can be achieved for other languages when trained with respective language-specific datasets. 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. The results from generalization tests highlight the framework's potential for wide-ranging applications in creating talking head videos. By demonstrating an adept ability to generalize across languages, genders, and speech speeds, the framework sets a promising precedent for future advancements in the field. This paves the way for developing more flexible and efficient systems for synthesizing talking head videos.

Keywords: talking head, image-based animation, speech-driven

Resumo

A animação facial orientada por fala, uma técnica que emprega sinais de fala como entrada, tem como objetivo gerar animações realistas e expressivas de cabeças falantes. Apesar dos avanços nos métodos de síntese de falantes, persistem desafios em termos de obtenção de controle preciso, generalização robusta e adaptabilidade a vários cenários e características do locutor. Além disso, a maioria das abordagens existentes são implementadas para uma gama restrita de idiomas, sendo o inglês o idioma predominante. Este trabalho apresenta uma nova estrutura de dois estágios para a geração de animações facias 2D, combinando os pontos fortes das arquiteturas *Transformers* e das Redes Adversariais Generativas (em inglês, *Generative Adversarial Networks*, ou GANs). No primeiro estágio, o modelo baseado *Transformer* extrai informações contextuais ricas da entrada de fala de áudio, sintetizando pontos de referência faciais. Na segunda etapa, emprega-se uma modelagem baseada em GAN para traduzir as representações faciais em quadros de vídeo fotorrealistas. Esta estrutura separa a modelagem de variações dinâmicas de forma da aparência realista, abordando parcialmente o desafio da generalização. Além disso, torna-se possível atribuir múltiplas aparências ao mesmo alto-falante ajustando os pesos treinados do segundo estágio. Métricas objetivas foram usadas para avaliar a fala facial sintetizada, mostrando que elas se aproximas das métricas de vídeos reais gravados. Esta estrutura foi projetada para ser independente de linguagem. O modelo de prova de conceito foi treinado usando um conjunto de dados audiovisuais do português brasileiro, ilustrando sua aplicação inicial. O trabalho é baseado na hipótese de que este trabalho semelhante introduz uma nova estrutura de dois estágios para geração de falantes, e a eficácia pode ser alcançada para outras linguagens quando treinada com os respectivos conjuntos de dados específicos da linguagem. Os resultados dos testes de generalização destacam o potencial do abordagem proposta para aplicações abrangentes na criação de vídeos. Ao demonstrar uma capacidade hábil de generalizar entre idiomas, géneros e velocidades de fala, a estrutura estabelece um precedente promissor para avanços futuros neste campo. Isso abre caminho para o desenvolvimento de sistemas mais flexíveis e eficientes para sintetizar vídeos de animações faciais 2D.

Palavras-chave: animação facial 2D, animação baseada em imagem, orientada por fala.

List of Figures

List of Tables

List of Acronyms

2D Two-Dimensional

3D Three-Dimensional

ABD Average Blink Duration

ACD Average Color Difference

AD-NeRF Audio-Driven Neural Radiance Field

AE Autoencoder

ALiBi Attention with Linear Biases

APC Autoregressive predictive coding

BERT Bidirectional Encoder Representations for Transformers

CA centro acadêmico

CGAN Conditional GAN

CNN Convolutional Neural Network

CPBD Circular Pattern-Based Deviation

CSIM Cosine Similarity

DCGAN Deep Convolutional GAN

EAR Eye Aspect Ratio

FAN Face Alignment Network

FEEC Faculdade de Engenharia Elétrica e de Computação

FID Fréchet Inception Distance

FPS Frames per Second

GAN Generative Adversarial Network

GELU Gaussian Error Linear Unit

GRU Gated Recurrent Unit

HG Hourglass

HMM Hidden Markov Model

LMD Landmark Distance

LPIPS Learned Perceptual Image Patch Similarity

LSGAN Least Squares Generative Adversarial Network

LSTM Long Short-Term Memory

MFCC Mel-Frequency Cepstral Coefficients

MMD Maximum Mean Discrepancy

MM-LLM MultiModal Large Language Model

MSE Mean Squared Error

MSN Multiple Synergy Network

NeRF Neural Radiance Field

NLP Natural language processing

PE Positional Encoding

PPE Periodic Positional Encoding

PSNR Peak Signal-to-Noise Ratio

RMSE Root Mean Squared Error

RNN Recurrent Neural Network

SPADE Spatially-adaptive normalization

SSIM Structural Similarity Index

TCN Temporal Convolutional Network

UNICAMP Universidade Estadual de Campinas

VAE Variational Autoencoder

WGAN Wasserstein GAN

Summary

Chapter 1

Introduction

The continuous evolution of human-machine interfaces is rapidly driving us towards more natural and intuitive ways of interacting with computational devices. Remarkable progress in artificial intelligence, including natural language processing, visual computing, and speech processing, along with new interaction scenarios like smart homes, autonomous vehicles, and the metaverse, is fueling the creation of advanced virtual assistants. These assistants can conduct natural conversations in various contexts, offering a more intuitive user experience (DIEDERICH *et al.*, 2022; PUSHPAKUMAR *et al.*, 2023).

In Computer Graphics, "talking heads" are animated virtual human heads that mimic human speech, facial expressions, and lip movements (MATTHEYSES; VERHELST, 2015). They can have a wide range of applications across various fields. In education, they may serve as virtual assistants that enhance learning by adding visual engagement to spoken content. They can improve accessibility in communication for those with limited reading or writing skills, aiding people who are illiterate or less familiar with technology by breaking down information into simple, interactive explanations. In the entertainment sector, talking heads create dynamic, interactive virtual characters, offering a more personalized and immersive experience for users (WANG *et al.*, 2022).

However, the challenge of creating videorealistic facial animations, so realistic they could be mistaken for real video footage, underscores the complexity of human visual speech processing — a research frontier yet to be fully explored. Bridging the gap between synthesized and authentic human expressions and natural movements remains a persistent goal in this rapidly advancing field. This chapter starts by outlining the problems associated with synthesizing speech-driven talking heads, including the motivations behind this work and its challenges. We then discuss our approach, research question, and this study's primary goals and contributions. Lastly, we outline the text structure, providing a roadmap for the reader to navigate the nuances of our work.

1.1 Problem Definition

In speech-driven facial animation, making realistic and expressive talking heads using speech signals as the primary input presents a multifaceted set of challenges.

Human speech perception is complex and naturally bimodal, as evidenced, for example, by the well-known McGurk effect established several decades ago (MCGURK, 1976). It highlights that human speech perception relies not solely on auditory information but also significantly on visual cues, such as lip movements. When there is a discrepancy between what we hear (auditory information) and what we see (visual information), it can confuse or reduce the intelligibility of speech, which is why synchronizing the audio with the corresponding lip movements in videos is crucial. Discrepancies like this are especially noticeable in poorly dubbed movies or video calls with latency issues.

Despite significant advancements in synthesizing talking heads, the pursuit of achieving meticulous control, robust generalization across diverse scenarios, and adaptability to the unique characteristics of different speakers remains filled with challenges. In other words, we may be managing to drag ourselves out of the *uncanny valley*[1]. Still, experiments such as the revival of internationally known Brazilian singer Elis Regina in a Volkswagen Kombi advertisement aired in Brazil in 2023, "sing" us that there is still a challenging road ahead (Figure 1.1). The core issue lies in capturing the nuanced interplay between speech and facial movements, which varies significantly across languages and individual speakers. This variability demands a model capable of understanding and replicating the subtle dynamics and texture that define realistic and natural-looking facial animations. The quest for a solution to overcome these barriers drives the need for innovative approaches that leverage the latest advancements in machine learning techniques.

1.2 Motivation

The fusion of human communication nuances with technological interfaces has always been a frontier of digital innovation, particularly in the domain of speech-driven facial animation. As digital interactions evolve to be more immersive, realistic talking head animations capable of mimicking human-like expressions offer enhanced affinity in interactions (SEYMOUR *et al.*, 2021). This demand is not just driven by the entertainment industry but also by applications in virtual reality, telepresence, language learning platforms, and assistive technologies, underscoring the versatility and societal impact of advancements in this field (ZHEN *et al.*, 2023).

[1]The term "uncanny valley" was coined by Masahiro Mori, in 1970 (MORI *et al.*, 2012). Based on informal observations, Mori claimed that human affinity for robots and toys increases as realism increases. However, when a robot or toy resembles and imitates a real human being but falls short of perfection, the human observer feels disgusted. The term is also adopted to refer to avatars, animations, and synthetic videos.

Figure 1.1: Uncanny Valley example. Frames from a Brazilian advertisement employing facial animation techniques to replicate the renowned singer Elis Regina. The frames do not exhibit natural textures. Despite its quality, this advertisement generated significant controversy due to the use of artificial intelligence techniques to simulate the iconic singer, who has already passed away. Extracted from (VOLKSWAGEN, 2023).

This research is motivated by the limitations of existing speech-driven facial animation methods, particularly their lack of language diversity. The prominence of English in existing models marginalizes non-English speakers and limits the animations' culture. Brazilian Portuguese, the sixth most spoken language globally, represents a significant linguistic demographic currently underserved in the realm of talking head animations. By focusing on this language, our work seeks to contribute to a more inclusive and diverse technological landscape (PORTUGUESA, 2024).

Moreover, our framework utilizes advanced machine learning techniques, specifically Transformers and Generative Adversarial Networks (GANs), aligning with the core principles of MultiModal Large Language Models (MM-LLMs) (ZHANG *et al.*, 2024). This integration is essential for addressing the fundamental MM-LLMs challenge of effectively connecting models across different modalities—such as speech and video in our case—to enable collaborative inference. By efficiently combining audio inputs with visual outputs, our framework enhances the precision and adaptability of talking head synthesis and exemplifies the MM-LLMs approach to achieving robust generalization across different speaker characteristics.

1.3 Challenges

This research addresses the significant challenges in creating realistic and expressive talking head animations from speech signals, highlighting the complexity of blending linguistic details with visual elements. The key challenges addressed in this work are as follows:

- **Speech to Facial Landmarks**: Developing a model capable of capturing the intricate relationship between speech and facial movements by synthesizing facial landmarks that maintain temporal coherence.

- **Realism and Expressiveness**: Finding a balance between technical accuracy in lip synchronization and facial movements and the expressiveness inherent in natural human interactions.
- **Performance Evaluation**: Establishing objective metrics for comparing synthesized facial speech to ground-truth videos is vital to accurately measure the model's success.

1.4 Our Approach

When analyzing the deep learning methods of talking head generation, it is possible to observe two overall framework architectures employed in the synthesis process: end-to-end (one-stage) or two-stage (see Chapter 4).

In this work, we adopt a two-stage framework, more precisely, a novel speech-driven two-stage framework with the Brazilian Portuguese language as a case study. This proposed framework leverages the contextual processing abilities of Transformers to map raw audio to facial representations, combined with the generative power of GANs to transform these representations into photorealistic video frames, which is illustrated in Figure 1.2.

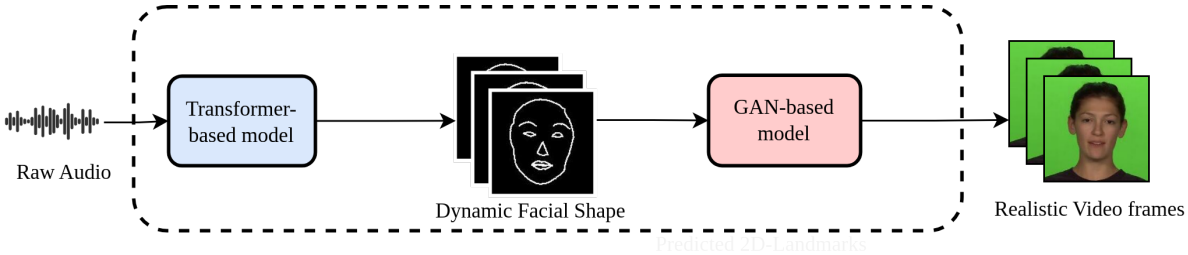


Figure 1.2: Framework Overview. The Transformer-based model captures a sequence of dynamic facial shapes from the raw audio. These facial shapes are transformed into realistic video frames by a GAN-based model.

The two-stage approach — separating the modeling of dynamic facial shape variations from the generation of realistic appearances — proposes a promising avenue towards enhancing the adaptability and generalization of talking head synthesis. Therefore, the practical realization of this potential framework and its empirical validation through objective metrics and real-world applicability forms the core problem our research seeks to solve.

The first stage is an audio-to-face representation, for which we employed the *FaceFormer* model implementation (FAN *et al.*, 2022). The second stage is a neural renderer, the *vid2vid* model implementation, which converts face representations into realistic visual-speech frames (WANG *et al.*, 2018).

FaceFormer modeling approach adopts a Transformer encoder-decoder architecture to process raw audio data and produce a sequence of animated Three-Dimensional

(3D) face meshes (VASWANI *et al.*, 2017; FAN *et al.*, 2022). In our model modification, we changed the motion encoder dimensions to allow *FaceFormer* to produce Two-Dimensional (2D) landmarks with dimensions of 68×2. This generation depends on the audio's contextual information and the sequence of previously predicted facial landmarks.

1.5 Research Questions

We have formulated a series of questions to encapsulate the motivation behind our research and outline the intended approach for achieving our goals in this dissertation.

- Considering that the *FaceFormer* model was originally designed to synthesize 3D meshes, our central research question is: **How does *FaceFormer* model perform in translating speech audio signals to realistic dynamic behavior of 2D facial landmarks?**
- **Can our framework synthesize high-quality talking heads with the available dataset volume?**
- **How far can our framework generalize to other speech agents and styles?**

1.6 Objectives

This research aims to advance the field of speech-driven facial animation by developing a novel framework that addresses the challenges of realism, expressiveness, and generalization, with Brazilian Portuguese as study case. To achieve this aim, the study is guided by the following specific objectives:

- **To explore advanced neural techniques for facial landmark synthesis from speech**: Investigate and implement cutting-edge neural network approaches to effectively capture and convert complex speech audio signals into dynamic facial expressions. This objective focuses on advancing the state-of-the-art in sequence-to-sequence speech processing techniques.
- **To generate adaptable and photorealistic talking head animations**: Develop and utilize neural architectures, potentially incorporating elements like Generative Adversarial Networks, to transform synthesized facial expressions into lifelike and adaptable video sequences. This involves refining the technology to support customization and versatility in visual outputs.
- **To assess and framework through rigorous evaluation**: Implement a robust evaluation strategy using a blend of objective metrics to quantify the performance and

generalization capabilities of the developed framework. This will facilitate continuous improvements and ensure the effectiveness of the framework in synthesizing realistic talk show videos.

- **To explore the model generalization across different speakers**: Explore the framework capacity to generalize across different speakers.

1.7 Contributions

The main contributions of this work are:

- **Novel Framework for Speech-Driven Facial Animation**: We have developed an innovative two-stage framework that uniquely integrates the capabilities of Transformer-based models and GANs. In the first stage, Transformer models are employed to meticulously extract facial landmarks from speech, capturing subtle expressions and nuances. The second stage utilizes GAN-based models to transform these landmarks into dynamic and photorealistic facial animations.

  In our refined approach, we adjust the output of the *FaceFormer* model, which originally generates a comprehensive face mesh consisting of five thousand 3D points. While this level of detail provides high fidelity, it comes with significant drawbacks: capturing such a detailed 3D mesh requires specialized equipment and extensive time commitments from actors, making the process expensive and logistically complex. Instead, we simplify this output to focus on 68 critical 2D facial landmarks. This transformation significantly streamlines the process by reducing the complexity and computational demand. To achieve this, we employ a specialized method that efficiently extracts these key landmarks from the photorealistic 2D images. This method involves identifying and isolating essential points that represent core facial features—such as the eyes, nose, mouth, and jawline—effectively capturing the expressive elements of the face with far fewer data points. This not only makes the process more efficient but also tailors the output to better suit real-world applications where simplicity and speed are valued alongside accuracy.

  The capacity of the *FaceFormer* to accurately isolate facial landmarks is essential not only for synthesing realistic facial animations but also offers significant value for broader applications through its adaptability and reusability. For instance, the output of the *FaceFormer* can be repurposed beyond digital animation to enhance interactions in robotic environments. Robots equipped with the capability to interpret and replicate human facial expressions can utilize the 2D landmarks generated by the *FaceFormer*.

- **Language Diversity and Method Generalization**: Our framework advances language diversity in talking head synthesis by effectively incorporating the diverse viseme set from CH-Unicamp dataset, which captures a wide range of phonetic nuances, to synthesizes a Brazilian Portuguese talking head. This methodological foundation is not only applicable to Brazilian Portuguese but also engineered to be replicated across any language that shares a similar diversity in viseme representation. By providing a reproducible recipe in our open-source code, we facilitate the adaptation of our framework to additional languages, broadening its utility and enhancing its linguistic versatility.

- **Open-Source Code and Demo Videos**: As part of our commitment to transparency and community engagement, we have made the source code and demo videos publicly available. This not only facilitates reproducibility and further research but also emphasizes our dedication to open science. The code can be viewed at <ai-unicamp.github.io/2StageTalkingHead>.

This work also resulted in the following publication:

- Brayan Bernardo and Paula Costa. 2024. A Speech-Driven Talking Head based on a Two-Stage Generative Framework. In Proceedings of the 16th International Conference on Computational Processing of Portuguese, pages 580–586, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics (ACL) (BERNARDO; COSTA, 2024).

1.8 Applications

The development of speech-driven talking heads holds a significant potential for a wide range of applications. By providing a more natural and engaging way to interact with machines, these technologies can transform user experiences across various fields. This section explores the practical applications of speech-driven talking heads, illustrating the broad impact of this research.

- **Customer Service Automation** One of the most immediate applications of talking heads is in the field of customer service. Virtual customer service agents can utilize talking head technology to provide users with a more personable and engaging interaction. This can significantly enhance user satisfaction and efficiency in resolving queries, especially in scenarios where visual and emotional engagement plays a crucial role in communication. As virtual assistants in smart homes and offices, talking heads can manage daily tasks and provide reminders or entertainment, all while maintaining a visually engaging and interactive presence that enhances

user experience. In sales, talking heads can be used for outreach by simulating real-life sales pitches and presentations. They provide a consistent and engaging representation of sales personnel, potentially increasing customer engagement and conversion rates.

- **Healthcare Communication** In healthcare, talking heads can be employed to simplify complex medical instructions or provide personalized patient support. By delivering information in a visually engaging and empathetic manner, these systems can improve patient understanding and compliance with medical guidelines, especially for those with reading difficulties or cognitive impairments. In elderly care, talking heads can provide personalized care by interacting with patients or the elderly in a compassionate and engaging manner. They can offer companionship, health reminders, and assist in daily routines.

- **Educational Tools** Talking heads can serve as personalized teaching assistants, offering one-on-one support to students. They can adapt to individual learning speeds and styles, providing explanations, feedback, and encouragement in a more interactive way. This can be particularly beneficial in language learning, where the accurate lip synchronization of talking heads can aid in better pronunciation and comprehension.

- **Entertainment and Media Production** Talking heads technology enhances the entertainment industry by enabling the creation of virtual celebrities and digital avatars that interact in real-time with audiences, hosting shows and performing at virtual concerts. This dynamic interaction boosts fan engagement and opens new possibilities for content delivery. In interactive media, talking heads serve as narrative drivers or hosts, adapting content to user interactions for a more personalized experience. In gaming, they enhance character development and interaction, providing more realistic and emotionally engaging characters that react to player decisions and progress, thereby improving immersion and gameplay experience.

- **Virtual Reality Environments** Virtual reality environments, designed to provide the most immersive experiences, can significantly benefit from the incorporation of talking heads. These avatars can act as guides, instructors, or companions in VR settings, enhancing the realism of virtual interactions. For example, in a VR educational program, a talking head could simulate a historical figure, providing first-person narratives and reacting to a learner's questions. This not only makes the educational content more engaging but also allows for a form of interaction that is closer to real-life conversations.

- **Training and Coaching Tools** Talking heads can assist in interview preparation by simulating various interview scenarios. They can provide feedback on responses,

body language, and speech, helping candidates practice and improve their interview techniques in a controlled environment. Talking heads can guide employees through new software, company policies, or training materials for corporate training and enablement, making the learning process more interactive and engaging. In executive coaching, talking heads can offer personalized advice and coaching, allowing for flexible scheduling and privacy, which are crucial for busy professionals.

- **Language and Cultural Preservation** Finally, talking heads can play a significant role in language and cultural preservation. By creating virtual avatars that speak less dominant languages, these technologies can help in teaching and preserving cultural heritage, making language learning accessible and engaging for new generations.

1.9 Organization

The text is organized as follows. Chapter 2 introduces core concepts about Transformers and GAN and discusses the current state of the art in talking head generation systems. This discussion focuses on deep learning approaches, particularly speech-driven two-stage architectures that utilize landmarks as intermediary representations. Chapter 3 details the resources and tools necessary for building our framework, including the dataset, landmark extraction method, objective evaluation methods, and computational resources. Chapter 4 presents the speech-driven, two-stage, landmark-based talking head generation framework proposed in this work. Chapter 5 presents the results obtained, including an ablation study, an exploratory study, and discussions. Chapter 6 offers some final remarks and directions for future work.

Chapter 2

Basic Concepts and Related Works

This chapter provides an overview of the main model architectures used in this work and reviews the state of the art in talking head generation systems, focusing on deep learning methods based on speech-driven two-stage architectures. Section 2.1 explains the basic concepts of the Transformer model. Section 2.2 introduces GAN. Section 2.3 provides a historical perspective, introducing the talking head's appearance, control mechanisms, and learning process developed in this work. Section 2.4 delves into further aspects of this work and describes state-of-the-art works related to the current work, summarizing state-of-the-art methods in terms of the model used, objective evaluation methods, and the datasets employed. Finally, in Section 2.5, we discuss how the current work relates to and contributes to existing approaches.

2.1 Transformers

Transformer architectures have significantly impacted the field of deep learning since their introduction by Vaswani *et al.* (2017). This innovative architecture has established new benchmarks in processing sequential data, especially in Natural language processing (NLP) tasks (LIN *et al.*, 2022). Unlike earlier models that predominantly used recurrent or convolutional layers, Transformers rely on attention mechanisms (BAHDANAU *et al.*, 2016). This approach has resulted in substantial enhancements in performance and training efficiency.

2.1.1 Encoder and Decoder

The standard Transformer model features an encoder-decoder structure designed to handle a wide range of sequence-to-sequence tasks, where the objective is to transform an input sequence into an output sequence.

The encoder maps a sequence of symbol representations expressed as $X = (x_1, ..., x_n)$, where n denotes the sequence length, into a series of continuous representations,

which are represented as $Z = (z_1, ..., z_n)$. In this phase, the encoder captures and encodes the contextual information present in the input sequence. Upon receiving the continuous representations Z, the decoder generates the output sequence denoted as $Y = (y_1, ..., y_n)$, symbol by symbol. A distinctive feature of the model during this phase is its autoregressive property. At each step of generating the output sequence, the decoder considers the continuous representations Z from the encoder and the symbols it has already generated. The autoregressive characteristic of the decoder ensures that each symbol in the output sequence is generated with a comprehensive understanding of the preceding elements, thereby maintaining consistency and context relevance in the output. Both the encoder and decoder utilize multiple layers, each consisting of self-attention mechanisms and feed-forward neural networks, to process the data.

2.1.2 Scaled Dot-Product Attention

The attention mechanism allows a neural network to focus on different parts of the input sequence when performing a task, akin to how humans pay attention to specific parts of an input when comprehending or responding. It helps the model to weigh and use the most relevant parts of the input data for making predictions or generating outputs. The Transformer attention mechanism is built by two components: the Scaled Dot-Product Attention and the Multi-Head Attention.

To understand the Scaled Dot-Product component, consider an input word sequence encoded into a set of vectors, typically through an embedding layer, as $X = x_1, ..., x_n$. For each word, three vectors are generated: a Query vector Q, a Key vector K, and a Value vector V. These vectors are produced by multiplying the word's embedding by respective matrices that are trained during the learning process, as illustrated below:

$$Q = XW^Q, K = XW^K, V = XW^V.$$

An attention score is computed for each Query-Key pair. This score determines how much focus the output element should put on each input part. The score is derived by taking the dot product of the query vector with the key vector of the respective word being scored. The scores are divided by the square root of the key vectors' dimension, which helps achieve more stable gradients. The results are passed through a softmax operation, determining how much each word will be expressed at the respective position. Finally, each value vector V is multiplied by the softmax scores, and then these weighted value vectors are summed up, producing the output of the self-attention layer. The process is shown visually in Figure 2.1 and described mathematically as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{2.1}$$

The self-attention layers in the decoder operate with a key difference compared to those in the encoder. To preserve the autoregressive property, they are restricted to only considering preceding elements in the output sequence. This restriction is achieved by masking subsequent positions with a value close to negative infinity, effectively excluding them from consideration during the softmax operation in the self-attention computation. Meanwhile, the "Encoder-Decoder Attention" layer functions as usual, with the unique aspect being that it generates its Queries matrix from the preceding layer while receiving the Keys and Values matrices from the final output of the encoder's layers.

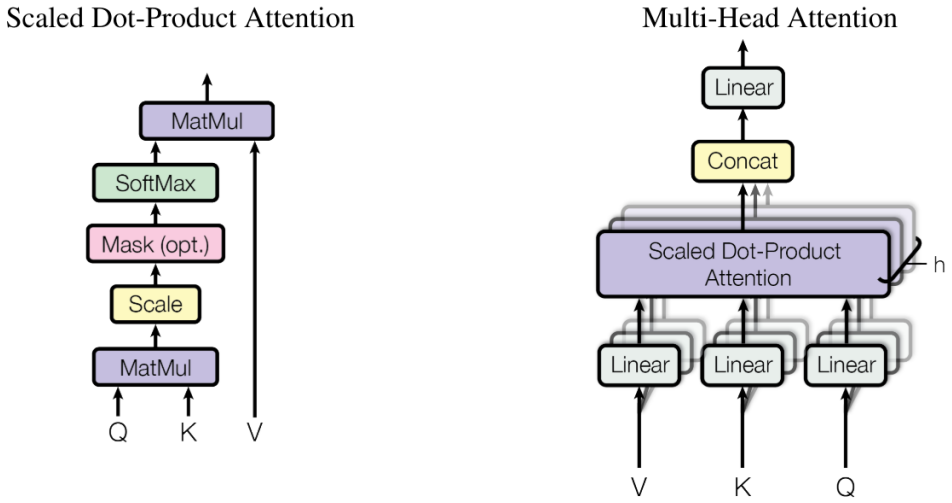


Figure 2.1: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention. Source: Images extracted from (VASWANI *et al.*, 2017)

2.1.3 Multi-Head Attention

The Attention Mechanism operates through h parallel heads, each employing distinct sets of Query/Key/Values weights. The different matrices give to attention layer multiple representation subspaces, also it expands the model's capacity to focus on different positions. The dimensions of the matrices are reduced according to the number of heads, d_{model}/h, ensuring that the computational cost remains comparable to that of a single-head attention with full dimensionality. Finally, the outputs of each self-attention head are concatenated and then multiplied by an additional weights matrix W_O. Figure 2.1 visually illustrates this process, which is defined mathematically as:

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \\ \text{where head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V). \end{aligned}$$

2.1.4 Positional Encoding

It is worth mentioning the Positional Encoding (PE) mechanism. In this work, the PE will be adapted to improve the positional information for speech synthesis (Section 4.3.2).

Unlike recurrent neural networks, Transformers process input sequences in parallel, which leads to the loss of positional information. PEs reintroduce this information, allowing the model to consider the position of each element in the sequence. PEs is added to the input embeddings before feeding them into the Transformer model. This addition enables the model to preserve the order of the sequence throughout the network and learn temporal or sequential relationships between the elements of the input sequence.

$$\begin{aligned} PE(pos, 2i) &= \sin\left(\frac{pos}{10000^{2i/d}}\right) \\ PE(pos, 2i+1) &= \cos\left(\frac{pos}{10000^{2i/d}}\right) \end{aligned} \tag{2.2}$$

where pos is the word position, i represents the dimension index and d the embedding dimension. The variable i varies from 0 to $d-1$. The sinusoidal functions alternate between sine for even indices and cosine for odd indices. By using this scheme, each dimension of the PE vector gets a unique sinusoidal wave based on its index i. This ensures that the Transformer model can distinguish different positions in the sequence and dimensions within the embeddings.

2.2 Generative Adversarial Networks

Upon their introduction, Generative Adversarial Networks GANs represented a significant advancement in generative modeling and machine learning. Initially introduced by Goodfellow *et al.* (2020), GANs are designed to map a specific distribution and generate new data with similar characteristics. A GAN comprises two neural networks, the generator and the discriminator, which are trained simultaneously through adversarial processes. The generator aims to produce data that is indistinguishable from real data, while the discriminator attempts to differentiate between the generator's fake data and true data. In this work, we utilize a GAN-based network to map the distribution of realistic speech video frames.

2.2.1 Generator

The generator network is responsible for creating data that mimics the real-world distribution. This data can be images, audio, text, or another data form. The generator in a GAN, typically structured as a deep neural network, is designed to create a probability distribution p_g over data x. This distribution is not explicitly provided in the form of

p_g; rather, it is defined implicitly through the generation of samples. Given a noise vector z, the generator G outputs a sample $G(z)$, designed to resemble real data x from the distribution p_{data}. The noise vector z comes from a predefined noise distribution p_z, typically Gaussian or uniform. The collection of all samples $G(z)$, created as z, is varied and represents the generated data distribution p_g. This is the distribution that G learns to approximate p_{data}. The training process aims to make p_g converge to p_{data} so that the generated samples become indistinguishable from actual data. Mathematically, the generator's operation is defined by the transformation:

$$\hat{x} = G(z; \theta_g), \quad z \sim p_z(z),$$

where θ_g symbolizes the generator's learnable parameters. The ultimate aim for G is to closely approximate the authentic data distribution $p_{data}(x)$ such that the discriminator cannot reliably distinguish between true and synthesized samples.

2.2.2 Discriminator

The discriminator D operates as a binary classification neural network that discerns the probability that a sample originates from the actual data rather than the synthetic data. Essentially, it acts as the adversary that the generator competes against. For each sample x, the discriminator outputs a probabilistic value $D(x; \theta_d)$, reflecting the likelihood that x is a genuine sample from the dataset, where θ_d represents the discriminator's parameters. The discriminator is optimized to identify both real and synthetic data accurately.

2.2.3 Adversarial Training

The training process in GANs is central to their functionality; it involves a game-theoretic scenario where the discriminator tries to maximize the probability of classifying the data correctly, and the generator aims to minimize this probability. This process continues until the generator produces outputs that are, in theory, indistinguishable from the real data.

The discriminator's goal is to accurately distinguish real data from fake data generated by the generator. It aims to maximize the probability that the generator seeks to minimize, effectively working to reduce the combined error on both real and fake data. The generator wants to maximize the probability that the discriminator incorrectly classifies its output. Considering both networks within the GAN framework, the general loss function is represented by:

$$\min_G \max_D \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]. \tag{2.3}$$

The training process of GANs is an iterative and competitive optimization that alternates between improving D and G. The discriminator is trained by feeding it batches of real data and batches of fake data generated by the generator. The discriminator adjusts its parameters to improve its classification accuracy. Once the discriminator is updated, it's the generator's turn. Here, the generator produces new data that aims to be classified as real by the discriminator. The generator then updates its parameters based on how well it tricks the discriminator. It essentially tries to increase the probability that the discriminator is making a mistake on its data.

2.2.4 Conditional Generation

The introduction of random noise as input in traditional GANs can yield impressive results in terms of generating new data. However, this approach cannot precisely control the generated output's characteristics. Addressing this limitation, Mirza e Osindero (2014) introduced an innovative method using segmentation maps as inputs in the GAN framework. The architecture of Conditional GAN (CGAN) involves both the generator and discriminator networks receiving additional input in the form of labels or other data, influencing the generation process. The primary advantage of segmentation maps is the added control level over the generated output. This control manifests in the network's ability to produce results that are realistic and accurately aligned with the input maps. The work presented in this document will guide the generation of realistic speech video frames based on simple facial landmarks.

Since their introduction in 2014 Goodfellow *et al.* (2020), GANs have experienced significant advancements, revolutionizing the fields of artificial intelligence and machine learning. Notable developments include the emergence of Deep Convolutional GANs (DCGANs) and PatchGAN for improved stability, image quality and generation efficiency (RADFORD *et al.*, 2016; ISOLA *et al.*, 2017). Innovations like Wasserstein GAN (WGAN) and gradient penalty methods have resolved initial challenges such as training instability and mode collapse. Continual advancements and a strong emphasis on ethical considerations, particularly in addressing issues like deepfakes, place GANs at the forefront of AI research, continually expanding the capabilities of generative modeling (GUI *et al.*, 2023a).

2.3 Talking Heads: A Historical Perspective

The synthesis of talking heads has evolved into a strong field of research since the pioneering work of Parke (1972). Beyond achieving photorealistic representations of the human face, advancements in speech-synchronized facial animation demand meticulous attention to diverse aspects. This includes precise lip synchronization with speech audio, faithful reproduction of speech articulatory movements specific to a target language, and coarticulation patterns that describe the dynamic interactions between these articulatory movements in the context of a connected speech. Additionally, these aspects extend to non-verbal signs, such as speech prosody, involving intonational head movements, and physiological factors like eye blinking.

When analyzing the history of different approaches do facial animation synthesis, it is possible to identify multiple classification dimensions. For example, considering the output facial appearance, talking heads can be broadly classified based on their facial appearance into two categories: 2D (image-based) and 3D (3D animations) (WANG *et al.*, 2022; ZHEN *et al.*, 2023). While 3D animations offer explicit controllable spatial information, they typically require specialized equipment, making data collection more complex and time-consuming. On the other hand, 2D models can be easily built from videos, making them more accessible and applicable. Furthermore, they can achieve a high level of photorealism due to their image-based properties. However, 2D models lack depth information relying on limited labeled image datasets, which can be challenging to capture the full range of facial movements. This work focuses on 2D synthesis methods.

Numerous approaches have emerged over the years to synthesize talking heads, initially emphasizing text-based visual speech modeling, followed by speech-driven, video-driven, and hybrid approaches (MATTHEYSES; VERHELST, 2015). This work focuses on speech-driven methods. Speech-driven visual speech synthesis involves generating a new visual speech signal based on an input auditory speech signal. These speech-driven systems predict the desired facial expressions by analyzing features extracted from the auditory input signal. The pioneering work of Bregler *et al.* (1997) introduced the Video Rewrite method, which involves editing the mouth movements of an existing video to match new speech. Video Rewrite generates new videos through two steps: analysis of a training database and synthesis of new visual speech. In the analysis stage, Video Rewrite automatically segments the audio track of the training database into phonemes. In the synthesis stage, the speech input is also segmented, and the method selects the mouth images in the database that most closely match the target phonemes. The resulting sequence of mouth images is integrated into the existing video.

Classifying visual speech models according to their synthesis strategy is also possible. Historically, facial animation approaches can be categorized into three main groups: Rule-based systems, Concatenative systems, and Statistical prediction (MATTHEYSES;

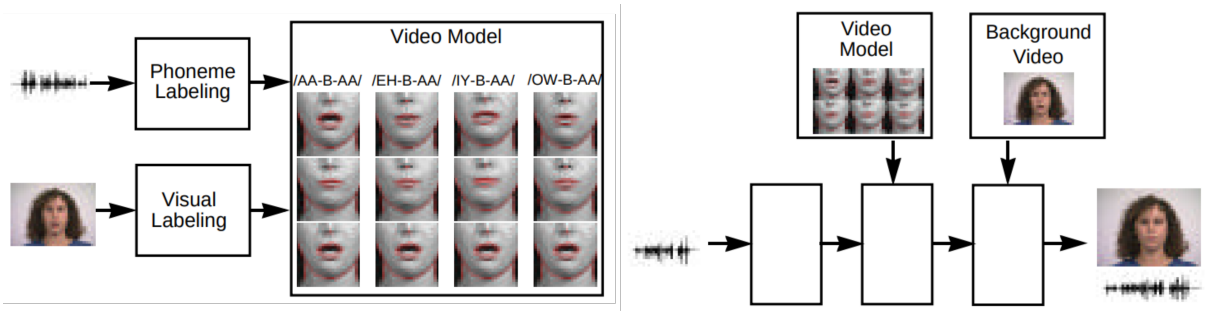


Figure 2.2: Overview of analysis and synthesis stage of Video Rewrite method. On the left is the analysis stage. It uses the audio track to segment the video into phonemes. On the right is the synthesis stage. It segments new audio and uses it to select phonemes from the video model. Source: Images extracted from (BREGLER *et al.*, 1997)

VERHELST, 2015).

Rule-based visual speech synthesis approaches employ predefined rules to anticipate output speech properties. This method, often called keyframe-based synthesis, directly predicts a few frames of the output video signal, typically positioned at the midpoint of each phoneme or viseme. Then, interpolation methods between predicted keyframes are used to ensure smooth and duration-synchronized visual speech signals, aiming to simulate visual coarticulation effects for realism. Various techniques have been employed for keyframe interpolation, such as morphing techniques with Scott *et al.* (1994), optical flow by Ezzat e Poggio (2000), and Radial Basis Functions by Noh e Neumann (2000).

Bregler *et al.* (1997), for instance, represented a Concatenative system. For this synthesis method, a speech synthesizer necessitates a database containing original speech recordings from a single speaker. When generating new speech, the system searches this database for segments that partially match the target phoneme sequence. These selected segments are then concatenated to produce the final synthetic speech signal.

However, despite the historical significance of rule-based and concatenative systems in facial animation, recent years have seen a dominance of statistical prediction approaches. This shift is attributed to advancements in deep learning methods and the availability of increased computational resources, leading to the emergence of a specific branch of deep neural-based methods. The text will continue focusing on statistical prediction.

2.3.1 Statistical prediction-based synthesis

Statistical prediction-based synthesis adopts machine learning techniques to build a mathematical model by analyzing a training dataset. This involves a training phase where the model establishes connections between observed features of original speech and the corresponding visual speech sequences. Following training, the model can predict visual

speech sequences based on unseen speech input. This approach combines the benefits of rule-based and concatenative synthesis by reusing observed articulations without explicit modeling, maintaining a small data footprint, and not storing original speech data post-training. However, a drawback is the need to parameterize the original speech data for model training, leading to the synthetic speech signal being regenerated from predicted parameter values, potentially resulting in degraded signal quality.

Brand (1999) suggested transferring visual speech modeling from a collected corpus to static images or photographs. This methodology employs recorded video as training data to construct a finite-state machine. Each state in the machine has an output probability distribution over facial configurations and their corresponding acoustic features. Coarticulation is modeled through a Hidden Markov Model (HMM) to predict facial configuration sequences. Cosker *et al.* (2004) investigated the correlation among speech, articulatory movements, and non-verbal signaling by combining visual parameters from an active appearance model with speech signal parameters represented by the standard Mel-Frequency Cepstral Coefficients (MFCC) (ABDUL; AL-TALABANI, 2022). These parameters are then used to train an HMM speech-driven synthesis model.

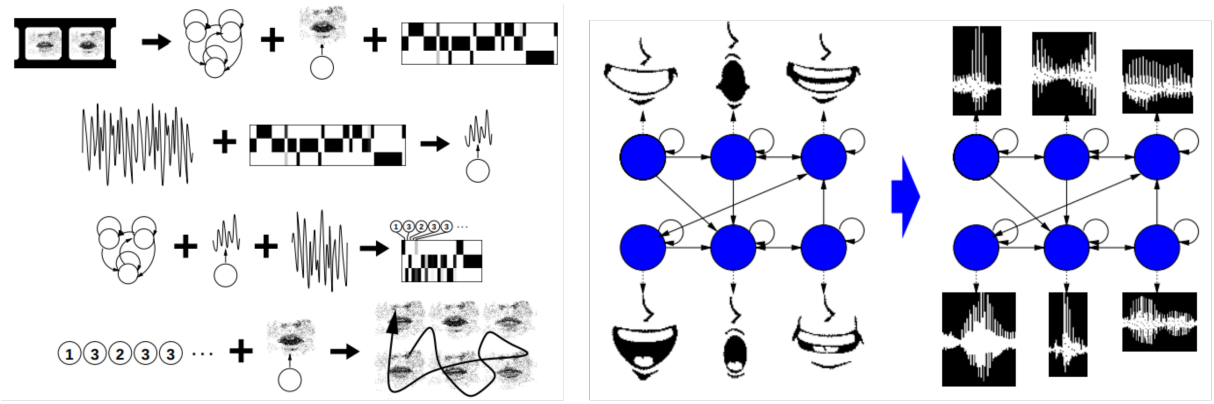


Figure 2.3: Overview of Voice Puppetry method. On the left, Schematic of the training, remapping, analysis, and synthesis steps. On the right, reuse the facial HMM's internal state machine in constructing the vocal HMM. Source: Images extracted from (BRAND, 1999)

HMM were the predominant approach for statistical synthesis. However, a significant advancement emerged when Fan *et al.* (2015) demonstrated the superiority of deep bidirectional Long Short-Term Memory (LSTM) networks over HMM-based methods. This work showcased the effectiveness of leveraging deep learning techniques, specifically bidirectional LSTM, in surpassing the performance of traditional HMM-based approaches in the domain of statistical synthesis. The problem is similar to the HMM synthesis case. In this context, the deep bidirectional LSTM is employed to model the trajectory of visual attributes. The "bidirectional" aspect indicates that the networks can acquire knowledge from the preceding context and subsequent context. This nuance holds particular importance in the visual modeling of coarticulation, given that the phonemes that follow it

impact the articulatory configuration of a specific phoneme. They also emphasize that, when compared to HMM synthesis, the training procedure for a recurrent LSTM network incurs a notably increased computational cost. Nevertheless, as the succeeding sections reveal advancements in deep learning models, this challenge will be addressed, leading to a surpassing of HMM in the domain of visual animation synthesis.

At the same time, another revolutionary class of networks emerged to address generative modeling tasks. GAN, a class of deep learning models introduced by Goodfellow *et al.* (2020), offer unique capabilities for generating realistic and expressive facial animations synchronized with speech signals. The adversarial training process pushes the generator to improve its ability to produce visually convincing speech-related facial movements, such as lip synchronization, facial expressions, and coarticulation effects. The adversarial training enables GAN to capture the subtle nuances and dynamic aspects of visual speech, enhancing the realism and naturalness of the synthesized content. Moreover, GANs can be integrated into the visual speech synthesis pipeline, for instance, they can be used to refine and enhance the visual quality of the output generated by other components of the system, such as LSTM-based models. This combination of LSTM networks for capturing temporal dependencies and GANs for high-fidelity visual generation creates a powerful synergy in the field of visual speech synthesis.

Deep learning research is advancing with new architectures that challenge traditional models. LSTM are being surpassed by Transformers (LIN *et al.*, 2022; VASWANI *et al.*, 2017). Unlike LSTM, which processes data sequentially, Transformers can handle entire data sequences in parallel. This capability accelerates training and captures long-range dependencies in the data more effectively. The self-attention mechanism weights the importance of different parts of the input data, regardless of their position. Additionally, Transformers scale more efficiently with volume data, thanks to their ability to manage larger context windows and their effectiveness in leveraging large-scale datasets, as demonstrated by models like GPT and BERT. Furthermore, Transformers are inherently more adaptable to a variety of tasks beyond text processing, such as image recognition and time-series analysis, making them a versatile tool in the AI toolkit.

In generative modeling, the dominance of GANs is contested by emerging models. Diffusion models have rapidly gained traction, celebrated for their ability to generate high-quality images while avoiding the training stability issues commonly associated with GANs, such as mode collapse (CROITORU *et al.*, 2023). These models operate through a process that gradually adds noise to data and then learns to reverse this noise addition to generate coherent outputs. Their success is highlighted in various applications, from creating detailed artworks to synthesizing realistic textures. Meanwhile, Variational Autoencoders (VAEs) continues to offer significant value, particularly in terms of controllability and interpretability of the generated images, by modeling the latent space through a normal distribution (KINGMA *et al.*, 2019). Although they may not achieve the crisp detail of

GAN-generated images, their strength lies in their ability to manipulate and understand the underlying factors of variations in data (GUI *et al.*, 2023b).

Some works were developed employing Brazilian Portuguese as a case study. In this context, it is notable to cite De Martino *et al.* (2006), where a 3D rule-based talking head generator combines a non-linear transition function strategy with the identification of visemes that are dependent on phonetic context. This approach defines a viseme not just as the visual representation of a single speech segment but also considers the sequence that includes the visemes before and after it. In Martino (2013), this idea is adapted to a 2D model, using Radial Basis Functions for smoother transitions between facial expressions. Subsequently, enhancing this approach, (COSTA, 2015) incorporated expressive speech face modeling based on the Ortony, Clore, and Collins (OCC) model of emotions. This work is also notable by the construct of the CH-Unicamp dataset, which is employed in the present work. More recently, Jesus Filho (2021) proposed a facial animation synthesis system leveraging Hidden Markov Models, with context-dependent phonemes and audio as inputs, producing facial 2D landmarks as outputs. Additionally, Reis (2020) suggested modifications to a GAN-based model to achieve expressive synthesis, utilizing facial 2D landmarks as inputs and generating photorealistic images as outputs. The current work aims to build upon these foundational studies to advance a Brazilian Portuguese speech-driven photorealistic talking head generation system.

The rapid advancement of deep learning technology has provided technical support and promoted the robust development of talking-head video generation methods. In summary, the recent surge in deep learning technologies, including advancements in architectures, generative models, computational resources, and dataset availability, has catalyzed the creation of sophisticated and realistic talking-head video generation methods. Among these advancements, MM-LLMs stand out by integrating capabilities from various domains, such as natural language processing and computer vision, to foster seamless intermodal interactions (ZHANG *et al.*, 2024). The integration of generative models with MM-LLMs represents an exciting frontier. MM-LLMs, which excel in handling and synthesizing information across different modalities, can be enhanced by generative models' capabilities in generating visually compelling outputs from textual or auditory inputs. This synergy could lead to more sophisticated systems capable of tasks such as generating video from text descriptions or improving the realism and detail of visual content produced in response to multimodal prompts. This integration allows for more effective synthesis of talking-head videos that are not only realistic but also capable of adapting to a wide range of linguistic inputs and visual contexts, thus pushing the boundaries of what is achievable in automated video content generation.

2.4 Deep Facial Animation Synthesis

When analyzing the deep learning methods of talking head generation, it's possible to observe two overall framework architectures employed in the synthesis process: end-to-end (one-stage) or two-stage. End-to-end approaches aim for a direct mapping from audio to video frames. Speech2Vid was one of the initial projects investigating a single-step approach using four subnetworks. It includes an audio encoder that extracts features from the audio, an identity encoder that identifies features from a reference image, and an image decoder that creates images by combining speech and identity features. These subnetworks work together like Autoencoder (AE), trained with L1 reconstruction loss. Additionally, it uses a pre-trained deblurring Convolutional Neural Network (CNN) as a post-process to enhance the quality of the images. Inspired by the Neural Radiance Field (NeRF) breakthrough, Guo *et al.* (2021) introduced the Audio-Driven Neural Radiance Field (AD-NeRF) for Talking Head generation. AD-NeRF uses audio features from DeepSpeech as a condition, learning to transform these audio features into dynamic neural radiance fields for creating talking face visuals. Unlike other models, AD-NeRF captures both the head and upper body by learning two separate neural radiance fields, providing a more comprehensive approach to video speech generation.

Two-stage architectures for video generation from audio typically follow a dual-step process: first, converting audio input into facial parameters, and second, transforming these parameters into video frames (SHENG *et al.*, 2024). This approach, which will be the focus of our discussion, involves first understanding how audio cues are mapped onto facial movements and expressions and then how these facial parameters are used to generate corresponding video frames. The upcoming sections will delve into the details of this method and highlight various studies that have successfully utilized this technique.

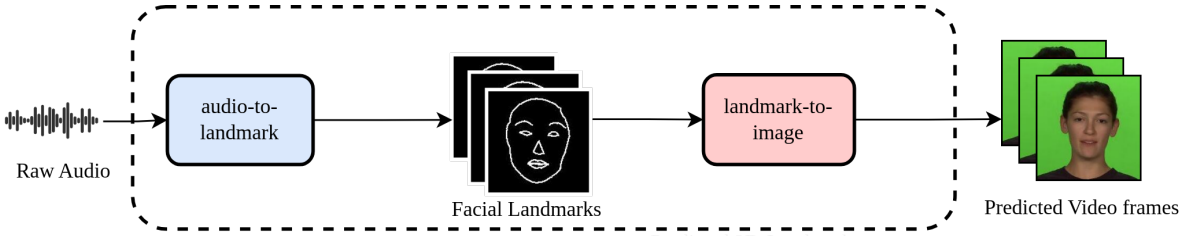


Figure 2.4: Overview of Two-Stage Landmark-Based Methods: Initially, the audio-to-landmark network maps the raw audio to a sequence of corresponding facial landmarks. Subsequently, the landmark-to-image network renders each facial landmark into a photorealistic video frame.

2.4.1 Two-Stage Methods

Based on the data type of the facial parameters, two-stage methods in talking head synthesis can be categorized into Landmark-based, Coefficient-based, or Vertex-based methods:

- Landmark-based methods in talking head generation involve using facial landmarks as a key component in synthesizing realistic and expressive facial animations. Facial landmarks are specific points on the face that correspond to distinct features, such as the corners of the eyes, the tip of the nose, and the corners of the mouth. These methods leverage these landmarks' spatial information to drive the animation synthesis process. This work specifically concentrates on two-stage landmark-based methods.

- Coefficient-based methods typically involve representing facial expressions or features using a set of coefficients. These coefficients serve as numerical values that capture the characteristics of facial movements or expressions. The idea is to encode the essential information about facial dynamics into a compact set of coefficients, which can then be used to drive the animation synthesis process.

- Vertex-based methods in the context of computer graphics generally refer to techniques that operate on the vertices (corners or points) of a 3D model. In the realm of talking head generation, this could involve manipulating the positions of vertices in a 3D facial mesh to create facial animations. Vertex-based methods allow for detailed control over facial expressions, capturing fine-grained movements at the level of individual vertices. They can be employed in the synthesis of both 2D and 3D talking head animations. The complexity of managing a large number of vertices in a facial mesh can pose computational challenges. Achieving natural-looking animations may require sophisticated algorithms to ensure smooth transitions between expressions (SHENG *et al.*, 2024).

The strategy of first mapping audio to high-level structures, such as facial landmarks, before generating video frames conditioned on these landmarks can prevent the capture of false correlations between audiovisual signals unrelated to the speech content (CHEN *et al.*, 2019). Moreover, employing independent models facilitates the transfer of lip movements from other speakers to different specific target identities (JALALIFAR *et al.*, 2018). Additionally, employing facial landmarks improves interpretability, as they directly correlate with observable facial features, thereby facilitating intuitive comprehension and manipulation of aspects like eye blinking and head movement (SINHA *et al.*, 2020; LU *et al.*, 2021).

2.4.2 Two-Stage Landmark-Based Methods

This section aims to define the state of the art of two-stage landmark-based talking head synthesis methods. The first column of Table 2.1 highlights the most relevant papers we review in this section. In the remaining columns, we summarize some aspects of their synthesis approaches that are going to be discussed, such as target identity, face region, head motion, and eye blinking.

Table 2.1: Summary of synthesized talking head target aspects presented in Sections 2.4.2 "S" stands for Specific-person. "A" stands for Arbitrary.

Reference	**Target Identity**	**Face Region**	**Head Motion**	**Blink eyes**
Suwajanakorn *et al.* (2017)	S	Mouth	X	X
Jalalifar *et al.* (2018)	S	Mouth		
Chen *et al.* (2019)	A	Face		
Sinha *et al.* (2020)	A	Face		X
Das *et al.* (2020)	A	Face		X
Zhou *et al.* (2020)	A	Face	X	
Lu *et al.* (2021)	S	Face/ Upper Body	X	
Zheng *et al.* (2021)	A	Face		
Yu *et al.* (2022)	S	Lip/Jaw		
Zhong *et al.* (2023)	A	Lip/Jaw		
This work	S	Face	X	

The pioneering work of Suwajanakorn *et al.* (2017) mapped the speech audio and mouth shape representation through a time-delay LSTM, matching it with a specific set of 18 landmark points outlining the contours of both the outer and inner lips. The audio input is represented by audio features extracted using standard MFCC. A three-step pipeline from the mouth landmark is employed to render realistic speech texture. First, a frame selection algorithm was employed to identify and process target video frames that closely matched the landmark, integrating them with a teeth proxy derived from the target video to synthesize a highly detailed mouth region. Secondly, even though the goal is to synthesize just the mouth area and reuse the rest, it was noted that if Obama pauses his speech and the head keeps moving, it looks unnatural. To address this, a dynamic programming algorithm was implemented to synchronize audio with visual pauses, avoiding head movements during periods of silence. Finally, a jaw correction method was devised for frame composition to adjust the new motion, and a Laplacian pyramids technique was applied to blend all the last steps. While this approach produces convincing videos with accurate lip synchronization, it requires a substantial 17 hours of Obama speech video training, a considerable duration when compared to other works utilizing HMM.

Jalalifar *et al.* (2018) introduced the LSTM + CGAN architecture, where audio

MFCC features are input into a bidirectional LSTM, followed by two hidden layers to map 8 points lip landmarks. These landmarks are then used as input to condition the CGAN for synthesizing realistic frames. The author emphasizes that the two stages are nearly independent, allowing the transfer of lip movements from other speakers. A simple affine transformation is sufficient to align the source and target facial landmarks. Even though it was designed without specific attention to teeth details, CGAN demonstrated promising results with a simpler pipeline. The study utilized two hours of Obama speech videos to achieve the desired quality.

Chen *et al.* (2019) introduced a novel LSTM + Convolutional-Recurrent Neural Network (RNN) structure, with a specific emphasis on the correlation between adjacent frames during the rendering stage. The facial landmarks map 68 facial points, adding more facial detail points such as eyes, nose, and jaw. Given the emerging topic of attention mechanism (GUO *et al.*, 2022), Chen incorporated it into the network to enhance robustness against visual variations and noisy audio conditions. Pixel jittering may not be apparent in single-image generation, but it becomes a significant issue for video generation, given that humans are sensitive to any temporal discontinuities. This issue was addressed with two components. A proposed novel dynamically adjustable pixel-wise loss with an attention mechanism; and a regression discriminator based on the perceptual loss (JOHNSON *et al.*, 2016). The framework models were trained separately on the GRID dataset (COOKE *et al.*, 2006), comprising 1000 short videos spoken by 33 different speakers, totaling 27 hours. The method effectively learns facial movement representations and synthesizes a talking head based on an audio speech and a single face source image.

Facial landmarks for different subjects contain individual-specific facial attributes, such as distinct face structures, sizes, shapes, and diverse head positions. Speech-driven lip movements for a given audio segment are independent of these variations. To ensure landmark prediction invariance to these factors, Sinha *et al.* (2020) proposed a novel three-step landmark prediction method. Instead of using audio MFCC features, DeepSpeech features are introduced (HANNUN *et al.*, 2014) to guarantee robustness due to the different audio sources, accents, and noise. Firstly, a convolutional encoder-decoder architecture with temporal loss is trained to map canonical 68-point facial landmarks from the DeepSpeech features. Secondly, eye blinks are imposed on the facial landmarks to add realism. An LSTM learns to predict the eye landmarks from a noise vector. Finally, the canonical facial landmarks with blinking eyes are retargeted to the person-specific facial landmarks. To achieve realistic image rendering, Least Squares Generative Adversarial Network (LSGAN) is employed to generate facial texture from person-specific facial landmarks, incorporating an attention mechanism to preserve identity-related texture. The model is trained on the GRID and TCD-TIMIT datasets (HARTE; GILLEN, 2015). TCD-TIMIT, comprising 62 speakers, contains much more phonetic variability than the GRID dataset.

Following the same three-step pipeline idea of (SINHA *et al.*, 2020), Das *et*

al. (2020) employed a GAN to map facial landmarks from DeepSpeech features. The model-agnostic meta-learning (MAML) approach (FINN *et al.*, 2017) is then utilized to train the rendering GAN, enabling rapid adaptation to an unknown face at inference time using only a few images. Compared with transfer learning, fine-tuning the meta-learned network requires a significantly smaller requirement than fine-tuning based on transfer learning. The model is trained on the TCD-TIMIT dataset.

In the pursuit of a more realistic talking head, Zhou *et al.* (2020) are concerned about facial expressions and head motions. AutoVC (QIAN *et al.*, 2019), a voice style transfer algorithm, is employed to learn disentangled speech content and identity features. The predicted facial landmarks are derived from the combination of landmarks from two models. The first, a content animation component, utilizes an LSTM-based encoder to map the voice conversation features to facial landmarks with a neutral style. The second, a speaker-aware component, captures more specific features not so related to oral speech as facial expressions and head motion. They observed that generating cohesive head motions and facial expressions necessitates capturing longer temporal dependencies than the speech content animation module. So, a LSTM-based encoder followed by a self-attention network is adopted to compute the face landmarks. To translate the facial landmarks into realistic images, it was employed a U-Net architecture (RONNEBERGER *et al.*, 2015). The landmarks generator components were separately trained with different datasets focusing on your specific goals. The content animation component train is conducted on the VCTK corpus (VEAUX *et al.*, 2016), which includes 109 English speakers with various accents. For the speaker-aware component, to learn head motion and facial expressions, and for the rendering model, the VoxCeleb2 dataset (CHUNG *et al.*, 2018) was selected due to the variety of speakers.

Similar to (ZHOU *et al.*, 2020), Lu *et al.* (2021) also extracted audio features and processes them with two modules in parallel to obtain separate features for mouth and head/upper-body motion. The first stage involves an Autoregressive predictive coding (APC) network (CHUNG; GLASS, 2020), which extracts deep audio features and utilizes a manifold projection to map these features to the target person's speech space. In the second stage, mouth and head features are learned independently using a time-delay LSTM and a multi-dimensional Gaussian inspired by (OORD *et al.*, 2016). These features are then combined to synthesize facial and upper-body landmarks. In the final stage, the landmarks and a candidate image set are inputted into a U-Net architecture training in a GAN style to generate photorealistic renderings. For training, different datasets are employed for each model. To train the APC was used the Mandarin Chinese part of the Common Voice dataset (ARDILA *et al.*, 2020), the authors observe that the system continues to perform effectively in other languages, attributed to the model's ability to learn high-level and semantic information. To train the landmark generator, 30 minutes of Obama's speech were used, and for the rendering network, 3 minutes of the specific

target person were used. The authors emphasize that this work is the first capable of live generation. The entire system operates with an inference time of approximately 27.4 ms, achieving over 30 Frames per Second (FPS) with a latency of 300 ms.

Zheng *et al.* (2021) proposed a talking face generation approach incorporating word semantic supervision and cross-modal temporal synchronization for landmark learning. A word detector is trained to identify a word based on sequential landmarks and employed as a classification loss. This ensures that the predicted landmarks closely reflect the content of the corresponding audio words. To address temporal consistency, a cross-modal synchronization, inspired by the concept in (BELGHAZI *et al.*, 2018), is implemented to enforce coherence between adjacent audio segments, helping in the smooth transition of landmarks. To synthesize realistic frames, a GAN is used with a U-Net serving as the generator and a reconstruction loss to optimize it. It is trained on the LRW dataset (CHUNG; ZISSERMAN, 2017) to demonstrate the model generalization and then evaluated using the GRID dataset.

Yu *et al.* (2022) presented a novel landmark prediction model, Multiple Synergy Network (MSN), aimed at enhancing the accuracy of landmark prediction by integrating multimodal inputs (audio and text) and considering jaw movements to ensure synergy between articulators, specifically lips and jaw. MFCC and corresponding text inputs are fed into their respective encoders, constructed using a Temporal Convolutional Network (TCN) (BAI *et al.*, 2018). In contrast to traditional recurrent architectures like LSTM, TCN demonstrates significantly longer memory when processing sequential data. The encoder outputs are then input into a multilayered bidirectional Gated Recurrent Unit (GRU) network for mouth-jaw landmark prediction (CHO *et al.*, 2014). A GAN-based model is proposed for the video synthesis stage with a generator and two discriminators. Adopting a coarse-to-fine-grained style, the generator utilizes an encoder-decoder with residual blocks to generate the hallucinated target frame. Simultaneously, another encoder-decoder with self-attention modules employs FlowNet2 to predict optical flow (ILG *et al.*, 2017). The outputs from these processes are merged to produce the final frame. The model is trained on 1 hour of Obama speech videos and to test generalization, it was employed the FaceForensics(RöSSLER *et al.*, 2018). It's worth mentioning that this model can generalize to other targets if its dataset is available.

To address the generic person target problem, Zhong *et al.* (2023) proposed a transformer-based landmark generator and a two-module GAN with Spatially-adaptive normalization (SPADE) layers to modulate the encoded features. The transformer-based generator utilizes MFCC audio features, reference landmarks, and pose prior landmarks as input to predict lip and jaw landmarks. These predicted landmarks are then combined with pose prior landmarks to generate the target landmarks. The GAN-based renderer incorporates alignment and translate modules. The alignment module takes multiple reference images and their landmarks as input to obtain motion fields. These motion fields

are then used to warp the reference images and their features to the target head pose and expression. The translation module, assisted by MFCC features, warped images, and features from the alignment module, translates the predicted landmarks to generate the target face image. The model is trained on LRS2 dataset (CHUNG *et al.*, 2017) em tested on LRS3 dataset (CHUNG; ZISSERMAN, 2017) to evaluate the generalization.

Although not based on landmarks, the work of Fan *et al.* (2022) is noteworthy. *FaceFormer* is a transformer-based model dedicated to creating a sequence of animated 3D face meshes from speech audio, utilizing a vertex-based approach. It employs *wav2vec 2.0* for speech feature extraction and introduces a transformer decoder with biased self-attention layers to convert these features into a sequence of face meshes.

2.4.3 First-Stage Models

Considering the works discussed in the previous section, Table 2.2 summarizes the details of the first stage model, which involves audio-to-landmark translation.

Table 2.2: Summary of first stage framework aspects presented in Sections 2.4.2. "Reg" stands for regularization. "Adv" stands for adversarial. "Rec" stands for reconstruction. "NLL" stands for Negative Log-likelihood. "GL" stands for Graph Laplace.

Reference	**Speech Feature**	**Feature to Landmark**	**Landmark**	**Loss**
Suwajanakorn *et al.* (2017)	MFCC	LSTM	[18,2]	L2
Jalalifar *et al.* (2018)	MFCC	LSTM	[8,2]	L2
Chen *et al.* (2019)	MFCC	LSTM	[68,2]	L2
Sinha *et al.* (2020)	DeepSpeech	AE/CNN	[68,2]	L2,Reg,MMD
Das *et al.* (2020)	DeepSpeech	GAN/CNN	[68,2]	L2,Reg,Adv,MMD
Zhou *et al.* (2020)	AutoVC	GAN/Att-LSTM	[68,2]	L2,GL,Adv
Lu *et al.* (2021)	APC	LSTM/MGauss	[73,3]	L1,L2,NLL
Zheng *et al.* (2021)	MFCC	LSTM	[68,2]	Rec,C,Reg
Yu *et al.* (2022)	MFCC	TCN-GRU	[33,2]	RMSE
Zhong *et al.* (2023)	MFCC	Transformer-Enc	[33,2]	L1,Reg
This work	wave2vec2.0	Transformer	[68,2]	GAN

The table is organized into four columns detailing the distinctive components of each framework:

- **Speech Feature**: This column lists the audio feature extraction techniques. Most studies employ MFCC, which is a traditional signal processing method providing a static representation of the spectral features of sound. The calculation involves breaking the audio signal into short frames and applying a sequential mathematical pipeline that includes applying the Fourier transform, mapping the powers of the spectrum obtained to the Mel scale, and then computing the log of the powers and the discrete cosine transform. Two studies have applied DeepSpeech, which uses deep learning techniques to convert spoken language into text. One work has used

AutoVC, a deep learning-based method focused on transforming the speaker identity in audio signals, emphasizing voice conversion. Finally, another has employed APC, focusing on learning representations from audio data in an unsupervised manner by predicting future audio samples to understand temporal structures within the data.

- **Feature to Landmark**: This column details the computational models and architectures used to process speech features and produce corresponding landmarks. The LSTM network, a type of recurrent neural network, features in numerous studies, highlighting its suitability for sequential data. Other architectures include GANs, favored for generating highly realistic landmarks, and Transformers, known for their efficient training and nuanced context capture via self-attention mechanisms. Additionally, various combinations of AEs are employed for dimensional reduction and denoising, CNNs for feature extraction and local pattern recognition, and attention mechanisms to focus on pertinent features, thereby enhancing contextual understanding.

- **Landmark**: This column indicates the number of landmarks or key points used to animate the talking head, which varies across studies. The landmarks are typically represented in pairs, with the first number representing the count of landmarks and the second the dimensionality of the landmark space.

- **Loss**: The last column specifies the loss functions used to train the models, which are critical for guiding the learning process toward accurate landmark generation. Common loss functions include L1 and L2 norms, which measure prediction errors, Regularization Loss, where L1 or L2 is used not only on the current target but on the past synthesized data, and more specialized losses like Maximum Mean Discrepancy (MMD), a difference between features, Adversarial Loss, from the adversarial training process, Reconstruction Loss, which compares the original image to the reconstructed image, Root Mean Squared Error (RMSE), measures the average deviation, among others (GRETTON *et al.*, 2012). For more details about the loss function, please refer to Terven *et al.* (2023).

2.4.4 Second-Stage Models

Table 2.3 summarizes the second stage model details, focusing on landmark-to-image rendering, as related to the works discussed in Section 2.4.2.

Due to the predominance of GANs, the table is organized based on the following aspects:

- **Generator**: This column describes the computational models and architectures that generate synthetic realistic images according to the corresponding landmarks. process the speech features and generate corresponding landmarks. GANs are commonly

Table 2.3: Summary of second stage framework aspects presented in Sections 2.4.2. "C" stands for classification. "Adv" stands for adversarial. "RD" stands for regression-based discriminator. "LS" stands for least squares. "PG" stands for PatchGAN. "Reg" stands for regularization. "P" stands for perceptual. "Rec" stands for reconstruction. "FM" stands for feature matching. "W" stands for warped.

Reference	**Generator**	**Discriminator**	**Loss Funciton**
Suwajanakorn *et al.* (2017)	Statistical pipeline	-	-
Jalalifar *et al.* (2018)	GAN/CNN	C	Adv
Chen *et al.* (2019)	GAN/CNN-glsRNN	C,RD	Pixel-wise,Adv
Sinha *et al.* (2020)	GAN/Att	PG	Pixel Intensity,LS,Reg
Das *et al.* (2020)	GAN/Att	C	Rec,Adv,P
Zhou *et al.* (2020)	U-Net	-	L1,P
Lu *et al.* (2021)	U-Net/GAN	PG	Adv,Color,P,FM
Zheng *et al.* (2021)	U-Net/GAN	C	L1,Adv
Yu *et al.* (2022)	U-Net/GAN	PG	Adv,Image,Video,Flow,P
Zhong *et al.* (2023)	SPADE/AE	-	P,Rec,Style,FM,W
This work	GAN	PG	

employed in landmark-to-image models. The evolution of generator models in this context has progressed from simple CNNs to more sophisticated architectures that can incorporate U-Net structures, an architecture known for its effectiveness in image segmentation tasks, and attention mechanisms, which focus on specific areas of the image for better detail generation, optical flow, which is used to capture the motion between consecutive frames of a video, and SPADE layers, to modulate the synthesis process with semantic information of the scene.

- **Discriminator**: This column details the discriminators utilized to assess the synthesized data generated by the generator model. Consolidated in the field, PatchGAN discriminators are tailored to evaluate local patches within an image, providing a more detailed and fine-grained assessment than conventional discriminators. Typically, two types of discriminators are employed with this technique: a classifier to analyze the authenticity of the image and a video discriminator used on the latest synthesized images to ensure temporal coherence.

- **Loss**: Many loss functions have been employed to achieve realistic facial synthesis. These include Perceptual Loss, which measures the perceptual similarity between generated and real images, Reconstruction Loss, aimed at reproducing the input faithfully, Style Loss, focusing on capturing artistic style, Matching Features to ensure consistency in feature representation, Adversarial Loss for enhancing realism through the adversarial training process, Pixel-size Loss for preserving spatial dimensions, Pixel Intensity Loss for maintaining accurate color representation, among others.

2.4.5 Evaluation

Chen *et al.* (2020), in an extensive review of the literature on quantitative metrics of talking-head generative models, reveals four essential characteristics that high-quality synthesized talking-head videos should possess: Visual Quality, Identity Preservation, Lip Synchronization, and Natural and Spontaneous Motion. Table 2.4 encapsulates a synthesis of objective evaluation methods utilized in the studies discussed in Section 2.4.2, categorized according to these modalities.

Table 2.4: Summary of objective evaluation methods presented in Section 2.4.2

Reference	**Visual Quality**	**Identity Preservation**	**Audio-visual Synchronization**	**Natural Motion**
Suwajanakorn *et al.* (2017)	-	-	-	
Jalalifar *et al.* (2018)	-	-	-	
Chen *et al.* (2019)	PSNR,SSIM		LMD	
Sinha *et al.* (2020)	PSNR,SSIM CPBD		LMD	EAR
Das *et al.* (2020)	PSNR,SSIM CPBD	FaceNet, ACD	LMD,AVo,AVc	EAR
Zhou *et al.* (2020)			LMD	D-L,D-V, D-Rot/Pos
Lu *et al.* (2021)	PSNR,SSIM LPIPS			D-L,D-V, D-Rot/Pos
Zheng *et al.* (2021)	PSNR,SSIM		LMD	
Yu *et al.* (2022)	PSNR,SSIM		LMD	
Zhong *et al.* (2023)	PSNR,SSIM, FID,LPIPS	CSIM	LMD	
This work	SSIM,FID, LPIPS		LMD	EAR

- **Visual Quality**: Unlike still image synthesis, video generation demands seamless transitions between frames to avoid any perception of inconsistency, such as temporal discontinuities and subtle distortions, which viewers can readily detect. For evaluation, the following metrics are used: Peak Signal-to-Noise Ratio (PSNR), quantifying the ratio of signal power to noise; Structural Similarity Index (SSIM), evaluating structural similarities between generated and real images; Circular Pattern-Based Deviation (CPBD), assessing circular patterns in images; Learned Perceptual Image Patch Similarity (LPIPS), capturing perceptual differences; and Fréchet Inception Distance (FID).

- **Identity Preservation**: The synthesized video must maintain the individual's identity, as viewers are particularly attuned to changes in perceived identity. For evaluation, the following metrics are used: Cosine Similarity (CSIM), comparing

identity vectors extracted by a face recognition network; Average Color Difference (ACD), quantifying color disparities; and similarity between FaceNet features for reference identity image and the predicted frames.

- **Audio-visual Synchronization**: A significant challenge in talking-head generation lies in ensuring that the visual dynamics, including facial and lip movements, are in sync with the audio modality, such as speech or corresponding landmarks, given that viewers are sensitive to even minor discrepancies between facial movements and the accompanying audio. For evaluation, the following metrics are used: Landmark Distance (LMD), measuring the distance between predicted and ground truth facial landmarks; and AV Offset and AV confidence produced by Syncnet, a deep learning model designed to asses audio-visual synchronization (CHUNG; ZISSERMAN, 2016).

- **Natural Motion**: In natural speech, individuals exhibit involuntary movements like head nods, blinks, or various facial expressions, conveying non-verbal information that is critical for listeners to understand the spoken content fully. For evaluation, the following metrics are used: Head motion is evaluated using D-L (Landmark Distance), D-V (Landmark Velocity Difference), and D-Rot/Pos (Head Rotation and Position Difference). Eye blinking is assessed using Average Blink Duration (ABD), calculated by the number of consecutive frames from the start to the end of a blink, and this is compared with a dataset of natural human blinks.

These diverse metrics reflect a comprehensive approach to objectively evaluating the performance of talking head generation models across various dimensions.

2.4.6 Datasets

Tables 2.2 and 2.6 present a summary of the datasets used related to works discussed in Section 2.4.2

Research in talking head generation increasingly relies on expansive datasets for model training and evaluation. The scope and complexity of these datasets have multiplied significantly. For instance, the number of subjects and the total duration of recordings in the LRS2 and LRS3 datasets have increased by orders of magnitude compared to the GRID dataset, including many accents and linguistic subtleties across different speakers. Moreover, there has been a shift from datasets comprising speech videos recorded in controlled indoor environments to datasets that include any form of talking head speech sourced from television or online platforms. This expansion has introduced a rich diversity in content features, such as varied lighting conditions and camera angles. These advancements contribute substantially to enhancing the model's capabilities, enabling them to produce high-fidelity and contextually relevant animated sequences adaptable to a wide variety of scenarios and speaker profiles.

Table 2.5: Summary of the datasets employed on the works discussed in Section 2.4.2.

Reference	**Name**
Suwajanakorn *et al.* (2017)	17hrs Obama speech
Jalalifar *et al.* (2018)	2hrs Obama speech
Chen *et al.* (2019)	GRID,LRW
Sinha *et al.* (2020)	GRID,TCD-TIMIT
Das *et al.* (2020)	TCD-TIMIT
Zhou *et al.* (2020)	VCTK,6hrs Obama speech,VoxCeleb2
Lu *et al.* (2021)	Mandarin Common Voice, 20m Obama speech
Zheng *et al.* (2021)	GRID,LRW
Yu *et al.* (2022)	1h Trump/Obama speech,FaceForensics
Zhong *et al.* (2023)	LRS2,LRS3
This work	CH-Unicamp

Table 2.6: Dataset details.

Year	**Name**	**Lang.**	**Hrs**	**Subj.**	**Env.**	**Description**
2006	GRID	EN	27.5	33	Lab	High-quality video recordings of spoken sentences.
2015	TCD-TIMIT	EN	11.1	62	Lab	Speakers reading sentences from two camera angles.
2015	CH-Unicamp	PT-BR	0.25	1	Lab	An actress performed everyday dialogues.
2017	LRW	EN	173	1k+	Wild	Variations of words spoken by numerous speakers.
2017	VCTK	EN	44	109	Lab	Speech data uttered by English speakers with various accents.
2018	LRS2	EN	224	500+	Wild	Spoken sentences from BBC television.
2018	LRS3	EN	438	5k+	Wild	Spoken sentences from TED and TEDx videos.
2018	VoxCeleb2	EN	2.4k	6k+	Wild	Utterances of celebrities from YouTube videos.
2019	FaceForensics	EN	5.7	1k	Wild	Youtube videos manipulated with face manipulation methods.
2020	Common Voice	EN	26	889	Lab	Volunteers who record sample sentences with a microphone and review recordings of other users.

2.5 Concluding Remarks

In the previous sections, we followed the historical perspective, understood the main concepts of talking head generation methods, saw examples of landmark-based methods, and finally analyzed these works by detailing aspects such as the first and second-stage architectures, objective metrics applied to evaluate, and datasets employed to train and validate the models. Now, we discuss these aspects and present our methodology choices for the work detailed in this document.

First Stage

Deep learning-based speech feature extraction techniques have emerged as compelling alternatives to traditional MFCC, eliminating the need for handcrafted features and providing more flexibility and adaptability in various audio processing applications.

The predominant choice of LSTM models in landmark synthesis has given way to a range of advanced deep-learning techniques. Contemporary methods like CNNs, GANs, TCN, and Transformer-based models have proven highly effective in discerning the complex relationships between input features and corresponding facial landmarks.

The representation of facial landmarks has evolved over time, transitioning from an initial mapping of lips using 18 points to a more comprehensive and widely adopted standard of 68 mapping points on the facial frame. Increasing the number of landmark points in facial mapping significantly contributes to the quality of synthesized talking head videos, enhancing the realism and expressiveness of the generated animations.

Moreover, facial landmark synthesis methods have expanded their range of loss functions, incorporating diverse measures. This diversified approach reflects a nuanced understanding of challenges in landmark synthesis, leveraging specific functions to address temporal dependencies, distribution matching, input reproduction, prevention of overfitting, realism enhancement, and accuracy assessment between predicted and ground truth landmarks.

We have selected the implementation of the *FaceFormer* as our primary stage architecture (FAN *et al.*, 2022). Drawing inspiration from Sinha *et al.* (2020), Zhou *et al.* (2020), we recognize that a non-MFCC approach for speech feature extraction can enhance robustness. *FaceFormer* employs *wav2vec2.0*, which utilizes TCN and the Transformer encoder to extract deep features (BAEVSKI *et al.*, 2020). To transform these speech features into landmarks, the adapted *FaceFormer* Transformer-based decoder provides significant advantages over LSTM. These include the capability to handle long-range dependencies, the incorporation of PE, the use of attention mechanisms, and ease of training (ISLAM *et al.*, 2024).

Second Stage

Integrating new mechanisms into GANs collectively enhances their performance and versatility, generating high-quality and realistic outputs. The strategy of evaluating local patches using PatchGAN has shown significant results. Additionally, the video discriminator taking in count the past frames assess temporal coherence. Adopting various loss functions underscores the multifaceted nature of facial realism, necessitating a comprehensive and tailored approach to address different image aspects.

Taking inspiration from (YU *et al.*, 2022), this work will employs a GAN framework equipped with optical flow and PatchGAN image and video discriminators to translate landmarks into high-quality, realistic images while maintaining temporal coherence. The *vid2vid* implementation includes all these features (WANG *et al.*, 2018).

Evaluation and Dataset

A relevant aspect of the reviewed works is how they were evaluated. Objective metrics are evolving from pixel-based functions to deep feature-based functions, aiming for an assessment closer to human perception. In addition, other aspects of natural motion are evaluated, such as head movements and eye blinking.

In evaluating image quality, this study utilizes SSIM, a traditional metric, and modern deep-feature-based metrics, namely LPIPS and FID. Lip synchronization is assessed using the established LMD metric. While other measures like Av and LSRD (Lip Synchronization Relative Distance), proposed by Chen *et al.* (2020), provide a more perceptual analysis of lip synchronization from the images, these deep learning methods are tailored for English-speaking videos. Adapting them for Portuguese videos would require additional efforts to retrain the models. The eye blinking frequency will be used as the metric to assess natural and spontaneous motion. However, the head motion will not be evaluated due to the limited expressive movements in the dataset used. Finally, the criterion of Identity Preservation will not be part of our evaluation as our model focuses on synthesizing a specific individual.

Despite the extensive hours, varied subjects, and uncontrolled "wild" conditions, a common limitation of the datasets reviewed is their exclusive focus on English. This study centers on the Portuguese language and will employ the CH-Unicamp dataset tailored to this linguistic context.

Chapter 3

Materials and Methods

This chapter details the necessary resources for building our framework, including the dataset, computational resources, landmark extractor, and evaluation methods.

3.1 Dataset

The proposed method is trained on a subset of videos from CH-Unicamp, a Brazilian Portuguese dataset featuring expressive speech (COSTA, 2015). Our research group built this dataset in the context of previous works, attending to all the ethical requirements for research in Brazil. High-quality recordings, diversity of visemes, and emotional features characterize it.

The 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 actress first enunciated the text during the recording sessions with a neutral expression. Then, she performed the text according to a specified emotional state, following some guidelines inspired by the Ortony, Clore and Collins (OCC) model of emotions.

In this work, we aimed to validate the methodology on neutral videos, which are more straightforward, before enhancing it to include emotional conditioning, thereby enabling the use of the entire expressive dataset. The training dataset we used contains 124 video clips, while the validation and test datasets contain 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.

Figure 3.1: Example of video frames from the dataset.

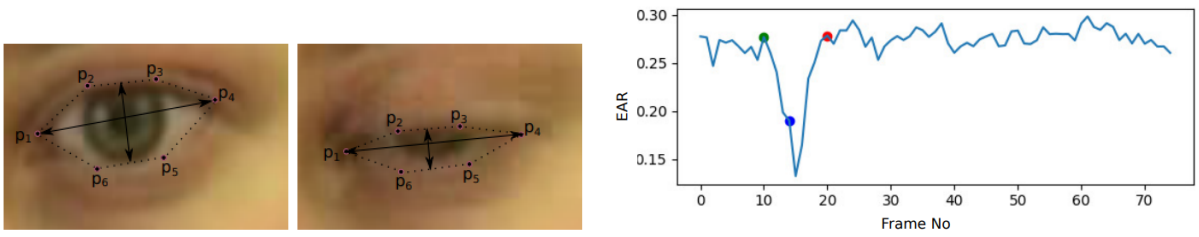
3.2 Objective Evaluation Methods

A set of popular objective methods in the computer vision area were employed to evaluate our results' quality objectively. While having limitations, objective metrics provide consistent, reproducible, and quantifiable results. This quantification allows researchers to assess improvements or declines in performance in a standardized manner, facilitating benchmarking and comparisons between different models or approaches over time.

3.2.1 Eye Blink

As discussed in Section 2.4.5, eye blinks represent a natural and spontaneous facial motion that is important for achieving realistic talking heads. To determine the occurrence of eye blinks, Cech e Soukupova (2016) introduced a blink detector based on the Eye Aspect Ratio (EAR). Given facial landmarks (we extracted using Face Alignment Network (FAN)), the EAR is calculated for each frame based on the distance between the upper and lower eye landmarks. The average human blink rate is 0.4 blinks per second, so a number close to this is considered good (DAS *et al.*, 2020). The EAR calculus is illustrated by Figure 5.1 and defined as:

$$\text{EAR} = \frac{\|p2 - p6\| + \|p3 - p5\|}{\|p1 - p4\|}.$$



(a) (b)

Figure 3.2: On the left, landmarks are used for EAR calculation. On the right, a chart shows the EAR score over the video frames. The space between the green and red dots indicates the occurrence of a blink. The blue dot represents the blinking eye threshold. Extracted from (CECH; SOUKUPOVA, 2016)

After computing the EAR to all frames, the approach counts the blinks by detecting deep drops in the EAR; see Figure 5.1. To detect significant drops, two constants

are defined: the blinking eye threshold, set at 0.6, which indicates a blink, and the blinking duration, set at six frames, which determines the number of consecutive frames for which the EAR must fall below the threshold. These values were established after extensive testing to ensure optimal performance.

3.2.2 Structural Similarity Index Measure

The SSIM is based on the premise that the human visual system is highly adapted for extracting structural information from a visual scene. Thus, a measure of structural similarity should provide a good approximation of perceived image quality (WANG *et al.*, 2004). Unlike traditional metrics such as Mean Squared Error (MSE) or PSNR, which assess absolute errors, SSIM undertakes a comprehensive analysis of two images by evaluating their luminance similarity, contrast similarity, and structural similarity within their local neighborhoods, see Figure 3.3. By analyzing the local variations and spatial dependencies within these neighborhoods, SSIM produces a score that ranges from 0 to 1, with 1 indicating perfect similarity. The calculation involves analyzing various windows within an image, with the final score being the average across all these windows.

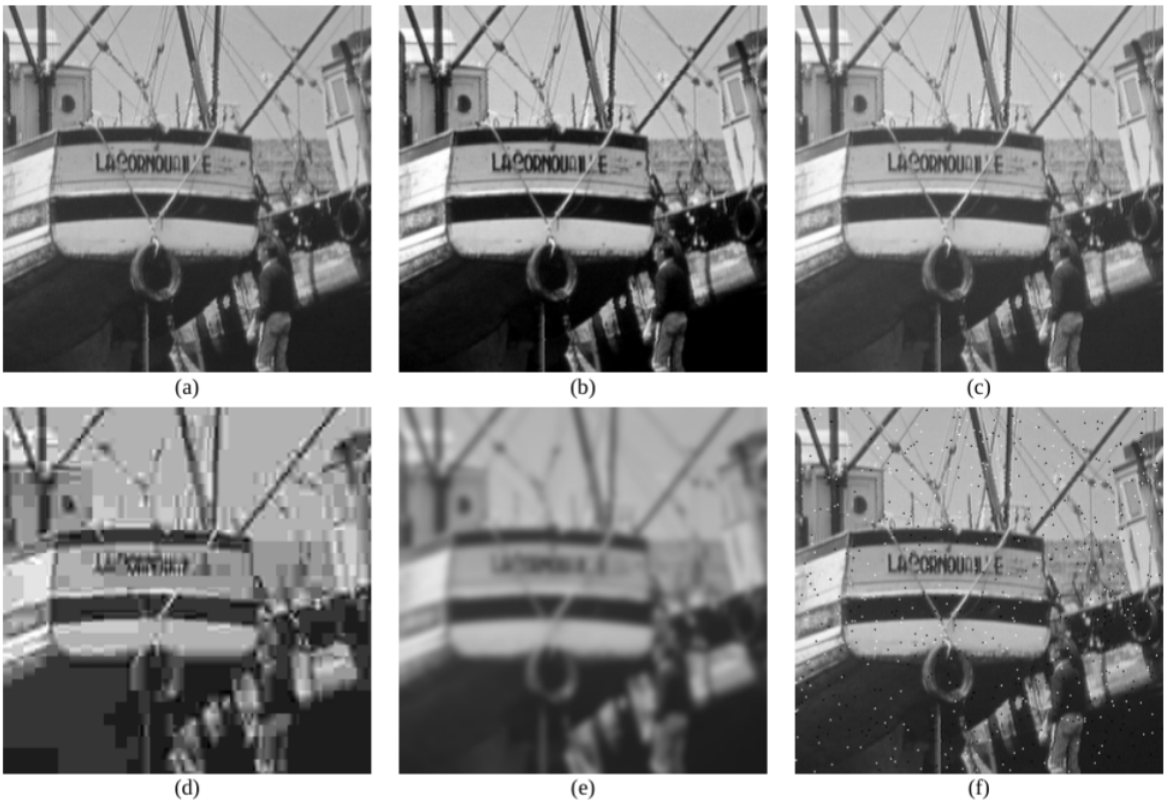


Figure 3.3: SSIM comparison. The method is employed on an image with different types of distortions. (a) Original image; (b) Contrast stretched image, SSIM = 0.9168; (c) Mean-shifted image, SSIM = 0.9900; (d)JPEG compressed image, SSIM = 0.6949; (e) Blurred image, SSIM = 0.7052; (f) Salt-pepper impulsive noise-contaminated image, SSIM = 0.7748. Extracted from (WANG *et al.*, 2004)

3.2.3 Fréchet Inception Distance

The FID is a widely used evaluation metric for assessing the quality and diversity of generated images (HEUSEL *et al.*, 2017). FID is designed to capture the similarity between the distribution of generated images and the distribution of actual images, marking an improvement over the Inception score, see Figure 3.4. While the Inception score uses a classifier network to evaluate the quality and diversity of images by outputting a score, FID calculates the distance between feature embeddings extracted from real and generated images by a pre-trained Inception network. This captures their high-level semantic information. FID compares the multivariate Gaussian distributions of these embeddings by computing the Fréchet distance, offering a quantitative measure of the discrepancy between the distributions. A lower FID score indicates better image quality and diversity, reflecting a closer resemblance between the distributions of generated and authentic images.

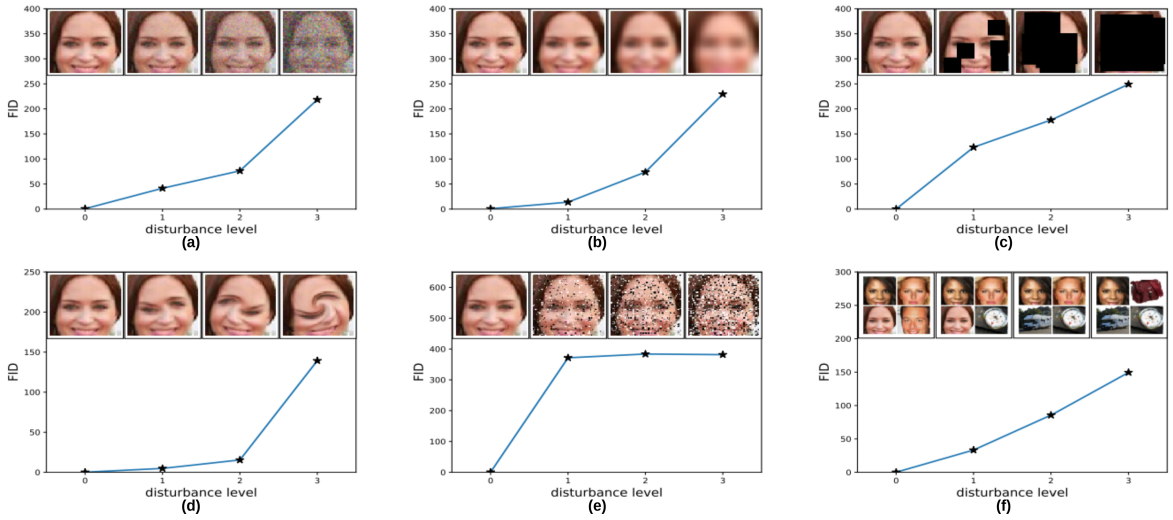


Figure 3.4: FID disturbance comparison. (a) Gaussian noise; (b) Gaussian blur; (c) Implanted black rectangles; (d) Swirled images; (e) salt and pepper noise; (f) CelebA dataset contaminated by ImageNet images. FID scores increase as the level of disturbance grows. Extracted from (HEUSEL *et al.*, 2017)

3.2.4 Learned Perceptual Image Patch Similarity

LPIPS, similar to FID, evaluates the quality of images using feature representations based on deep learning models (ZHANG *et al.*, 2018). Unlike FID, which compares distributions, LPIPS evaluates the perceptual similarity between two images by computing the cosine distance from their respective feature representations (Figure 3.5). This metric focuses on patches of images rather than the entire image, assessing local similarities and integrating them to yield an overall perceptual similarity score. A higher score indicates more significant dissimilarity, while a lower score signifies higher similarity.

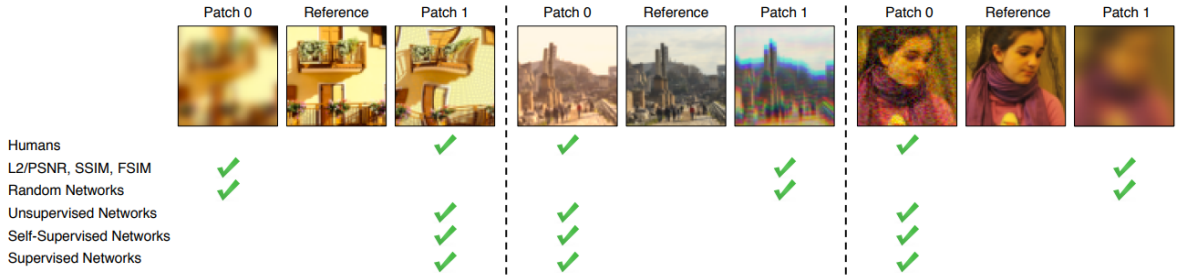


Figure 3.5: Objective metrics comparison on human perceptual. The traditional metrics (L2/PSNR, SSIM, FSIM) do not align with human judgments. However, random deep networks with different supervision types are more effective at capturing human-like perceptual similarities in images. Extracted from (ZHANG *et al.*, 2018)

3.2.5 Landmark Distance

To evaluate the accuracy of generated lip landmarks, Chen *et al.* (2018) computed the average Euclidean distance between each corresponding ground truth and generated landmark point. The LMD is defined as:

$$\text{LMD} = \frac{1}{T} \times \frac{1}{P} \sum_{t=1}^{T} \sum_{p=1}^{P} \|LR_{t,p} - LF_{t,p}\|_2 \,,$$

where T represents the quantity of video frames, and P denotes the total number of landmark points in each frame.

Lower LMD values indicate higher accuracy, reflecting a closer match between generated and actual facial landmarks.

3.3 Computational Resources

Tables 3.2 and 3.1 summarize the most important software and hardware specifications employed in each framework stage. This information is crucial to replicate the experiment. All stages were executed on a Linux server.

Table 3.1: Hardware Specifications used from the Artificial Intelligence Lab (Recod.ai).

Stage	**Processor**	**RAM Memory**	**Graphics card**
First	Intel i7-5820K, 3.30GHz	32GB	NVIDIA RTX A6000, 48GB
Second	Intel Xeon 5218R, 2.10GHz	32GB	NVIDIA TITAN X, 12GB

3.4 Concluding Remarks

This chapter outlines the resources and tools applied for training and validating our framework. It includes the video dataset employed for model training, the CH-

Table 3.2: Software Specifications.

Resource	**First Stage**	**Second Stage**
Python (<https://www.python.org/>)	3.7	3.5
Pytorch (<https://pytorch.org/>)	1.9	0.4.0
Torchvision (<https://pytorch.org/vision/stable/index.html>)	0.10	0.2.1
CUDA (<https://developer.nvidia.com/cuda-toolkit>)	11.1	9.0
cuDNN (<https://developer.nvidia.com/cudnn>)	-	7.1.2

Unicamp, which contains 15 minutes of daily Portuguese speech covering most of the main phonemes/visemes; the landmark extraction method, an enhanced state-of-the-art approach; computational resources, including the mainly computational software libraries and hardware specifications; and the most commonly used objective evaluation methods in the literature, employed to derive evaluation metrics. The next chapter presents the whole pipeline for synthesizing talking head animations from speech and details the framework's architectural model.

Chapter 4

Two-Stage Talking Head Generator Framework

This chapter presents the whole methodology of training a speech-driven, two-stage, landmark-based model capable of synthesizing a realistic talking head. It integrates different technologies (Transformer and GAN), leveraging the strengths of both to address the challenges in talking head generation. Additionally, the Transformer model *FaceFormer*, employed in this work, was originally built to generate 3D meshes. We adapted the first layer dimension to synthesize 2D landmarks. Figure 4.1 overviews the training pipeline.

The model employs raw audio and video frames sourced from the CH-Unicamp database. The processing of audio and images is detailed in Section 4.1. This framework is structured in a two-stage style. Section 4.3 describes the first stage of the framework, where the objective is to map raw audio to facial landmarks. Section 4.4 then delineates the second stage, mapping these landmarks to realistic video frames. Sections 4.5 and 4.6 cover experiment details, including training and inference information. Lastly, Section 4.7 presents the concluding remarks of the chapter.

4.1 Data Preprocessing

This work uses raw audio as input and a sequence of facial 2D landmarks as an intermediary representation. This sequence is the output of the first stage framework, which is a Transformer-based model and serves as the input for the second stage, a GAN-based model.

The audios are extracted from the dataset videos and reduced to 16kHz due to a pre-trained model used in this work, the *wave2vec 2.0* (GROSMAN, 2021).

Accurate facial keypoint detection is crucial for obtaining favorable outcomes in this work, as these keypoints are utilized to create the target output for the first-stage model and the input for the second-stage model. The task of recognizing facial keypoints

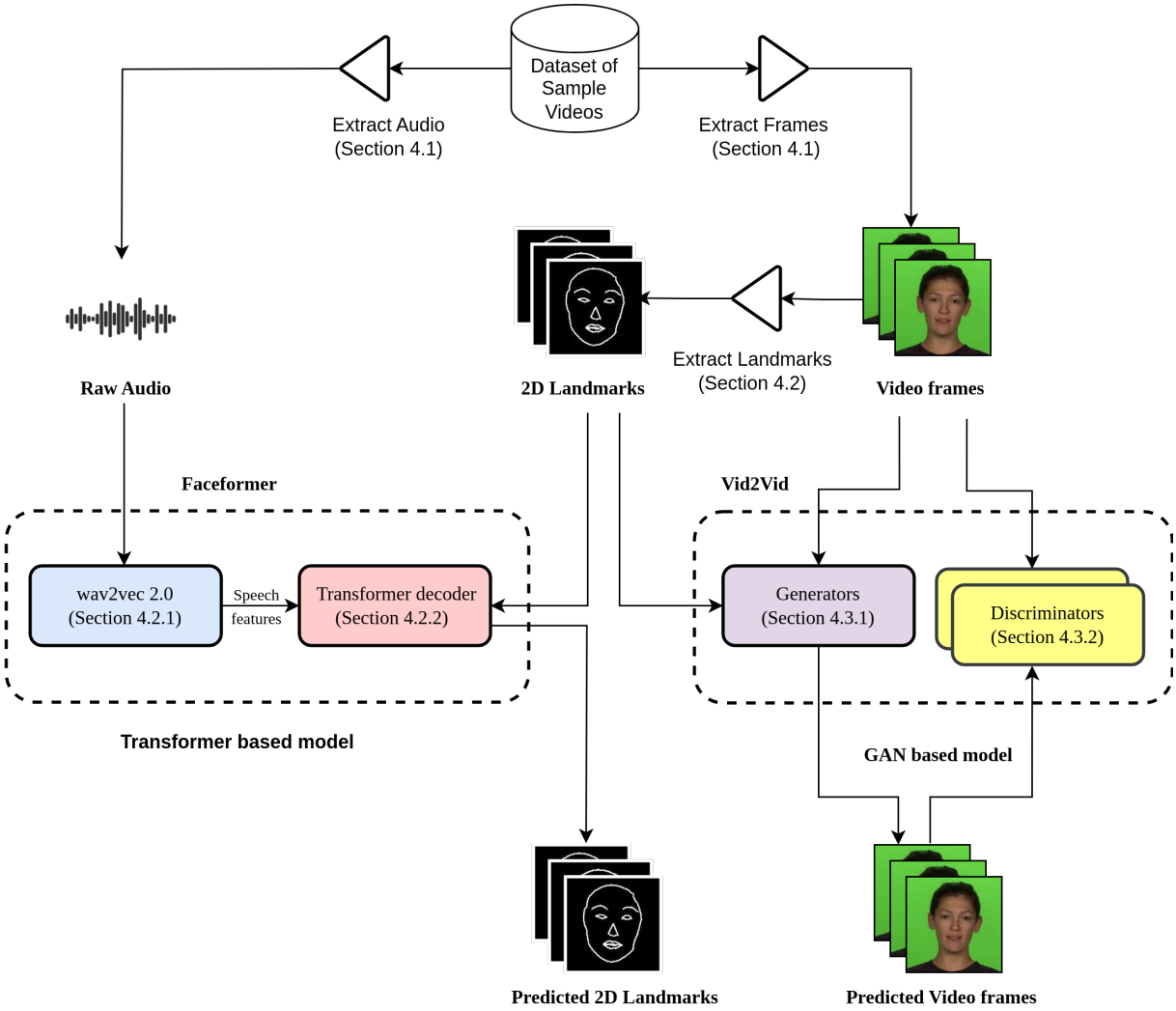


Figure 4.1: Training overview process with respective text sections.

is well acknowledged, and conventional software packages produce remarkable outcomes. Reis (2020) tested different approaches to identifying facial keypoints in the same dataset of this work, DLIB (KING, 2009) and FAN (BULAT; TZIMIROPOULOS, 2017b), more details in Section 4.2. The DLIB approach did not identify keypoints when the actress performed small, natural facial rotations while speaking. In contrast, the FAN approach achieves better results even in minor facial rotations. Therefore, this work uses the FAN approach to obtain the facial keypoints. Figure 4.2 exhibits the keypoints extracted.

To build the 2D landmark facial representations, first, frames were extracted from all videos at 30 FPS, followed by a center crop and downsampling process, resulting in a resolution of 256x256 pixels. The images were reduced due to the computational resources required to train the second stage model, the *vid2vid* model. Finally, the FAN is employed to extract the facial keypoints. A 2D facial landmark vector with dimensions (68×2) is obtained for each realistic video frame.

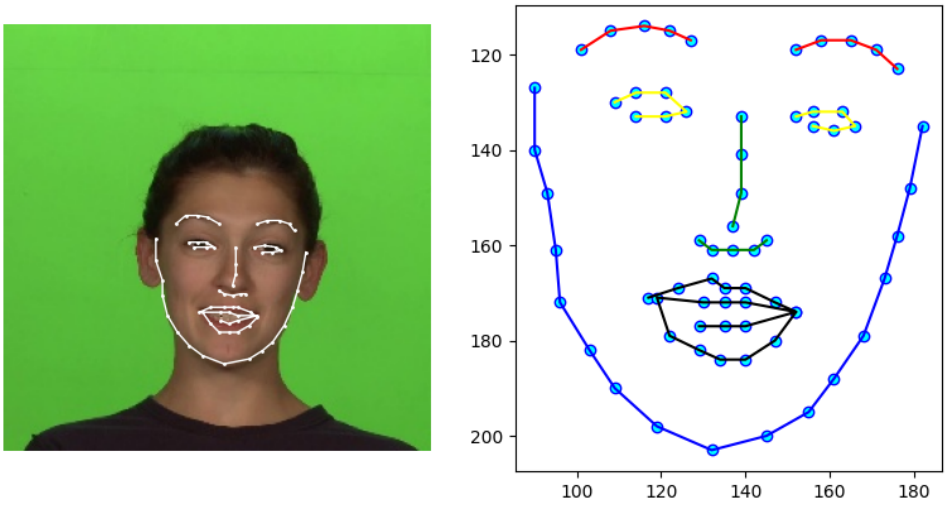


Figure 4.2: Example of keypoint extraction. On the left side, a video frame with overwritten keypoints. On the right side, the key points are grouped by colored lines indicating each facial region. Red eyebrows, yellow eyes, green nose, black lips, and blue chin.

4.2 Facial Landmark Extractor Method

We use the FAN to extract facial landmarks (BULAT; TZIMIROPOULOS, 2017b). Reis (2020) tested different approaches to identifying facial keypoints and chose FAN because it achieves better results even in minor facial rotations. FAN is based on a stack of four Hourglass (HG) networks, a state-of-the-art architecture for human pose estimation (NEWELL *et al.*, 2016).

Unlike HG, which uses ResNet as a bottleneck block, FAN employs a new advanced residual block (HE *et al.*, 2016). Bulat e Tzimiropoulos (2017a) proposed a novel hierarchical, parallel, and multi-scale residual architecture that significantly outperforms the standard residual block from ResNet. This approach increases the receptive field size, improves gradient flow, and is specifically designed to have a similar number of parameters as the original bottleneck. FAN was trained on 300W-LP, a very large facial landmark dataset containing 2D and 3D landmarks, and evaluated on other 2D and 3D datasets (approximately 230,000 images) (ZHU *et al.*, 2015).

4.3 First Stage: Audio to Landmarks

To map the landmarks from audio, we use the *FaceFormer* model implementation (FAN *et al.*, 2022). *FaceFormer* is based on the full encoder-decoder Transformer architecture.

FaceFormer delves deeper into the temporal dynamics of speech. Through a series

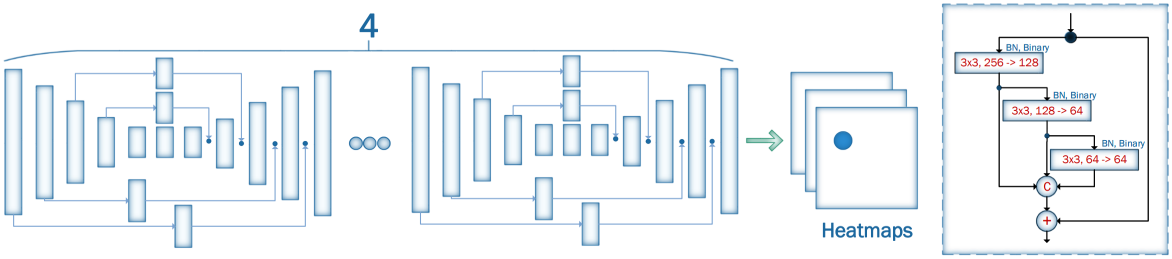


Figure 4.3: The Face Alignment Network (FAN) is designed as a stack of four HourGlass (HG) networks. Each rectangle is replaced by the new advanced residual block (shown on the right in the image). Extracted from (BULAT; TZIMIROPOULOS, 2017b)

of self-attention mechanisms and encoder-decoder structures, it outputs a detailed face representation in the form of facial landmarks. These landmarks encapsulate the essential movements and expressions that are synchronous with the spoken content, forming a robust foundation for the subsequent rendering stage.

The effectiveness of the *FaceFormer* is rooted in its Transformer-based design, which allows for processing long-range dependencies in speech patterns, a critical factor in maintaining the natural flow of expressions in generated talking heads. Furthermore, the model's ability to learn from vast amounts of data ensures that the synthesized facial representations are accurate and diverse, accommodating a wide range of speech articulations.

Therefore, following the encoder-decoder Transformer architecture, *FaceFormer* is presented in two steps: the encoder, which extracts features from the audio, and the decoder, which maps the speech features to a sequence of facial landmarks.

4.3.1 Extracting Features from Speech Audio

Instead of the common adoption of MFCC features, *FaceFormer* adapts as an audio feature extractor the self-supervised pre-trained speech model, *wav2vec 2.0*, see Figure 4.4.

wav2vec 2.0 can be described by three components:

- **Temporal Convolutional Layers**. The encoder's architecture comprises multiple blocks of TCNs, succeeded by layer normalization and the Gaussian Error Linear Unit (GELU) activation function. Before inputting into the encoder, the raw waveform is standardized to have a zero mean and unit variance.

- **Context Network**. A Transformer encoder with a series of multi-head self-attention and feed-forward layers that transform the audio feature vectors from TCN layers into enriched speech representations. Rather than utilizing fixed positional embeddings that provide absolute positional information, the model incorporates a

one-dimensional convolutional layer that functions as a relative positional embedding, capturing the position of each element in relation to others within the sequence.

- **Quantization Module**. The features from the TCN layer are discretized, in parallel to the Context Network process, to a finite set of speech unit representations through product quantization and the Gumbel softmax function. Product quantization involves the selection of quantized representations from an array of codebooks, followed by their concatenation. The Gumbel softmax function facilitates the selection of discrete entries from these codebooks, allowing full differentiability and, consequently, backpropagation.

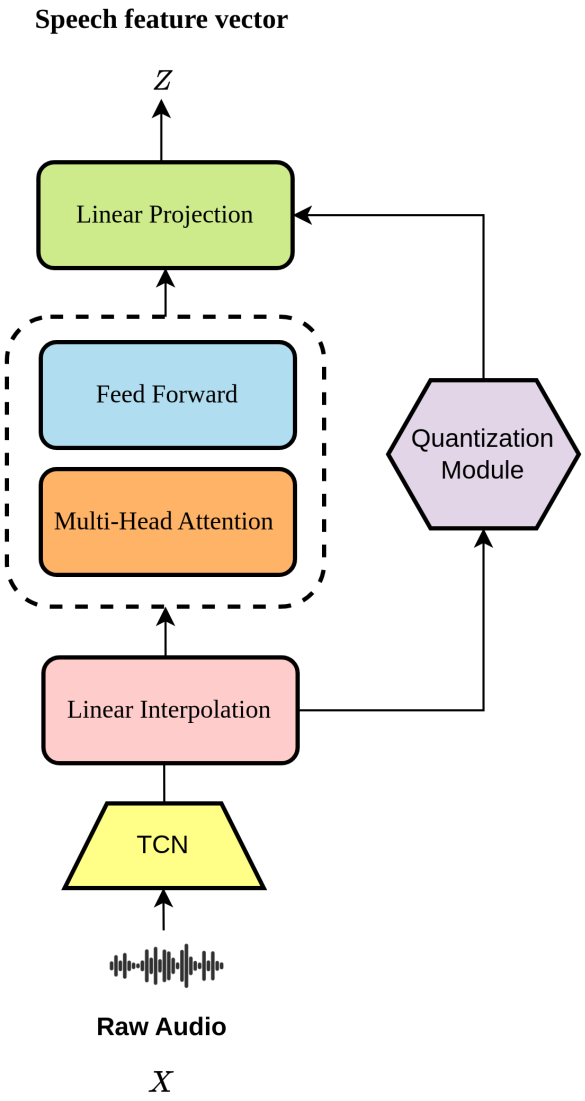


Figure 4.4: *wav2vec 2.0* adapted to crossmodal compatibility.

The model is pre-trained in a self-supervised manner on unlabeled speech data to learn representations of discrete speech units. The pre-training strategy is similar to the

masked language modeling used in Bidirectional Encoder Representations for Transformers (BERT) (DEVLIN *et al.*, 2018). It involves masking a portion of the feature encoder outputs, which are then passed through the context network and replaced with a pre-trained feature vector. The loss function is composed of a contrastive loss which aims to identify the true quantized latent speech representation for a masked time step within a set of distractors. A diversity loss is also introduced to promote equitable utilization of all values from the codebooks. After pre-training, the model is fine-tuned with labeled data for the desired task by adding a randomly initialized linear projection on top of the context network.

Linear Interpolation for Crossmodal Compability

Audio and video components typically operate at different frequencies. Audio data often has a higher sampling rate than video data's frame rate. This discrepancy presents a challenge when trying to synchronize the two modalities. The Multi-Head Attention mechanism in the Transformer's decoder is designed to draw information from the encoder's output and align it with the input sequence to the decoder for each timestep. This means aligning speech features with corresponding video frames representing the respective facial movements. If the dimensions of the speech features do not align with the motion frequency of the video frames, the model may struggle to synchronize the audio with the appropriate video frames accurately.

Fan *et al.* (2022) adapted *wav2vec 2.0* by incorporating a linear interpolation on the TCN output layer to align the dimensionality of the speech features with the number of video frames. The output from *wav2vec 2.0* is then utilized with the target facial landmark frames in the decoder's Multi-Head Attention through matrix multiplication. For this multiplication to be valid, the dimensions must be compatible. Section 4.3.2 will provide a detailed discussion on this topic. Given that the target video is captured at a frequency of f_v (e.g., $f_v = 30$ FPS) and the audio at a frequency of f_a (e.g., $f_a = 44$Hz), the length of features output from the linear interpolation is kT, where T represents the video frame size, and k is determined as $\lceil f_v/f_a \rceil$, a alignment constant. Consequently, the decoder Cross-Modal Multi-Head Attention can align the speech and motion modalities (Section 4.3.2). In this work $k = \lceil \frac{30}{16} \rceil = 1$

To adapt to our context, a fine-tuned *wav2vec 2.0* based on the Brazilian Portuguese language is used (GROSMAN, 2021).

4.3.2 From Audio Speech Features to Facial Landmarks

The *FaceFormer* Transformer-based decoder learns to transform the speech features Z in a predicted facial landmarks $\hat{Y}$ sequence. In an autoregressive manner, the decoder predicts the next facial landmark frame based on all previously predicted frames, and the encoder's output speech features Z. Fan *et al.* (2022) enhanced the classic

Transformer decoder to generalize for longer input sequences and handle different data modalities.

To improve the model's generalization ability for longer audio sequences, Fan *et al.* (2022), inspired by Attention with Linear Biases (ALiBi), enhanced the classic Transformer decoder by adding a temporal bias to the query-key attention score and designed a periodic positional encoding strategy (PRESS *et al.*, 2021). To align the different data modalities, they design the biased cross-modal Multi-Head Attention, see Figure 4.5. Additionally, originally *FaceFormer* was built to generate 3D meshes, then to synthesize 2D landmarks, we change the first input layer dimension, the Motion Encoder, to (68×2).

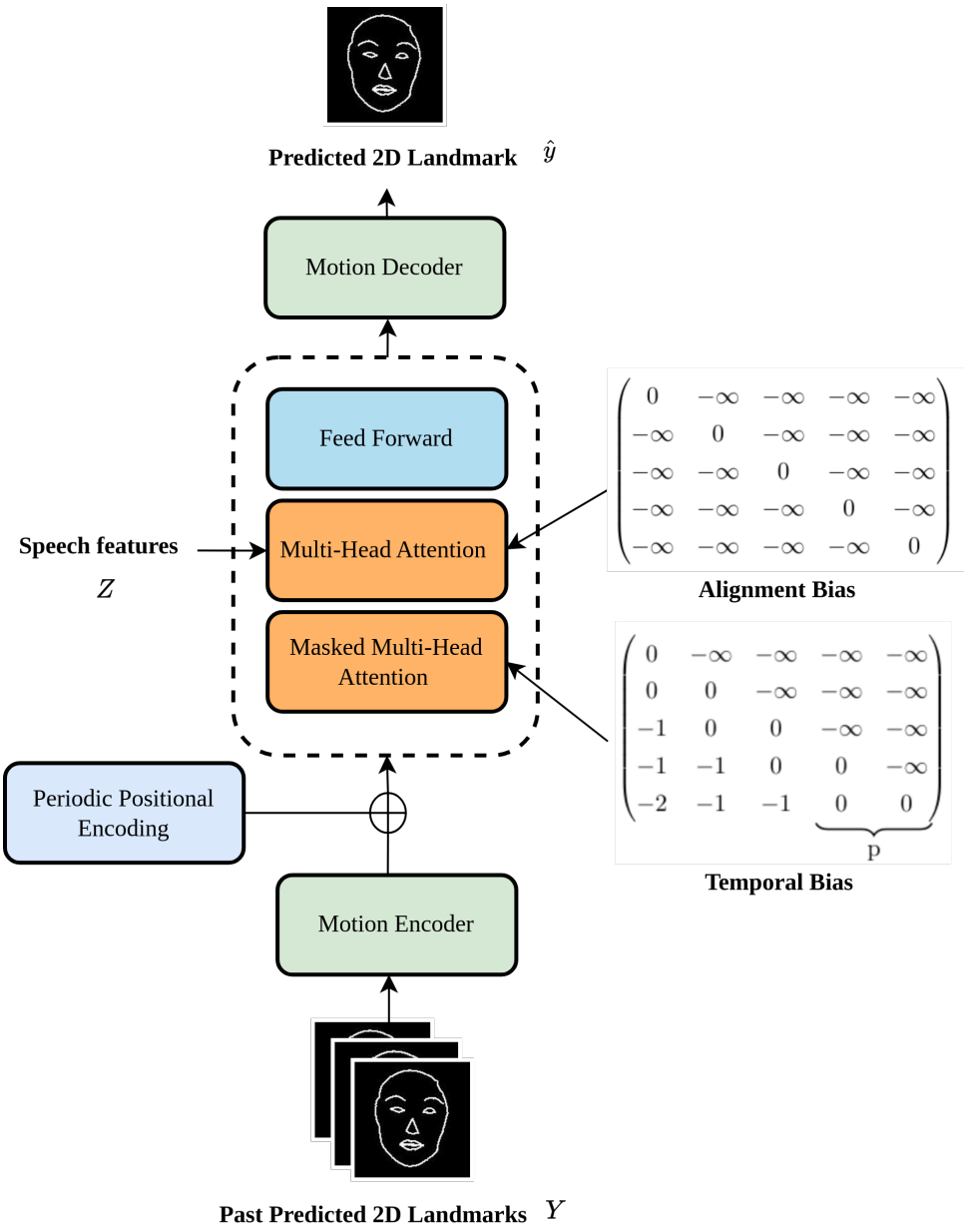


Figure 4.5: Transformer decoder adapted to long sequences.

Periodic Positional Encoding

In the context of natural language processing, ALiBi analyzed the Transformer's ability to extrapolate, that is, generate sequences in inference time longer than those used during training. The findings indicate that adding a bias to the attention scores in the query-key mechanism can be more efficient than employing conventional positional embeddings. This approach is advantageous regarding reduced memory usage, increased speed, and improved extrapolation ability. Inspired by this, Fan *et al.* (2022) attempted to apply the same mechanism to synthesizing 3D facial meshes. However, this was observed to lead to static facial expressions during inference. The ALiBi's mechanism fails to incorporate position information into the input representation, which is crucial to capturing the subtle variations in motion across sequential facial frames. To address this issue, Fan *et al.* (2022) reintroduced the sinusoidal positional embeddings (see Section 2.1.4) but adapted them to be periodic by adding a modulo operation to the positional object t:

$$\mathrm{PPE}_{t,2i} = \sin\left(\frac{(t \mod p)}{10000^{2i/d}}\right) PPE_{t,2i+1} = \cos\left(\frac{(t \mod p)}{10000^{2i/d}}\right)$$

The proposed Periodic Positional Encoding (PPE) strategy, which recurrently injects position information within each period p, demonstrated greater efficiency than assigning a unique position identifier for each token in the sequence.

Biased Causal Multi-Head Self-Attention

The ALiBi strategy, which substitutes the positional embedding with the bias in the attention scores, was inefficient for synthesizing facial mesh. However, Fan *et al.* (2022) adapted it to their work by making it periodic, similar to the PPE. To learn the dependencies between each frame in the context of the past facial frames sequence, a weighted contextual representation is calculated by adding a temporal bias to the scaled dot-product attention:

$$\mathrm{Att}\left(Q^F, K^F, V^F, B^F\right) = \mathrm{softmax}\left(\frac{Q^F (K^F)^T}{\sqrt{d_k}} + B^F\right) V^F,$$

where Q^F, K^F, V^F are the respective input linear projections of Query/Key/Value from the Masked Multi-Head Attention, and B^F is the temporal bias.

The main idea is to introduce a bias proportional to the distance of the target prediction. Unlike the original approach, this bias decreases with each p frame. In practice, B^F is a matrix where the upper triangle contains negative infinity to prevent future frames from influencing the prediction, and the lower triangle represents the temporal bias. This ensures that the closest facial frames have a greater impact on the target prediction:

$$B^F_{(i,j)} = \begin{cases} \frac{(i-j)}{p}, & \text{if } j \leq i, \\ -\infty, & \text{otherwise.} \end{cases}$$

where i and j are the indices of B^F.

Biased Cross-Modal Multi-Head Attention

As discussed in Section 4.3.1, the Transformer decoder must process speech features and the previously predicted facial landmarks. The Biased Cross-Modal Multi-Head Attention aims to combine the different modalities by adding an alignment bias B^A to the query-key attention score. The alignment bias B^A is defined as:

$$B^A_{(i,j)} = \begin{cases} 0, & \text{if } ki \leq j < k(i+1) \\ -\infty, & \text{otherwise} \end{cases}$$

where i and j are the indices of B^A and k the alignment constant.

Given the speech feature Z_{kT} output from the *wave2vec 2.0* and F_{kT}, the output of the Biased Causal Multi-Head Attention, which encoded the predicted face motions. Both Z_{kT} and F_{kT} serve as input to the Biased Cross-Modal Multi-Head Attention, where the speech feature Z_{kT} passes through Key and Value W matrices to be transformed into the Key K^A and Value V^A vectors, while the F_kT is converted into the Query $Q^{\hat{F}}$ vector, see Section 2.1.2. Consequently, the attention layer computation considers only the features from Z_{kT} and F_kT that are related to the target predicted token, aligning the speech features and the face motions. The Biased Cross-Modal Multi-Head Attention is defined as:

$$\text{Att}(Q^A, K^A, V^A, B^A) = \text{softmax}\left(\frac{Q^{\hat{F}}(K^A)^T}{\sqrt{d_k}} + B^A\right) V^A$$

Ultimately, the predicted target face landmark $\hat{y}$ is obtained through a linear transformation that projects the Biased cross-modal output into the landmark dimension (68×2).

Learning Objective

The decoder repeats the process to predict the next facial landmark $\hat{y}$ in an autoregressive manner until the end of the video sequence. In this moment, the model is trained by minimizing the MSE between the predicted outputs $\hat{Y} = (\hat{y}_1, ..., \hat{y}_{kT})$ and the ground truth landmarks $Y = (y_1, ..., y_{kT})$:

$$\mathcal{L}_{\text{MSE}} = \sum_{t=1}^{T} \sum_{v=1}^{V} \|\hat{y}_{t,v} - y_{t,v}\|^2$$

4.4 Second Stage: Facial Landmarks To Facial Image

Following the successful transformation of audio input into a detailed facial representation via the *FaceFormer* model, the next step in synthesizing a talking head is to render these facial landmarks into realistic video frames. To map the landmarks to realistic images, we use the *vid2vid* model implementation (WANG *et al.*, 2018). *vid2vid* is a general video-to-video synthesis framework based on CGAN and introduces a new spatio-temporal learning objective. The primary objective of *vid2vid* is to achieve high realism in the synthesized video frames. This entails capturing the fine details and nuances of facial expressions and ensuring that these expressions are in perfect sync with the audio input. The realism extends to the seamless transition of facial movements across frames, maintaining consistency and fluidity that mirrors natural human expressions.

One of the key challenges in video synthesis is maintaining temporal coherence across frames. *vid2vid* addresses this through sophisticated temporal modeling, ensuring that each generated frame is consistent with its predecessors, thereby avoiding jitter and unnatural movements. Wang *et al.* (2018) incorporated optical flow, specifically using FlowNet2, to achieve temporal consistency in video synthesis. Optical flow refers to the pattern of apparent motion of objects in a visual scene, as observed from a viewpoint. In the context of video, it represents the motion between two consecutive frames. For realistic video synthesis, it's crucial to maintain continuity and smoothness between frames. Optical flow provides the motion vectors that describe how each pixel in one frame moves to the next. This information is vital for ensuring that subsequent frames in a generated video are coherent and temporally consistent.

Therefore, the GAN-based framework utilizes different generators and discriminators, supported by FlowNet2, in conjunction with a spatio-temporal adversarial objective, to achieve highly realistic and temporally coherent frames.

4.4.1 Generator

The *vid2vid* network is a CGAN framework for video-to-video synthesis, where a source sequence of frames S is mapped to a corresponding sequence of real video frames X. The objective is to learn a mapping function G that can convert S to $\hat{X}$ such that the conditional distribution of $\hat{X}$ given S matches the real conditional distribution of frames:

$$p(\hat{x}_1^T | s_1^T) \approx p(x_1^T | s_1^T)$$

This is formulated as a minimax optimization problem with the generator G mapping the input sequence to the output frame sequence, trained according to the adversarial loss Equation 4.1.

The conditional distribution $p(\hat{x}|s)$ is simplified through a Markov assumption

to a product factorized:

$$p(\hat{x}_1^T|s_1^T) = \prod_{t=1}^{T} p(\hat{x}_t|\hat{x}_{t-L}^{t-1}, s_{t-L}^{t-1})$$

Wang *et al.* (2018) assumes that the video frames can be generated sequentially, in which the generation of the current frame only depends on the current source frame s_t, the past L source frames $(s_{t-1}, ..., s_{tL})$ and the past L generated frames $(\hat{x}_{t-1}, ..., \hat{x}_{tL})$. A feed-forward network G is trained to capture the conditional distribution $p(\hat{x}|s)$ and outputs the next frame, which is obtained recursively. Wang *et al.* (2018) experiments suggest $L = 2$ as the best trade-off between quality and computational resources.

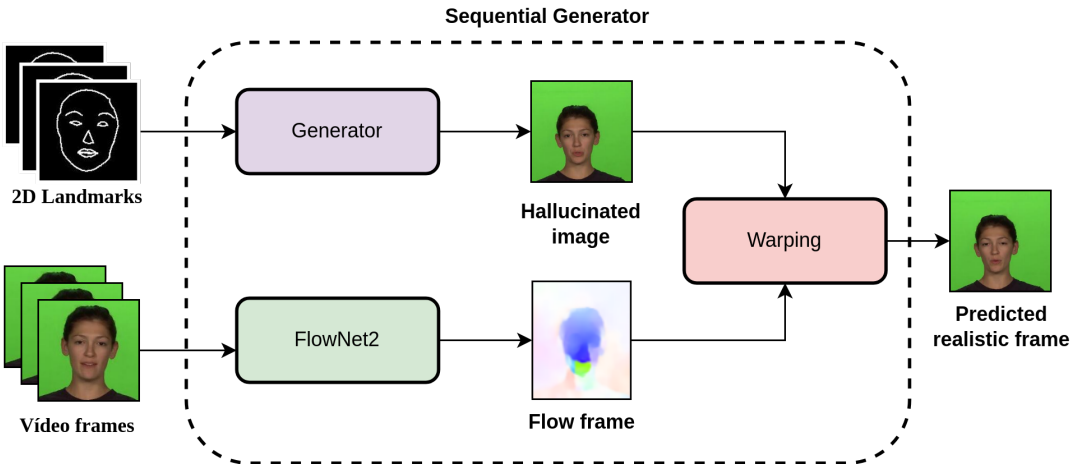


Figure 4.6: Sequential Generator Overview.

The Generator G is structured in a coarse-to-fine style by breaking down the generation process into hierarchical stages of refinements (see Figure 4.6). Video signals have a significant amount of repetitive information in successive frames. By understanding the optical flow, which is the pattern of apparent motion between frames, it is possible to predict the subsequent frame by adjusting the current one. This prediction method is generally accurate, except in obscured or not visible areas. Accordingly, this idea, the Generator G, is designed in three steps: firstly, it synthesizes the next frame in a hallucinatory manner; secondly, it estimates the optical flow; and finally, it combines the outcomes of the previous steps to synthesize the final predicted frame. Both the first and second steps consider the background information through a mask. This process is outlined in the subsequent equation:

$$G(\hat{x}_{t-L}^{t-1}, s_{t-L}^{t}) = (1 - \hat{m}_t) \odot \hat{w}_{t-1}(\hat{x}_{t-1}) + \hat{m}_t \odot \hat{h}_t, \tag{4.1}$$

where h_t is the hallucinated image; w_{t-1} is the estimated optical flow; m_t represents an occlusion mask employed to manage background information. Both h, w, and m are computed by residual networks. All network outputs are based on the input source

images s and previously synthesized images $\hat{x}$.

4.4.2 Discriminators

Two types of discriminators are designed: Conditional Image Discriminator (D_I) and Conditional Video Discriminator (D_V) (see Figure 4.7). These discriminators are based on PatchGAN architecture. Unlike a traditional discriminator that classifies an entire image as real or fake, PatchGAN focuses on smaller patches of an image. Each patch is independently assessed, and the discriminator's outputs are averaged to decide the image's authenticity. PatchGAN effectively captures and critiques the finer textures and details of images by concentrating on small areas. This local perspective allows it to enforce high-frequency correctness, ensuring the generated images are texturally realistic. PatchGAN discriminators generally have fewer parameters than those assessing the entire image. This efficiency makes them faster and less resource-intensive, facilitating their use in applications where detail and speed are crucial.

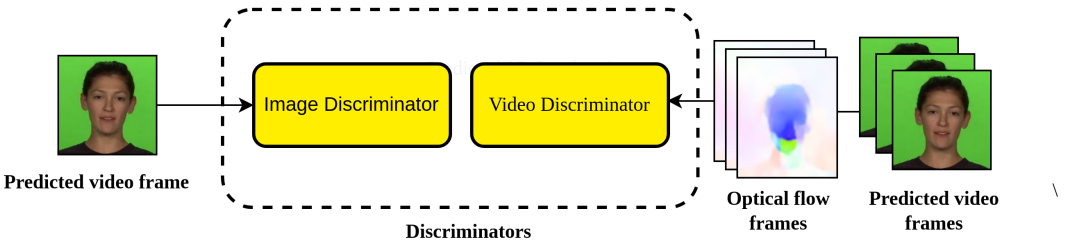


Figure 4.7: Image and Video Discriminators.

D_I aims to ensure that each generated output frame resembles the expected real frame. The discriminator output should be true for a very similar output or false for a fake output. Wang *et al.* (2018) enhances the PatchGAN architecture to multiple discriminators of different scales, this means that the image is processed by discriminators that look at patches of different sizes (ISOLA *et al.*, 2017). By using multiple scales, the GAN can more accurately assess the realism of images at various levels of detail, from coarse structures to fine textures. The generator receives richer feedback on different aspects of the image quality, which can guide it in producing more convincing outputs across resolutions.

D_V is designed to ensure that consecutive generated output frames follow the temporal dynamics considering the optical flow. This approach enables the discriminator to examine the temporal dynamics of the video and efficiently penalize unnatural or sudden changes within the sequence of frames. This discriminator also implements multi-scale as D_I to different image patch sizes. Other multi-scale techniques are employed but with a focus on the temporal aspect. At the most detailed level, the discriminator analyzes a sequence of K directly following frames from the original series. Moving to a broader scale, the discriminators skip $K - 1$ intermediate frames, still considering a total of K

frames in this new sequence. The authors find that this helps to ensure both short-term and long-term consistency.

4.4.3 Learning Objective

To train the Generator G, the following learning objective function is minimized:

$$\min_G \left(\max_{D_I} \mathcal{L}_I(G, D_I) + \max_{D_V} \mathcal{L}_V(G, D_V) \right) + \lambda_W \mathcal{L}_W(G),$$

where $\mathcal{L}_I$ is the adversarial loss related to the conditional image discriminator D_I, $\mathcal{L}_V$ is the adversarial loss on K frames related to the conditional video discriminator D_V, and $\mathcal{L}_W$ is the flow estimation loss.

The $\mathcal{L}_I$ loss is given by:

$$\mathcal{L}_I = E_{(x_1^T, s_1^T)}[\log D_I(x_i, s_i)] + E_{(x_1^T, s_1^T)}[\log(1 - D_I(\hat{x}_i, s_i))].$$

The $\mathcal{L}_V$ loss is computed recursively for all K frames:

$$\mathcal{L}_V = E_{(x_1^T, s_1^T)}[\log D_I(x_i, s_i)] + E_{(x_1^T, s_1^T)}[\log(1 - D_I(\hat{x}_i, s_i))].$$

The $\mathcal{L}_W$ loss is the sum of the differences between the ground truth and predicted optical flow and the ground truth and predicted final frame. The loss is given by:

$$\mathcal{L}_W = \frac{1}{T-1} \sum_{t=1}^{T-1} \left(\|\widetilde{w}_t - w_t\|_1 + \|\widetilde{w}_t(x_t) - x_{t+1}\|_1 \right).$$

4.5 Training

The models were separately trained with the Adam optimizer using a fixed learning rate of 10^{-4} to FaceFormer and 2.10^{-4} to *vid2vid*. Both models with batch size set to 1. FaceFormer model were trained for 2560 epochs with the facial landmarks extracted by FAN. The encoder parameters were initialized and fixed with the pre-trained *wav2vec 2.0* weights (GROSMAN, 2021). *vid2vid* was trained for 120 epochs with both realistic and 2D-facial landmarks video frames.

The FaceFormer network consumed around 13 GB of video memory in the training process. The training stage took approximately one week until epoch 2560 in this configuration. The *vid2vid* network consumed around 11 GB of video memory and required approximately two weeks to complete the training process until epoch 120.

4.6 Inference

Similar to the training process, inference is performed in two steps. Figure 4.8 shows the framework during the inference mode. In the initial step, the FaceFormer encoder takes the raw audio X as input and produces the speech features Z. Subsequently, the FaceFormer decoder autoregressively generates the 2D-landmark sequence. In the second step, for each 2D landmark, the *vid2vid* model generates a photorealistic frame. The *FaceFormer* model requires approximately 3.5 ms to synthesize each facial landmark, and a 9-second video takes approximately 1.8 s to process. Similarly, the *vid2vid* model takes about 4 ms to synthesize each realistic frame, with a 9-second video requiring around 2 s to complete.

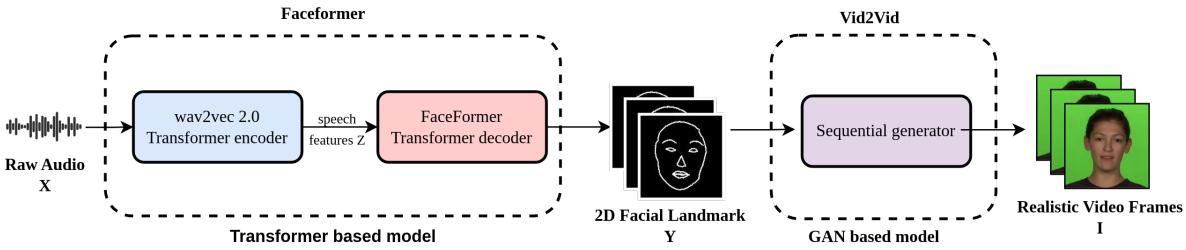


Figure 4.8: Framework overview during the inference stage.

The chosen synthesis approach initially involves selecting the optimal epoch of the *vid2vid* model while using ground truth 2D landmarks as input. Subsequently, we initiate the process of synthesizing 2D landmarks with *FaceFormer* model and assess its performance in conjunction with *vid2vid*.

4.7 Concluding Remarks

The current chapter describes the development of our talking head generation framework based on a two-stage landmark design. This approach is elaborated from dataset preprocessing through to the training and synthesis processes. The framework is capable of synthesizing talking head videos from raw audio.

The audio and video frames were preprocessed to fit our context. We downsampled the audio frequency for input into a pre-trained *wave2vec 2.0* model, which extracts features from the audio. The video frames' resolution was reduced to decrease computational resource consumption, as our rendering network, *vid2vid*, requires significant resources to synthesize realistic video frames.

Our framework applies the models proposed by Fan *et al.* (2022) and Wang *et al.* (2018) and proposes a new methodology to synthesize videorealistic image-based talking heads from speech. We adapted *FaceFormer* to synthesize *2D* landmarks from Portuguese audio by using a *wave2vec* model fine-tuned on the Portuguese language and adjusted

the dimensionality of the input and output decoder layers to generate 2D landmarks of dimensions (68×2). The *vid2vid* network, leveraging optical flow and PatchGAN for enhanced quality and temporal coherence, is used to render the 2D landmarks into realistic video frames.

The next chapter will present our results. Initially, we apply objective metrics to evaluate image quality and natural spontaneous motion. Subsequently, an ablation study is conducted to understand some framework components better. Finally, we undertake an exploratory test to synthesize talking heads with audio from outside our dataset. This chapter helps us to answer our research questions proposed in Section **??**.

Chapter 5

Results

As our framework employs a two-stage synthesis design, the outputs of both the first and second stages are evaluated using objective metrics. Following the characteristics described in Section 2.4.5 by Chen *et al.* (2020), we assess our results' image quality, audio synchronization, and natural spontaneous motion. Given that the framework is trained to synthesize a specific individual, we do not evaluate identity preservation. Section 5.1 assesses spontaneous motion by examining blinking eyes in the 2D facial landmarks using the EAR. Section 5.2 evaluates the quality of the synthesized realistic image frames by employing metrics such as SSIM, FID, and LPIPS. An ablation study is performed in Section 5.3, analyzing the impacts of modifications proposed by FAN *et al.* to the transformer decoder on landmark synthesis. We utilize the LMD metric to assess audio synchronization on the landmarks, focusing on lip landmarks. Finally, Section 5.4 presents a generalization test of our framework by synthesizing talking heads with random speech audios outside of our dataset. This test evaluates the framework's potential for generalization across different speakers. Examples of animations synthesized using our method can be seen at <ai-unicamp.github.io/2StageTalkingHead>.

The experiment adopted a k-fold cross-validation approach, 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 results presented in this section showcase the aggregate outputs from the k-fold cross-validation iterations, specifically capturing the mean (μ) and standard deviation (σ) of the objective evaluation metrics across different epochs. Each epoch's mean score is computed from all 13 test samples within a single iteration. Subsequently, these scores' means and standard deviations are calculated across all iterations for each epoch.

5.1 Facial Landmarks

We assess spontaneous natural motion on facial landmarks by evaluating the blinking eyes with Eye Aspect Ratio (EAR) (see Figure 5.1).

In the eye blink detection method evaluation, we selected a random k-fold test dataset containing 13 samples. These samples were manually annotated of eye blink occurrence to serve as a benchmark for assessing the detection method performance. Testing the detection method on this dataset yielded an accuracy of 67%. While indicative of the method's potential, this also underscores the challenges inherent in eye blink detection. We believe that the image's low resolution could negatively impact the landmark extractor method precision, where nuances in eye landmarks become critical for accurate blink detection.

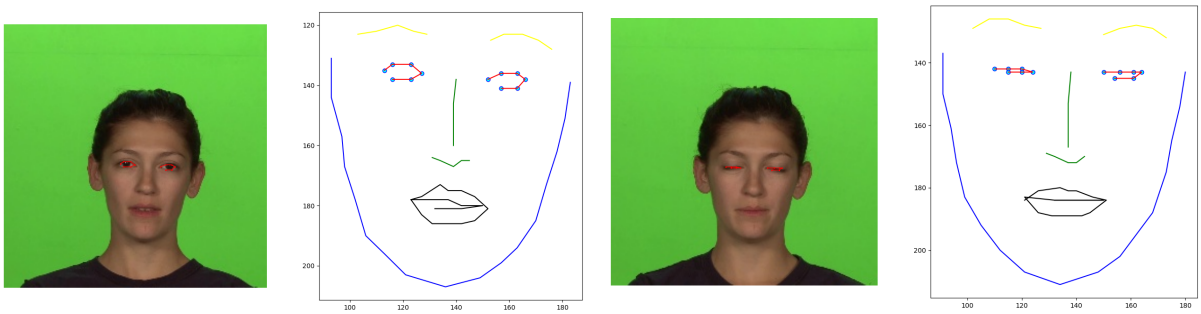


Figure 5.1: Video frame examples and their respective landmarks. The image on the left is an example of an open eye, and the image on the right is an example of a closed eye.

Table 5.1 presents the average number of eye blinks per video for the method applied to both ground truth and synthesized landmarks. The results confirm preliminary observations that *FaceFormer* struggles to capture temporal eye accurately blinks representation. This insight is crucial, highlighting a specific area where the method's performance could be enhanced.

Table 5.1: Blink detection method score. It was computed using both ground truth and synthesized 2D landmarks. The method is employed to count the eye blinks.

	Synthesized	**Ground Truth**
Average Eye Blink per video	0	4

5.2 Realistic Images

To conduct the experiments on the second-stage output, the realistic frame images, we fixed the model checkpoints of the second-stage (*vid2vid*), and we varied its inputs (landmarks) to assess 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. Figure 5.2 displays examples of ground truth and synthesized frames.

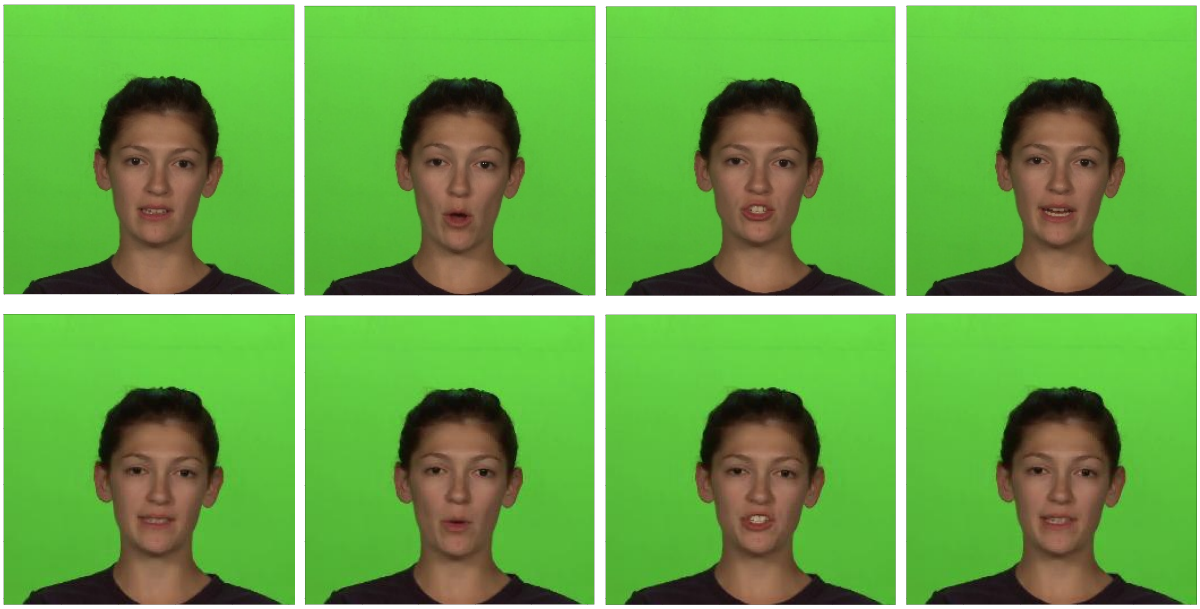


Figure 5.2: Video frame examples and their corresponding synthesized frames. The top row consists of ground truth images, while the bottom contains synthesized images.

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), Frechet Inception Distance (FID), and Learned Perceptual Image Patch (LPIPS) (details in Section 3.2).

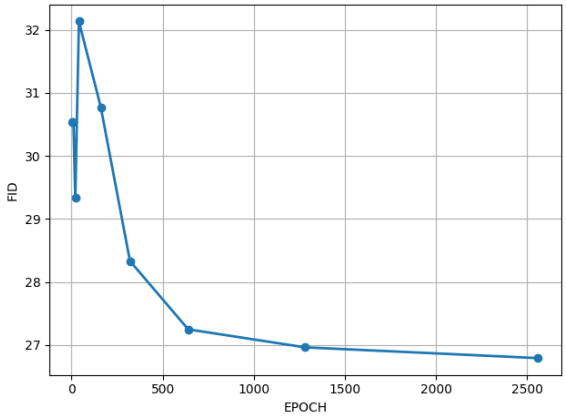
Table 5.2: Objective scores were computed using synthesized and ground-truth 2D landmarks as input to the second stage of our pipeline. The arrows up indicate that higher is better, while the arrows down indicate that lower is better. We see that *FaceFormer* training successfully learns facial shape 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.

Ep	**FID** ↓		**LPIPS** ↓		**SSIM** ↑	
	μ	$\pm\sigma$	μ	$\pm\sigma$	μ	$\pm\sigma$
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

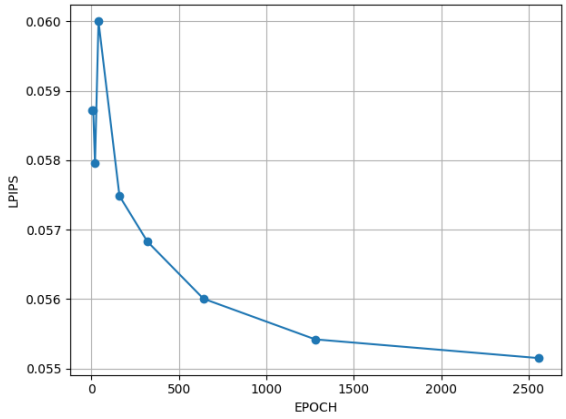
The initial rows of Table 5.2 display a consistent decrease in FID and LPIPS

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 the potential for even better results with extended training. Figure 5.3 graphically displays the average score metrics over the epochs.

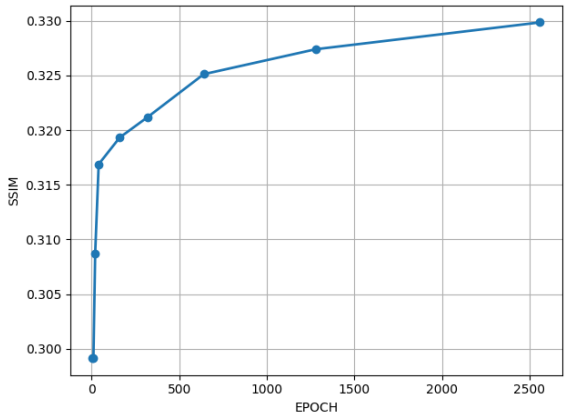
The final row of Table 5.2 presents the scores obtained when ground-truth landmarks are input to the second stage. Although using ground truth landmarks yields better photorealism in animations, the scores are comparatively close to those obtained using the fully synthetic pipeline.



(a) FID ↓



(b) LPIPS ↓



(c) SSIM ↑

Figure 5.3: Objective metrics score over the *FaceFormer* training epochs, the score refer to the mean of k-fold experiments, with $k = 4$ and each fold containing thirteen samples. Objective Metrics: (a) FID score, indicating the distance between distributions of generated and real images; (b) LPIPS score, reflecting perceptual similarity to human judgment; and (c) SSIM score, measuring the similarity between generated images and ground truth.

The graphs clearly show a tendency for further improvement if the model is trained over 2500 epochs. However, it is unclear what the benefits of perceptual evaluation would be while raising the training in 1000 epochs means training for more than 3 days.

5.3 Ablation Study

We conduct an ablation study on the first-stage decoder to analyze the impact of each component on landmark synthesis. Each test is designed to isolate and evaluate the significance of specific architectural elements. We utilized a k-fold test dataset and trained the model with various modifications. The components and the respective tests are outlined below, with justifications for each choice:

- Periodic Positional Embedding (PPE): In the Transformer architecture, PPE is added to the input embeddings at the initial stage to provide the model with information about the order or position of tokens in the sequence

  1. *Removing PPE from the pipeline.* Removing PPE helps determine if the model can still effectively synthesize accurate landmarks without recurrent positional information.
  2. *Replacing PPE with the original Positional Embedding (PE).* By removing the periodic aspect, we assess whether the periodicity in positional embeddings is crucial for the model's performance (see Equation 2.2). The original PE is described in the subsequent equation:

$$\begin{aligned} PE(pos, 2i) &= \sin\left(\frac{pos}{10000^{2i/d}}\right) \\ PE(pos, 2i+1) &= \cos\left(\frac{pos}{10000^{2i/d}}\right) \end{aligned} \tag{5.1}$$

- Temporal Bias: The Temporal Bias B^T is employed in the first Multi-Head Attention layer of the Transformer decoder as illustrated in the following equation:

$$\text{Att}\left(Q^F, K^F, V^F, B^T\right) = \text{softmax}\left(\frac{Q^F (K^F)^T}{\sqrt{d_k}} + B^T\right) V^F.$$

  1. *Removing temporal bias weights.* This test assigns equal weight to all past predicted landmarks, allowing us to understand the importance of weighting past frames differently for the prediction of current frames. The following matrix B^{T^1} exemplifies it:

$$B^{T^1} = \begin{pmatrix} 0 & -\infty & -\infty & -\infty & -\infty \\ 0 & 0 & -\infty & -\infty & -\infty \\ 0 & 0 & 0 & -\infty & -\infty \\ 0 & 0 & 0 & 0 & -\infty \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

  2. *Implementing the original ALiBi without periodicity.* By removing the periodic aspect, we assess if periodic temporal bias is more effective for synthesizing facial

landmarks than the traditional ALiBi method (see Section 4.3.2). The following matrix B^{T^2} exemplifies it:

$$B^{T^2} = \begin{pmatrix} 0 & -\infty & -\infty & -\infty & -\infty \\ -1 & 0 & -\infty & -\infty & -\infty \\ -2 & -1 & 0 & -\infty & -\infty \\ -3 & -2 & -1 & 0 & -\infty \\ -4 & -3 & -2 & -1 & 0 \end{pmatrix}$$

- Alignment Bias: The Alignment Bias B^A is applied in the second Multi-Head Attention layer of the Transformer decoder, as depicted in the equation below:

$$\text{Att}\left(Q^F, K^F, V^F, B^A\right) = \text{softmax}\left(\frac{Q^F (K^F)^T}{\sqrt{d_k}} + B^A\right) V^F.$$

  1. *Adding one more past feature to the Alignment Bias mask.* This test evaluates the effect of including additional past features in the alignment process, determining if more past information improves the alignment of audio features with facial landmarks. The following matrix B^{A^1} exemplifies it:

$$B^{A^1} = \begin{pmatrix} 0 & -\infty & -\infty & -\infty & -\infty \\ 0 & 0 & -\infty & -\infty & -\infty \\ -\infty & 0 & 0 & -\infty & -\infty \\ -\infty & -\infty & 0 & 0 & -\infty \\ -\infty & -\infty & -\infty & 0 & 0 \end{pmatrix}$$

  2. *Adding ten more past features to the Alignment Bias mask.* By significantly increasing the number of past features considered, we assess the impact of extensive past information on the alignment and overall performance.

To evaluate, we apply objective metrics to both the facial landmarks and the synthesized realistic frames. We employ LMD to evaluate the *FaceFormer* decoder output, the facial landmarks. Due to limitations in computer resources, we do not apply visual objective metrics on the Alignment Bias experiment (details in Section 5.3.2).

5.3.1 Periodic Positional Embedding and Temporal Bias

Figure 5.4 displays the LMD applied to lip landmarks over the epochs for the tests conducted on the PPE and Temporal Bias components. Compared with the complete *FaceFormer* decoder, modifications to the Temporal Bias have a more significant impact on LMD. It is observable that removing the weights from the Temporal Bias was the modification that most increased the LMD error. Additionally, using PEs without the periodic aspect and completely removing it has the same impact as the epochs increase.

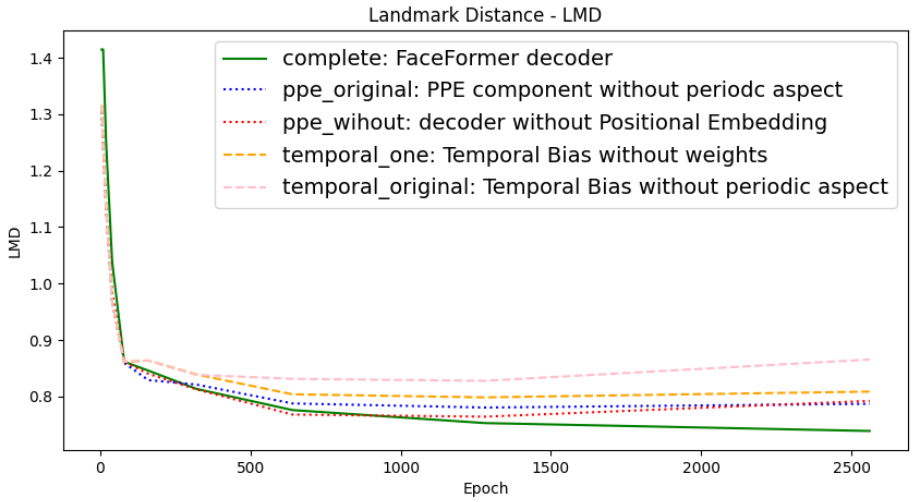
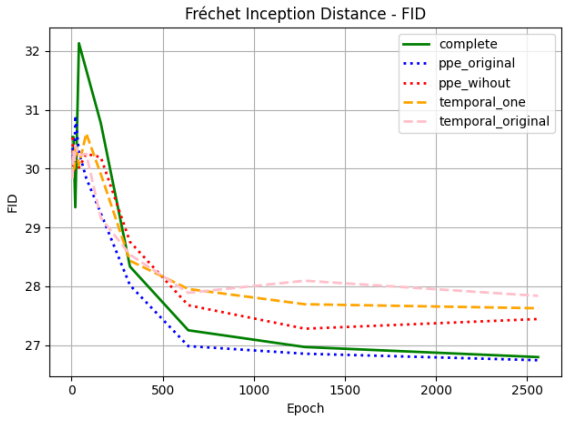
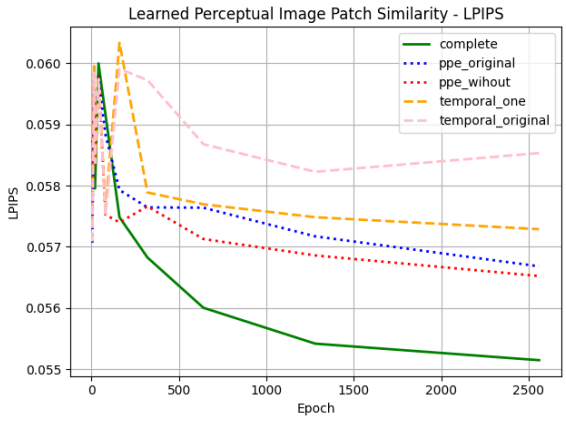


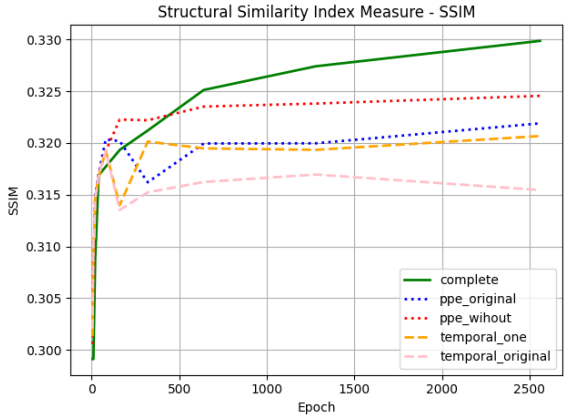
Figure 5.4: LMD of lip landmarks of each component experiment over the *FaceFormer* training epochs.



(a) FID ↓



(b) LPIPS ↓



(c) SSIM ↑

Figure 5.5: Visual objective metrics score of each component experiment over the *FaceFormer* training epochs. Objective Metrics: (a) FID score, (b) LPIPS score, and (c) SSIM score.

Figure 5.5 presents the scores from objective visual metrics. Following LMD, Temporal Bias shows a more significant impact. The greatest impact occurs when we employ the second test with the original ALiBi implementation (see Section 4.3.2).

5.3.2 Alignment Bias

We found that increased computational resources were needed when conducting the tests to evaluate the Alignment Bias. Adding one more context feature increased the training epoch time from 2 minutes to 15 minutes. When adding ten more, the time increased to almost 2 hours. Given this limitation, we trained the model for fewer epochs. Figure 5.6 displays the LMD scores over 360 epochs. The test of Alignment Bias with ten more context features was trained for only 160 epochs due to the increased time required. It is observable that with just one additional context, the score is below the complete framework results but closer, and with ten more context features, the LMD presents the worst score among the ablation tests. Given the time required to add more context features, it is not worth adding more.

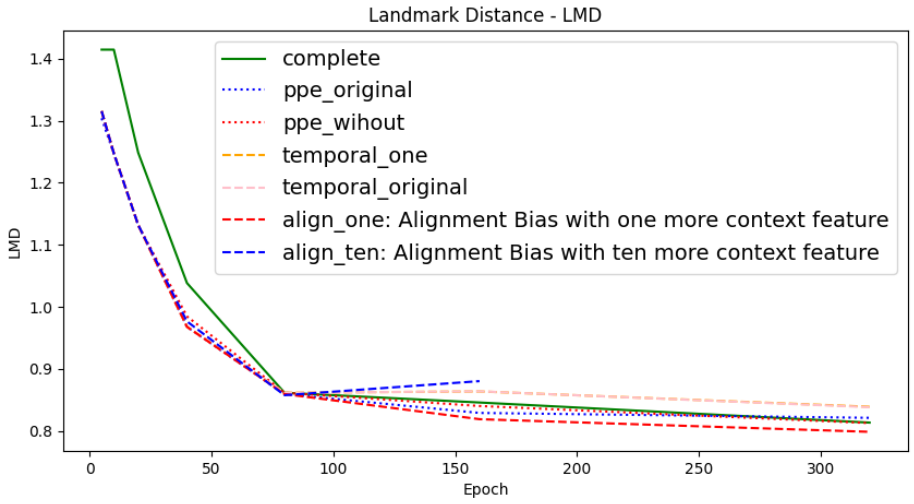


Figure 5.6: LMD of lip landmarks over the epochs with all ablation study components.

With 360 epochs, the framework cannot synthesize sufficiently accurate lip landmarks to perform lip synchronization even when complete. Therefore, we did not perform the objective visual metrics on the Alignment Bias test videos.

5.4 Exploratory Test

We conducted an exploratory approach to assess our model's ability to handle audio inputs from sources outside our dataset. The capability to effectively process and

synthesize talking head videos from unfamiliar audio sources indicates the framework's robustness and potential applicability in real-world scenarios. Additionally, these exploratory tests helped identify limitations and areas for improvement in our system's design and training process.

A diverse selection of audio samples from external sources was compiled for the exploratory tests. These samples were deliberately chosen to cover a broad spectrum of characteristics not present in the training dataset, including variations in language, gender, and speech speed. The videos can be viewed at

- **Language**: In the exploratory tests, the framework showed some potential in processing languages that were not part of its original training dataset. This potential was observed in the framework's use of visemes representations of phonemes—originally learned from Portuguese samples, which seemed to assist in creating talking head videos in English with a reasonable level of lip synchronization.

- **Gender**: Although our training used only a female voice, there were indications that the framework could also handle male voices. These initial results hint that the first-stage encoder output, the *wav2vec2.0* speech features, possess an element of gender neutrality, potentially allowing the framework to supply various genders despite not being explicitly trained on a diverse set of voices.

- **Speed**: Our framework displayed some promising tendencies in handling different speech speeds, as observed in videos where speech starts at a normal pace and ends with words pronounced more slowly, emphasizing each syllable. The system seemed to adapt the facial landmarks correspondingly, aiming for synchronization with the audio pace. Additionally, in videos with periods of silence, the framework accurately simulated the talking head closing its mouth, mirroring natural speech pauses.

These preliminary findings indicate a potential synergy in our setup between the *wav2vec2.0* model and the *FaceFormer* decoder. Together, they appeared to handle variations in languages, genders, and speech speeds to a certain degree, suggesting a capability of speech generalization for synthesizing talking head videos with a level of realism. However, further perceptual testing may be necessary to ensure our observations.

5.5 Concluding Remarks

The objective metrics results indicate that the synthesized facial landmarks score closely to the ground truth ones (see Section 5.2). One can see good lip synchronization with photorealistic texture by observing the generated talking head videos from the test dataset. However, despite these positive outcomes, it was observed that the talking heads do not blink (see Section 5.1).

Our Ablation Study on the *FaceFormer* decoder components shows that the Temporal Bias test most significantly increased the error in the synthesized facial landmarks (see Section 5.3). Despite the lower scores, the differences in the generated videos are subtle, with the ablation study videos maintaining good lip synchronization, similar to those produced by the complete framework. FAN *et al.* argued that these modifications were essential for synthesizing 3D meshes (five thousand 3D points), but our findings suggest that for synthesizing simple 2D landmarks (sixty-eight 2D points), they may not be so important. To confirm our findings, perceptual tests are necessary to determine which modules do not significantly impact subjective human assessment, potentially simplifying the architecture. Additionally, the perceptual test is crucial for delving deeper into our exploratory tests to assess the model's generalization capacity (see Section 5.4).

In the upcoming chapter, we'll summarize our findings and discuss research question conclusions and future directions. This final chapter aims to highlight key insights and chart a path forward, suggesting opportunities for advancing this field.

Chapter 6

Conclusions

To the best of our knowledge, our work builds the first two-stage speech-driven neural deep learning-based 2D talking head for Brazilian Portuguese. This approach focuses on the realistic synthesis of speech articulatory movements and facial appearance.

In Chapter 2, we started by introducing GAN and Transformer, the two based-architectures employed in this work. We then provide a brief historical overview of talking head synthesis, focusing on two-stage landmark-based approaches. We summarize the works reviewed, detailing aspects such as the first and second-stage architectures, objective evaluation metrics, and datasets used for training and validation. Additionally, we present some examples of Brazilian Portuguese talking heads. Finally, we discuss how our work aligns with the reviewed literature.

Building a photorealistic, speech-driven talking head is a challenging task that involves a complex engineering process. Chapter 3 outlined the dataset used to train our framework, the computational resources, the landmark extraction method, and the objective evaluation methods used to assess our results. In Chapter 4, we presented the methodology for constructing our framework. This includes the data preprocessing, detailing the models used (*FaceFormer* and *vid2vid*), and describing the training and inference processes.

Chapter 5 presents our results. First, we employed objective metrics (with methods such as SSIM, FID, and LPIPS) to evaluate the image quality of our synthesized talking head by applying our test dataset. Then, we employed an ablation study to try to understand the impact of each *FaceFormer* component. Finally, we explored the model's capacity to generalize to input audio inputs from sources outside our dataset. After conducting these experiments, we were able to answer the research questions proposed in Chapter 1:

- Considering that the *FaceFormer* model was originally designed to synthesize 3D meshes, our central research question is: **How does *FaceFormer* model perform in translating speech audio signals to realistic dynamic behavior of 2D**

facial landmarks?

The outcomes of our objective tests suggest that *FaceFormer* has been successfully adapted to synthesize 2D landmarks. We used *FaceFormer* to create 2D landmarks and input them into the *vid2vid* model to synthesize realistic talking head frames. Simultaneously, we input ground truth 2D landmarks into *vid2vid* and compared the results. The final row of Table 5.2 presents the scores. Although using ground truth landmarks results in better photorealism in animations, the scores are relatively close.

- **Can our framework synthesize high-quality talking heads with the available dataset volume?**

  This model was exclusively trained on a modest-sized dataset, just 15 minutes, consisting of Portuguese audio at a standard speech speed. Despite the Table 5.2 results, the synthesized talking head videos generated from the test dataset demonstrate good lip synchronization with words not seen in the training dataset.

- **How far can our framework generalize to other speech agents and styles?**

  We explore the model by testing it with audio outside our dataset; the model indicates the potential to handle variations in speech speed, gender, and other languages. Although the model suggests this potential, perceptual testing is needed to confirm our observations.

Despite the advancements, one notable limitation was observed: the models did not simulate eye blinking. However, when we used ground truth landmarks as input for the rendering second stage, the model was able to produce blinking eyes. Specifically, the first stage model struggled to capture the temporal dynamics of eye blinking. This suggests that while *vid2vid2* model can effectively render eye movements when given accurate landmarks, there is a gap in *FaceFormer* ability to learn and predict the natural blinking motion autonomously in the initial landmark detection phase. These results affirm the significance of well-defined facial landmarks and showcase the model's adaptability and potential for broader applications beyond its initial training constraints.

Even though advancements in rendering realistic features and expressions through algorithms and neural networks, the uncanny valley remains a critical boundary between the synthetic and the real. This psychological phenomenon, where the near-realistic representations of humans evoke feelings of strangeness, may highlight a possible fundamental limit to how human-like our creations can truly become. As we push the boundaries of what machines can achieve, reflecting on the uncanny valley forces us to consider the possibility that there might always be a gap between artificial creations and human authenticity. This introspection not only shapes our technological pursuits but also deepens our understanding of human perception and our intrinsic responses to replicas of ourselves.

6.1 Future Work

While this study marks a significant step forward in synthesizing talking head videos, it also opens several avenues for future research. First, exploring the integration of emotion and gesture synthesis could further enhance the realism and expressiveness of the talking heads, making them more relatable and engaging for users. Extending the framework to include a wider range of languages and dialects would reinforce its universality and accessibility. Another promising direction is optimizing the framework for real-time applications, such as live broadcasting or interactive virtual assistants. This would necessitate advancements in computational efficiency without compromising the quality of the output. Additionally, addressing the challenge of non-blinking in talking heads is crucial. Implementing a blink synthesis solution could improve the naturalism of facial animations, making the characters seem more alive.

In addition, while recognizing the valuable insights offered by objective metrics like SSIM, LPIPS, and FID in quantifying visual fidelity, we acknowledge their limitations in comprehensively evaluating the quality of synthesized talking heads. These metrics can be good 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. Therefore, we plan to complement objective metrics with subjective evaluation by human observers. Additionally, the results from the Ablation study suggest that some *FaceFormer* decoder components may not be crucial for synthesizing simple 2D landmarks. Subjective evaluation can help confirm our findings and determine which modules do not significantly impact subjective human assessment, potentially simplifying the architecture.

Lastly, ethical considerations and the potential to misuse talking head technologies call for research into safeguards and ethical guidelines. This includes developing methods to detect and flag synthetic video content to prevent misinformation and protect individual rights.

References

ABDUL, Z. K.; AL-TALABANI, A. K. Mel frequency cepstral coefficient and its applications: A review. *IEEE Access*, v. 10, p. 122136–122158, 2022. Available from Internet: <https://doi.org/10.1109/ACCESS.2022.3223444>.

ARDILA, R.; BRANSON, M.; DAVIS, K.; KOHLER, M.; MEYER, J.; HENRETTY, M.; MORAIS, R.; SAUNDERS, L.; TYERS, F.; WEBER, G. Common voice: A massively-multilingual speech corpus. In: CALZOLARI, N.; BÉCHET, F.; BLACHE, P.; CHOUKRI, K.; CIERI, C.; DECLERCK, T.; GOGGI, S.; ISAHARA, H.; MAEGAARD, B.; MARIANI, J.; MAZO, H.; MORENO, A.; ODIJK, J.; PIPERIDIS, S. (Ed.). *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2020. p. 4218–4222. ISBN 979-10-95546-34-4. Available from Internet: <https://aclanthology.org/2020.lrec-1.520>.

BAEVSKI, A.; ZHOU, Y.; MOHAMED, A.; AULI, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. Curran Associates, Inc., v. 33, p. 12449–12460, 2020. Available from Internet: <https://doi.org/10.48550/arXiv.2006.11477>.

BAHDANAU, D.; CHO, K.; BENGIO, Y. *Neural Machine Translation by Jointly Learning to Align and Translate*. 2016. Available from Internet: <https://doi.org/10.48550/arXiv.1409.0473>.

BAI, S.; KOLTER, J. Z.; KOLTUN, V. *An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling*. 2018. Available from Internet: <https://doi.org/10.48550/arXiv.1803.01271>.

BELGHAZI, M. I.; BARATIN, A.; RAJESWAR, S.; OZAIR, S.; BENGIO, Y.; COURVILLE, A.; HJELM, R. D. Mutual information neural estimation. In: DY, J.; KRAUSE, A. (Ed.). *INTERNATIONAL CONFERENCE ON MACHINE LEARNING, VOL 80*. [s.n.], 2018. (Proceedings of Machine Learning Research, v. 80). 35th International Conference on Machine Learning (ICML), Stockholm, SWEDEN, JUL 10-15, 2018. Available from Internet: <https://doi.org/10.48550/arXiv.1801.04062>.

BERNARDO, B.; COSTA, P. A speech-driven talking head based on a two-stage generative framework. In: GAMALLO, P.; CLARO, D.; TEIXEIRA, A.; REAL, L.; GARCIA, M.; OLIVEIRA, H. G.; AMARO, R. (Ed.). *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*. Santiago de Compostela, Galicia/Spain: Association for Computational Linguistics, 2024. p. 580–586. Available from Internet: <https://aclanthology.org/2024.propor-1.64>.

BRAND, M. Voice puppetry. In: *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*. USA: ACM Press/Addison-Wesley Publishing

Co., 1999. (SIGGRAPH '99), p. 21–28. ISBN 0201485605. Available from Internet: <https://doi.org/10.1145/311535.311537>.

BREGLER, C.; COVELL, M.; SLANEY, M. Video rewrite: Driving visual speech with audio. In: *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*. USA: ACM Press/Addison-Wesley Publishing Co., 1997. (SIGGRAPH '97), p. 353–360. ISBN 0897918967. Available from Internet: <https://doi.org/10.1145/258734.258880>.

BULAT, A.; TZIMIROPOULOS, G. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. *CoRR*, abs/1703.00862, 2017. Available from Internet: <http://arxiv.org/abs/1703.00862>.

BULAT, A.; TZIMIROPOULOS, G. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). IEEE, oct 2017. Available from Internet: <https://doi.org/10.1109%2Ficcv.2017.116>.

CECH, J.; SOUKUPOVA, T. Real-time eye blink detection using facial landmarks. *Cent. Mach. Perception, Dep. Cybern. Fac. Electr. Eng. Czech Tech. Univ. Prague*, p. 1–8, 2016. Available from Internet: <https://vision.fe.uni-lj.si/cvww2016/proceedings/papers/05.pdf>.

CHEN, L.; CUI, G.; KOU, Z.; ZHENG, H.; XU, C. What comprises a good talking-head video generation? In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. [s.n.], 2020. Available from Internet: <10.48550/arXiv.2005.03201>.

CHEN, L.; LI, Z.; MADDOX, R. K.; DUAN, Z.; XU, C. Lip movements generation at a glance. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. [s.n.], 2018. Available from Internet: <https://doi.org/10.48550/arXiv.1803.10404>.

CHEN, L.; MADDOX, R. K.; DUAN, Z.; XU, C. *Hierarchical Cross-Modal Talking Face Generation With Dynamic Pixel-Wise Loss*. Los Alamitos, CA, USA: IEEE Computer Society, 2019. 7824-7833 p. Available from Internet: <https://doi.ieeecomputersociety.org/10.1109/CVPR.2019.00802>.

CHO, K.; MERRIENBOER, B. van; GULCEHRE, C.; BAHDANAU, D.; BOUGARES, F.; SCHWENK, H.; BENGIO, Y. *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*. 2014. Available from Internet: <https://doi.org/10.48550/arXiv.1406.1078>.

CHUNG, J.; ZISSERMAN, A. Lip reading in profile. In: BRITISH MACHINE VISION ASSOCIATION AND SOCIETY FOR PATTERN RECOGNITION. *British Machine Vision Conference, 2017*. 2017. Available from Internet: <https://doi.org/10.48550/arXiv.1406.1078>.

CHUNG, J. S.; NAGRANI, A.; ZISSERMAN, A. VoxCeleb2: Deep Speaker Recognition. In: *Proc. Interspeech 2018*. [s.n.], 2018. p. 1086–1090. Available from Internet: <https://doi.org/10.48550/arXiv.1806.05622>.

CHUNG, J. S.; SENIOR, A.; VINYALS, O.; ZISSERMAN, A. Lip reading sentences in the wild. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [s.n.], 2017. p. 3444–3453. Available from Internet: <https://doi.org/10.1109/CVPR.2017.367>.

CHUNG, J. S.; ZISSERMAN, A. Out of time: automated lip sync in the wild. In: *Workshop on Multi-view Lip-reading, ACCV*. [s.n.], 2016. Available from Internet: <https://doi.org/10.1007/978-3-319-54427-4_19>.

CHUNG, J. S.; ZISSERMAN, A. Lip reading in the wild. In: SPRINGER. *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*. 2017. p. 87–103. Available from Internet: <https://doi.org/10.1007/978-3-319-54184-6_6>.

CHUNG, Y.-A.; GLASS, J. Generative pre-training for speech with autoregressive predictive coding. In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [s.n.], 2020. p. 3497–3501. Available from Internet: <https://doi.org/10.1109/ICASSP40776.2020.9054438>.

COOKE, M.; BARKER, J.; CUNNINGHAM, S.; SHAO, X. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, v. 120, n. 5, p. 2421–2424, 11 2006. ISSN 0001-4966. Available from Internet: <https://doi.org/10.1121/1.2229005>.

COSKER, D.; MARSHALL, D.; ROSIN, P.; HICKS, Y. Speech driven facial animation using a hidden markov coarticulation model. In: *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* [s.n.], 2004. v. 1, p. 128–131 Vol.1. Available from Internet: <https://doi.org/10.1109/ICPR.2004.1334024>.

COSTA, P. D. P. *Two-Dimensional Expressive Speech Animation*. Tese (Doutorado) — Universidade Estadual de Campinas, 2015. Available from Internet: <https://doi.org/10.13140/RG.2.1.3131.6968>.

CROITORU, F.-A.; HONDRU, V.; IONESCU, R. T.; SHAH, M. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 45, n. 9, p. 10850–10869, 2023.

DAS, D.; BISWAS, S.; SINHA, S.; BHOWMICK, B. 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*. Berlin, Heidelberg: Springer-Verlag, 2020. p. 408–424. ISBN 978-3-030-58576-1. Available from Internet: <https://doi.org/10.1007/978-3-030-58577-8_25>.

De Martino, J. M.; Pini Magalhães, L.; VIOLARO, F. Facial animation based on context-dependent visemes. *Computers Graphics*, v. 30, n. 6, p. 971–980, 2006. ISSN 0097-8493. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0097849306001518>.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. Available from Internet: <https://doi.org/10.18653/v1/N19-1423>.

DIEDERICH, S.; BRENDEL, A. B.; MORANA, S.; KOLBE, L. On the design of and interaction with conversational agents: An organizing and assessing review of human-computer interaction research. *Journal of the Association for Information Systems*, v. 23, n. 1, p. 96–138, 2022. Available from Internet: <https://doi.org/10.17705/1jais.00724>.

EZZAT, T.; POGGIO, T. Visual speech synthesis by morphing visemes. *International Journal of Computer Vision*, Springer, v. 38, p. 45–57, 2000. Available from Internet: <https://doi.org/10.1023/A:1008166717597>.

FAN, B.; WANG, L.; SOONG, F. K.; XIE, L. Photo-real talking head with deep bidirectional lstm. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [s.n.], 2015. p. 4884–4888. Available from Internet: <https://doi.org/10.1109/ICASSP.2015.7178899>.

FAN, Y.; LIN, Z.; SAITO, J.; WANG, W.; KOMURA, T. *FaceFormer: Speech-Driven 3D Facial Animation with Transformers*. 2022. 18749–18758 p. Available from Internet: <https://10.1109/CVPR52688.2022.01821>.

FINN, C.; ABBEEL, P.; LEVINE, S. Model-agnostic meta-learning for fast adaptation of deep networks. In: PRECUP, D.; TEH, Y. W. (Ed.). *Proceedings of the 34th International Conference on Machine Learning*. PMLR, 2017. (Proceedings of Machine Learning Research, v. 70), p. 1126–1135. Available from Internet: <https://proceedings.mlr.press/v70/finn17a.html>.

GOODFELLOW, I.; POUGET-ABADIE, J.; MIRZA, M.; XU, B.; WARDE-FARLEY, D.; OZAIR, S.; COURVILLE, A.; BENGIO, Y. *Generative adversarial networks*. ACM New York, NY, USA, 2020. 139–144 p. Available from Internet: <https://doi.org/10.1145/3422622>.

GRETTON, A.; BORGWARDT, K. M.; RASCH, M. J.; SCHÖLKOPF, B.; SMOLA, A. A kernel two-sample test. *The Journal of Machine Learning Research*, JMLR. org, v. 13, n. 1, p. 723–773, 2012. Available from Internet: <https://www.jmlr.org/papers/volume13/gretton12a/gretton12a.pdf?ref=https://githubhelp.com>.

GROSMAN, J. *Fine-tuned XLSR-53 large model for speech recognition in Portuguese*. 2021. <https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-portuguese>.

GUI, J.; SUN, Z.; WEN, Y.; TAO, D.; YE, J. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Transactions on Knowledge and Data Engineering*, v. 35, n. 4, p. 3313–3332, 2023. Available from Internet: <https://doi.org/10.1109/TKDE.2021.3130191>.

GUI, J.; SUN, Z.; WEN, Y.; TAO, D.; YE, J. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Transactions on Knowledge and Data Engineering*, v. 35, n. 4, p. 3313–3332, 2023.

GUO, M.-H.; XU, T.-X.; LIU, J.-J.; LIU, Z.-N.; JIANG, P.-T.; MU, T.-J.; ZHANG, S.-H.; MARTIN, R. R.; CHENG, M.-M.; HU, S.-M. Attention mechanisms in computer vision: A survey. *Computational Visual Media*, Springer Science and Business Media LLC, v. 8, n. 3, p. 331–368, mar. 2022. ISSN 2096-0662. Available from Internet: <http://dx.doi.org/10.1007/s41095-022-0271-y>.

GUO, Y.; CHEN, K.; LIANG, S.; LIU, Y.-J.; BAO, H.; ZHANG, J. *AD-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis*. 2021. Available from Internet: <https://doi.org/10.1109/ICCV48922.2021.00573>.

HANNUN, A.; CASE, C.; CASPER, J.; CATANZARO, B.; DIAMOS, G.; ELSEN, E.; PRENGER, R.; SATHEESH, S.; SENGUPTA, S.; COATES, A.; NG, A. Y. *Deep Speech: Scaling up end-to-end speech recognition*. 2014. Available from Internet: <https://ui.adsabs.harvard.edu/link_gateway/2014arXiv1412.5567H/doi:10.48550/arXiv.1412.5567>.

HARTE, N.; GILLEN, E. Tcd-timit: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia*, v. 17, n. 5, p. 603–615, 2015. Available from Internet: <https://doi.org/10.1109/TMM.2015.2407694>.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [s.n.], 2016. Available from Internet: <https://doi.org/10.48550/arXiv.1512.03385>.

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. Curran Associates, Inc., v. 30, 2017. Available from Internet: <10.48550/arXiv.1706.08500>.

ILG, E.; MAYER, N.; SAIKIA, T.; KEUPER, M.; DOSOVITSKIY, A.; BROX, T. Flownet 2.0: Evolution of optical flow estimation with deep networks. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [s.n.], 2017. p. 1647–1655. Available from Internet: <https://doi.org/10.1109/CVPR.2017.179>.

ISLAM, S.; ELMEKKI, H.; ELSEBAI, A.; BENTAHAR, J.; DRAWEL, N.; RJOUB, G.; PEDRYCZ, W. A comprehensive survey on applications of transformers for deep learning tasks. *Expert Systems with Applications*, v. 241, p. 122666, 2024. ISSN 0957-4174. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0957417423031688>.

ISOLA, P.; ZHU, J.-Y.; ZHOU, T.; EFROS, A. A. Image-to-image translation with conditional adversarial networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [s.n.], 2017. p. 1125–1134. Available from Internet: <https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.632>.

JALALIFAR, S. A.; HASANI, H.; AGHAJAN, H. *Speech-Driven Facial Reenactment Using Conditional Generative Adversarial Networks*. 2018. Available from Internet: <https://doi.org/10.48550/arXiv.1803.07461>.

Jesus Filho, S. R. d. Master's thesis, *HMM-based expressive facial animation synthesis*. 2021. Available at <https://hdl.handle.net/20.500.12733/1641994>.

JOHNSON, J.; ALAHI, A.; FEI-FEI, L. Perceptual losses for real-time style transfer and super-resolution. In: LEIBE, B.; MATAS, J.; SEBE, N.; WELLING, M. (Ed.). *Computer Vision – ECCV 2016*. Cham: Springer International Publishing, 2016. p. 694–711. ISBN 978-3-319-46475-6. Available from Internet: <https://doi.org/10.1007/978-3-319-46475-6_43>.

KING, D. E. Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.*, JMLR.org, v. 10, p. 1755–1758, dec 2009. ISSN 1532-4435. Available from Internet: <https://doi.org/10.5555/1577069.1755843>.

KINGMA, D. P.; WELLING, M. *et al.* An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, Now Publishers, Inc., v. 12, n. 4, p. 307–392, 2019.

LIN, T.; WANG, Y.; LIU, X.; QIU, X. A survey of transformers. *AI Open*, v. 3, p. 111–132, 2022. ISSN 2666-6510. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S2666651022000146>.

LU, Y.; CHAI, J.; CAO, X. Live speech portraits: Real-time photorealistic talking-head animation. *ACM Trans. Graph.*, Association for Computing Machinery, New York, NY, USA, v. 40, n. 6, dec 2021. ISSN 0730-0301. Available from Internet: <https://doi-org.ez88.periodicos.capes.gov.br/10.1145/3478513.3480484>.

MARTINO, J. M. D. Assessing the visual speech perception of sampled-based talking heads. In: *Proc. of the 12th International Conference on Auditory-Visual Speech Processing. (AVSP)*. [S.l.: s.n.], 2013.

MATTHEYSES, W.; VERHELST, W. Audiovisual speech synthesis: An overview of the state-of-the-art. *Speech Communication*, v. 66, p. 182–217, 2015. ISSN 0167-6393. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0167639314000818>.

MCGURK, M. J. Hearing lips and seeing voices. Nature, p. 746–748, 1976. Available from Internet: <https://doi.org/10.1038/264746a0>.

MIRZA, M.; OSINDERO, S. *Conditional Generative Adversarial Nets*. 2014. Available from Internet: <https://ui.adsabs.harvard.edu/link_gateway/2014arXiv1411.1784M/doi:10.48550/arXiv.1411.1784>.

MORI, M.; MACDORMAN, K. F.; KAGEKI, N. The uncanny valley [from the field]. *IEEE Robotics Automation Magazine*, v. 19, n. 2, p. 98–100, 2012. Available from Internet: <https://doi.org/10.1109/MRA.2012.2192811>.

NEWELL, A.; YANG, K.; DENG, J. Stacked hourglass networks for human pose estimation. In: SPRINGER. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*. 2016. p. 483–499. Available from Internet: <https://doi.org/10.1007/978-3-319-46484-8_29>.

NOH, J.-Y.; NEUMANN, U. Talking faces. In: IEEE. *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532)*. 2000. v. 2, p. 627–630. Available from Internet: <https://doi.org/10.1109/ICME.2000.871441>.

OORD, A. van den; DIELEMAN, S.; ZEN, H.; SIMONYAN, K.; VINYALS, O.; GRAVES, A.; KALCHBRENNER, N.; SENIOR, A.; KAVUKCUOGLU, K. Wavenet: A generative model for raw audio. In: *Arxiv*. [s.n.], 2016. Available from Internet: <https://arxiv.org/abs/1609.03499>.

PARKE, F. I. Computer generated animation of faces. In: *Proceedings of the ACM Annual Conference - Volume 1*. New York, NY, USA: Association for Computing Machinery, 1972. (ACM '72), p. 451–457. ISBN 9781450374910. Available from Internet: <https://doi.org/10.1145/800193.569955>.

PORTUGUESA, C. C. de Inteligência Artificial e Tecnologia da L. *Apelo pelos investigadores científicos sobre Inteligência Artificial e Tecnologia da Língua Portuguesa*. 2024. <https://propor.org/wp-content/uploads/2024/03/apeloPROPOR2024.pdf>. Accessed: 2024-05-10.

PRESS, O.; SMITH, N.; LEWIS, M. Train short, test long: Attention with linear biases enables input length extrapolation. 2021.

PUSHPAKUMAR, R.; SANJAYA, K.; RATHIKA, S.; ALAWADI, A. H.; MAKHZUNA, K.; VENKATESH, S.; RAJALAKSHMI, B. Human-computer interaction: Enhancing user experience in interactive systems. In: EDP SCIENCES. *E3S Web of Conferences*. 2023. v. 399, p. 04037. Available from Internet: <https://doi.org/10.1051/e3sconf/202339904037>.

QIAN, K.; ZHANG, Y.; CHANG, S.; YANG, X.; HASEGAWA-JOHNSON, M. Autovc: Zero-shot voice style transfer with only autoencoder loss. In: PMLR. *International Conference on Machine Learning*. 2019. p. 5210–5219. Available from Internet: <https://doi.org/10.48550/arXiv.1905.05879>.

RADFORD, A.; METZ, L.; CHINTALA, S. *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*. 2016. Available from Internet: <https://doi.org/10.48550/arXiv.1511.06434>.

REIS, F. A. d. B. Master's thesis, *A generative adversarial network approach to visual expressive speech synthesis with emotion control*. 2020. Available from Internet: <https://doi.org/10.47749/T/UNICAMP.2020.1149452>.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: NAVAB, N.; HORNEGGER, J.; WELLS, W. M.; FRANGI, A. F. (Ed.). *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Cham: Springer International Publishing, 2015. p. 234–241. ISBN 978-3-319-24574-4. Available from Internet: <https://doi.org/10.1007/978-3-319-24574-4_28>.

RöSSLER, A.; COZZOLINO, D.; VERDOLIVA, L.; RIESS, C.; THIES, J.; NIEßNER, M. *FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces*. 2018. Available from Internet: <https://ui.adsabs.harvard.edu/link_gateway/2018arXiv180309179R/doi:10.48550/arXiv.1803.09179>.

SCOTT, K. C.; KAGELS, D.; WATSON, S.; ROM, H.; WRIGHT, J.; LEE, M.; HUSSEY, K. Synthesis of speaker facial movement to match selected speech sequences. In: CITESEER. *Proceedings of the Fifth Australian Conference on Speech Science and Technology*. [S.l.], 1994. v. 2, p. 620–625.

SEYMOUR, M.; YUAN, L. I.; DENNIS, A.; RIEMER, K. *et al.* Have we crossed the uncanny valley? understanding affinity, trustworthiness, and preference for realistic digital humans in immersive environments. *Journal of the Association for Information Systems*, v. 22, n. 3, p. 9, 2021. Available from Internet: <https://doi.org/10.17705/1jais.00674>.

SHENG, C.; KUANG, G.; BAI, L.; HOU, C.; GUO, Y.; XU, X.; PIETIKäINEN, M.; LIU, L. *Deep Learning for Visual Speech Analysis: A Survey*. 2024. 1-20 p. Available from Internet: <https://doi.org/10.1109/TPAMI.2024.3376710>.

SINHA, S.; BISWAS, S.; BHOWMICK, B. Identity-preserving realistic talking face generation. In: *2020 International Joint Conference on Neural Networks (IJCNN)*. [s.n.], 2020. p. 1–10. Available from Internet: <https://doi.org/10.1109/IJCNN48605.2020.9206665>.

SUWAJANAKORN, S.; SEITZ, S. M.; KEMELMACHER-SHLIZERMAN, I. Synthesizing obama: Learning lip sync from audio. *ACM Trans. Graph.*, Association for Computing Machinery, New York, NY, USA, v. 36, n. 4, jul 2017. ISSN 0730-0301. Available from Internet: <https://doi.org/10.1145/3072959.3073640>.

TERVEN, J.; CORDOVA-ESPARZA, D. M.; RAMIREZ-PEDRAZA, A.; CHAVEZ-URBIOLA, E. A. *Loss Functions and Metrics in Deep Learning*. 2023.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L. u.; POLOSUKHIN, I. Attention is all you need. In: GUYON, I.; LUXBURG, U. V.; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. v. 30. Available from Internet: <https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>.

VEAUX, C.; YAMAGISHI, J.; MACDONALD, K. *SUPERSEDED - CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit*. University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2016. Available from Internet: <https://datashare.ed.ac.uk/handle/10283/2119>.

VOLKSWAGEN. *VW 70 anos | Gerações | VW Brasil*. 2023. <https://www.youtube.com/watch?v=aMl54-kqphE>. Accessed: 2024-04-14.

WANG, T.-C.; LIU, M.-Y.; ZHU, J.-Y.; LIU, G.; TAO, A.; KAUTZ, J.; CATANZARO, B. Video-to-video synthesis. p. 1152–1164, 2018. Available from Internet: <https://dl.acm.org/doi/abs/10.5555/3326943.3327049>.

WANG, Y.; SONG, L.; WU, W.; QIAN, C.; HE, R.; LOY, C. C. Talking faces: Audio-to-video face generation. In: _____. *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*. Cham: Springer International Publishing, 2022. p. 163–188. ISBN 978-3-030-87664-7. Available from Internet: <https://doi.org/10.1007/978-3-030-87664-7_8>.

WANG, Z.; BOVIK, A.; SHEIKH, H.; SIMONCELLI, E. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, v. 13, n. 4, p. 600–612, 2004. Available from Internet: <https://doi.org/10.1109/TIP.2003.819861>.

YU, L.; XIE, H.; ZHANG, Y. Multimodal learning for temporally coherent talking face generation with articulator synergy. *IEEE Transactions on Multimedia*, v. 24, p. 2950–2962, 2022. Available from Internet: <https://doi.org/10.1109/TMM.2021.3091863>.

ZHANG, D.; YU, Y.; DONG, J.; LI, C.; SU, D.; CHU, C.; YU, D. *MM-LLMs: Recent Advances in MultiModal Large Language Models*. 2024. Available from Internet: <https://arxiv.org/abs/2401.13601>.

ZHANG, R.; ISOLA, P.; EFROS, A. A.; SHECHTMAN, E.; WANG, O. The unreasonable effectiveness of deep features as a perceptual metric. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [s.n.], 2018. p. 586–595. Available from Internet: <https://doi.org/10.48550/arXiv.2310.05986>.

ZHEN, R.; SONG, W.; HE, Q.; CAO, J.; SHI, L.; LUO, J. Human-computer interaction system: A survey of talking-head generation. *Electronics*, v. 12, n. 1, 2023. ISSN 2079-9292. Available from Internet: <https://www.mdpi.com/2079-9292/12/1/218>.

ZHENG, A.; ZHU, F.; ZHU, H.; LUO, M.; HE, R. Talking face generation via learning semantic and temporal synchronous landmarks. In: *2020 25th International Conference on Pattern Recognition (ICPR)*. [s.n.], 2021. p. 3682–3689. Available from Internet: <https://doi.org/10.1109/ICPR48806.2021.9412425>.

ZHONG, W.; FANG, C.; CAI, Y.; WEI, P.; ZHAO, G.; LIN, L.; LI, G. Identity-preserving talking face generation with landmark and appearance priors. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [s.n.], 2023. p. 9729–9738. Available from Internet: <10.48550/arXiv.2305.08293>.

ZHOU, Y.; HAN, X.; SHECHTMAN, E.; ECHEVARRIA, J.; KALOGERAKIS, E.; LI, D. Makelttalk: Speaker-aware talking-head animation. *ACM Trans. Graph.*, Association for Computing Machinery, New York, NY, USA, v. 39, n. 6, nov 2020. ISSN 0730-0301. Available from Internet: <https://doi.org/10.1145/3414685.3417774>.

ZHU, X.; LEI, Z.; LIU, X.; SHI, H.; LI, S. Z. Face alignment across large poses: A 3d solution. *CoRR*, abs/1511.07212, 2015. Available from Internet: <http://arxiv.org/abs/1511.07212>.