Animote: Audio Infused Emotion Enhancement in Animated Portraits

Mahabaleshwar Poorvita

Master of Science in Computer Vision ,Robotics and Machine Learning

from the

University of Surrey

School of Computer Science and Electrical and Electronic Engineering

Faculty of Engineering and Physical Sciences

University of Surrey

Guildford, Surrey, GU2 7XH, UK

September 2024

Supervised by: Dr Marco Volino

DECLARATION OF ORIGINALITY

I confirm that the project dissertation I am submitting is entirely my own work and that any material used from other sources has been clearly identified and properly acknowledged and referenced. In submitting this final version of my report to the JISC anti-plagiarism software resource, I confirm that my work does not contravene the university regulations on plagiarism as described in the Student Handbook. In so doing I also acknowledge that I may be held to account for any particular instances of uncited work detected by the JISC anti-plagiarism software, or as may be found by the project examiner or project organiser. I also understand that if an allegation of plagiarism is upheld via an Academic Misconduct Hearing, then I may forfeit any credit for this module or a more severe penalty may be agreed.

Animote: Audio Infused Emotion Enhancement in Animated Portraits

Author Name Mahabaleshwar Poorvita

Author Signature Mahabaleshwar Poorvita Date:02/09/2024

Supervisor's name: Dr Marco Volino

WORD COUNT

Number of Pages: 63

Number of Words: 18,630

ABSTRACT

The field of audio-driven face animation has garnered significant attention in recent years, given its potential applications in entertainment, virtual reality, and human-computer interaction. This report provides a comprehensive literature review tracing the evolution of methodologies and innovations in this domain, from early research efforts to the latest advancements. Emphasis is placed on state-of-the-art (SOTA) audio feature extraction techniques, with a particular focus on the wav2vec model, and the impact of integrating emotional cues with phonetic features to enhance the expressiveness of generated animations.

A novel dataset, comprising one hour of meticulously recorded and preprocessed audio, is introduced and utilized to evaluate the effectiveness of contemporary audio feature extraction methods. The report explores the potential benefits of incorporating emotional features extracted from audio, alongside traditional phonetic features, to produce more nuanced and expressive facial animations. This integration aims to address the limitations of current techniques which often fail to capture the subtleties of human emotion.

The report also delves into the promising avenue of merging traditional feature extraction methodologies with cutting-edge innovations such as diffusion models. This hybrid approach is posited as a means to significantly improve the quality and realism of animated videos. By providing an in-depth analysis of the existing challenges and proposing potential solutions, this work lays the groundwork for future research in the field.

Key contributions of this report include a detailed survey of past and current research, the introduction of a new dataset for benchmarking, and a proposed framework for enhancing expressiveness in face animation through combined audio features and advanced modeling techniques. These achievements underscore the significance of this work in pushing the boundaries of audio-driven face animation and offer a clear path for future explorations and improvements.

CONTENTS

LIST OF FIGURES

1 INTRODUCTION

Imagine animated characters that not only speak flawlessly but also convey emotions as vividly as real people. This vision is at the heart of audio-driven face animation, a field poised to revolutionize digital interactions, storytelling, and entertainment. The central challenge is creating animations that move and emote as naturally as humans. Traditional methods often produce stiff, simplistic animations. Our project tackles this by exploring how state-of-the-art audio feature extraction methods and the inclusion of emotional cues can enhance the realism and expressiveness of animated characters. We begin with a survey of the field's evolution, highlighting key innovations. Building on this, we introduce a novel dataset of one hour of meticulously recorded audio. This dataset serves as a benchmark for evaluating feature extraction methods, with a focus on the wav2vec model. We investigate the potential of combining phonetic features with emotional cues to create more nuanced animations.

To push the boundaries further, we propose merging traditional feature extraction with diffusion models, promising unprecedented realism and quality in animations. By addressing these challenges, our project aims to advance the field and provide a foundation for future research. Embark on this journey with us as we explore the intersection of phonetic precision and emotional expressiveness, leveraging the latest machine learning advancements to bring animated characters to life like never before.

1.1 Background and Context

The field of audio-driven face animation has evolved significantly over the past decades, drawing from advancements in speech recognition, computer vision, and machine learning. Initially, efforts in this area relied on basic lip-sync techniques that often resulted in animations lacking realism and expressiveness. As the demand for more immersive digital interactions grew, so did the need for sophisticated methods capable of capturing the nuances of human speech and emotion.

A pivotal moment in this field was the introduction of self-supervised learning techniques for speech representation, exemplified by Baevski et al.'s Wav2Vec 2.0 model [3]. This model demonstrated a robust framework for converting raw audio into meaningful latent representations, significantly improving the accuracy and quality of audio-driven applications. Coupled with trans-

former architectures [26], these advancements enabled the mapping of audio features directly to head pose embeddings, paving the way for real-time synchronized audio-visual representations.

Further contributions to the field include high-resolution audio-visual datasets, such as the one introduced by Zhang et al. [33], which have been instrumental in training models for one-shot talking face generation. The importance of datasets in this domain cannot be overstated, as they provide the diverse and high-quality data necessary for developing and testing new algorithms. For instance, the MSP-IMPROV dataset [7] has been crucial for studying emotion perception, which is essential for enhancing the emotional responsiveness of generated animations.

Recent advancements have also seen the integration of emotion recognition tools like openSMILE [10], which enable the extraction of emotional cues from audio, adding an emotional dimension to the phonetic features extracted from speech. The emergence of generative models, particularly diffusion models [21], has further revolutionized the field, allowing for the synthesis of high-resolution images and animations conditioned on audio inputs. The motivation for new work in this area stems from the quest for more lifelike and emotionally resonant digital characters. Our project proposes a novel approach that combines traditional feature extraction methods with the latest innovations in diffusion models [2], offering a pathway to more expressive and realistic audio-driven face animations.

Upcoming sections in this report will provide a comprehensive literature survey tracing the history of research in the field of audio-driven face animation. It will detail the methodology proposed, the results achieved, and the groundwork laid for possible future work. The survey will highlight key innovations and gaps in the current landscape, leading into our proposed approach and its potential impact on the field.

1.2 Objectives

”The project aims to advance the field of audio-driven facial animation by conducting a comprehensive literature review, preparing and utilizing a unique dataset, training advanced neural network models, and exploring large audiovisual datasets. These efforts are focused on enhancing the quality, expressiveness, and realism of generated facial animations. The specific objectives are as follows:”

- Conduct a comprehensive literature review on recent advancements, challenges, and methodologies in audio-driven facial animation generation.
- Prepare and incorporate a 1-hour video dataset of speaker's face pronouncing a phonetically balanced sentences to explore novel techniques for audio feature extraction using wav2vec, aiming to enhance quality of generated video.
- Train a neural network to generate mesh offsets from audio, incorporating emotion and contextual features using the Model for Dimensional Speech Emotion Recognition based on Wav2vec 2.0.
- Explore large audiovisual datasets to improve performance and generalization in audio-driven facial animation.
- Evaluate and innovate techniques for integrating audio features, facial landmarks, and motion modules to create realistic and expressive animations.

1.3 Achievements

This project presents a comprehensive survey of research papers and technologies, spanning from traditional feature extraction methods such as MFCC to state-of-the-art models for audio feature extraction. Key highlights include the recording and presentation of a 1-hour video dataset with a steady head pose, which has been preprocessed to extract audio and segment it at 30fps. A neural network model was trained on this dataset to produce 3D mesh offsets using audio input. Emotional features were extracted from the audio using a pretrained model on the MSP-IMPROV dataset, enhancing the expressiveness of the animation. The effectiveness of this approach was demonstrated through metrics such as R2 for lip synchronization, showing promising results. Additionally, 3D meshes were converted to 2D using orthogonal projection, proposed for use as target meshes in training a diffusion model for lifelike animation. This paper provides an intuition for integrating the extracted 2D target poses with state-of-the-art generative models to produce high-quality animations. Although full results have not yet been achieved, evidence and logic suggest that this method could be successful with a larger, well-processed dataset and sufficient computational power.

1.4 Overview of Dissertation

This dissertation is organized to provide a clear and structured narrative of the work undertaken, showcasing the progression from foundational research to practical implementation and evaluation. It begins with a thorough literature review that explores the evolution of audio-driven facial animation, covering methodologies from traditional feature extraction techniques like MFCC to state-of-the-art models such as Wav2Vec 2.0. This review examines the latest advancements and ongoing research, setting the stage for the technical contributions in subsequent chapters. Following this, the technical report details the project's methodology, including the neural network architecture, data acquisition and preprocessing, and the implementation process. This section provides the technical foundation necessary to understand the implementation and results discussed later.

The dissertation then presents the evaluation metrics and results of the implemented models, demonstrating the effectiveness of the approach through various measures like R2 for lip synchronization. It includes visual and quantitative results, showcasing the quality of the generated 3D mesh offsets and their conversion to 2D using orthogonal projection. The concluding chapter reflects on the project's achievements and limitations, offering a critical analysis of the results and outlining potential future work to enhance the methodology and integrate the extracted 2D target poses into state-of-the-art generative models for lifelike animations. The final bibliography section ensures thorough documentation of all sources cited throughout the dissertation, allowing readers to explore the referenced literature for additional context and depth. This structured narrative offers a logical progression from theoretical foundations to practical implementation and critical evaluation, ensuring that the reader can easily follow the development and significance of the research.

2 BACKGROUND THEORY AND LITERATURE REVIEW

The flow chart presented in Figure 3.9 vividly illustrates the evolution of audio-driven face animation, capturing key milestones and groundbreaking innovations. Beginning in 1974 with the creation of a parametric model for human faces [1], this foundational work set the stage for future advancements. The field took a significant leap forward in 1997 with Bregler et al.'s pioneering work in driving visual speech with audio [6], a concept that opened new possibilities for synchronizing auditory and visual elements. The early 2000s witnessed further progress with the development of trainable videorealistic speech animation techniques [11], which allowed for more adaptable and lifelike facial animations.

By 2010, the introduction of OpenSMILE marked a turning point, enabling the extraction of emotional nuances from speech and thus adding depth to digital avatars [10]. The innovative approach of dynamic units for visual speech in 2012 refined lip synchronization, making animations smoother and more believable [24]. In 2017, the integration of pose and emotion learning by Karras et al. brought a new level of expressiveness to facial animations, blending physical movements with emotional cues seamlessly [16]. These developments were complemented by the emergence of Wav2Vec 2.0 in 2020, which significantly enhanced the process of converting raw audio into meaningful representations, further elevating the quality of audio-driven animations [3].

Recent years have seen the advent of GAN-based models for facial animation in 2022 and NeRF-based reconstruction techniques in 2023, pushing the boundaries of realism and efficiency [30, 31]. The culmination of these efforts is reflected in the 2024 achievement of photorealistic portrait animation driven by audio [29], showcasing the pinnacle of current technology. Each milestone in this timeline underscores the remarkable progress made possible by technological advancements and a deeper understanding of audio-visual integration.

Following this, we will delve into a comprehensive literature review, exploring the significant contributions and methodologies that have shaped the current landscape and envisioning the future directions of audio-driven face animation.

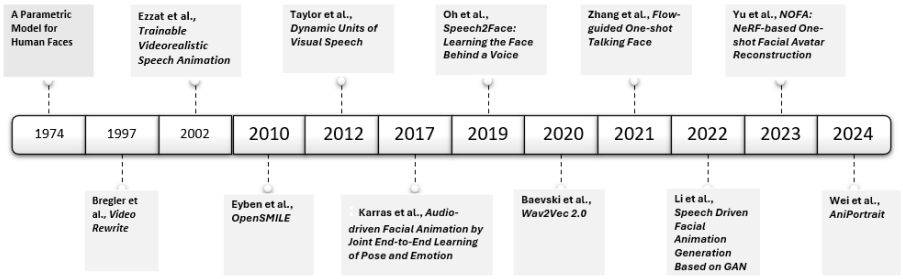


Figure 2.1: Timeline of Major Achievements in Audio-Driven Face Animation

2.1 Early methods in Audio driven face animation

Parametric models and rule-based approaches represent some of the earliest methods in the field of audio-driven facial animation. These techniques were foundational in exploring how audio input could be translated into realistic facial movements. Parametric models typically involve mathematical representations of facial features, allowing for controlled manipulation based on predefined parameters. Rule-based approaches, on the other hand, rely on a set of predefined rules derived from phonetic analysis to map audio features directly to facial motion parameters. These early methods, while innovative and pioneering, laid the groundwork for more advanced techniques by providing initial solutions to the complex problem of synchronizing facial animations with spoken audio. They highlighted both the potential and the limitations of using straightforward, manually defined systems to capture the nuances of human facial expressions in response to speech [1], [6], [8].

2.1.1 Parametric and Rule-Based Systems

One of the earliest methods, developed by Lewis in 1991, involved the use of predefined rules to map audio features directly to facial motion parameters. This approach was straightforward and intuitive, as it attempted to create a direct correlation between the sounds produced during speech and the corresponding facial movements. The rules were typically derived from phonetic features of the audio input, such as the positions of the lips, tongue, and jaw during the articulation of different phonemes. For example, Lewis's method might use a set of rules to determine that when a particular vowel sound is detected in the audio input, the system should generate a specific lip shape. Similarly, consonant sounds could be mapped to various positions and movements of the jaw and lips. While this method was innovative, it had significant limitations. The rule-based

approach struggled to capture the complexity and variability of natural facial expressions, often resulting in animations that appeared stiff and unnatural. Additionally, the reliance on manually defined rules meant that the system lacked the flexibility to adapt to different speakers or speaking styles.

In the field of parametric models, a significant contribution was made with the development of a parametric model for human faces [1]. This model laid the foundation for future rule-based systems by providing a structured way to represent facial features and movements. The parametric approach involved defining key facial points and using mathematical functions to simulate the movements of these points based on audio inputs. This method allowed for a more standardized way of creating facial animations, although it still faced challenges in achieving natural fluidity and expressiveness.

Another notable example of early rule-based systems is the work of Bregler, Covell, and Slaney [6], who developed the Video Rewrite system. This system drove visual speech by mapping audio features to facial movements using a set of predefined rules. The Video Rewrite system aimed to create realistic lip-sync animations by analyzing the phonetic content of the audio and generating corresponding facial movements. While it improved upon previous methods, it still faced difficulties in capturing the subtleties of natural facial expressions and adapting to different speaking styles.

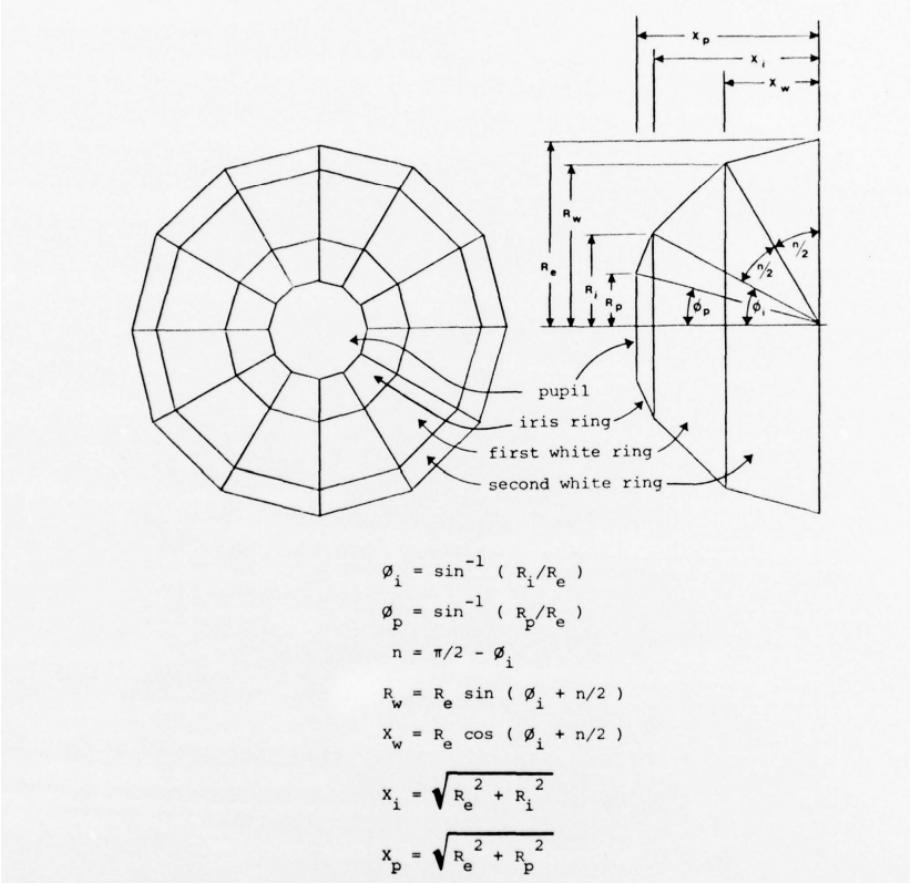


Figure 2.2: The construction of eyeball according to early parametric model[1]

2.1.2 Hidden Markov Models

Another early approach, introduced by Cohen and Massaro in 1993, utilized Hidden Markov Models (HMMs) to learn the mapping between audio and visual features. HMMs are statistical models that can represent the probabilistic relationships between observed data and hidden states. In the context of audio-driven facial animation, the observed data would be the audio features, while the hidden states would correspond to the underlying facial movements.

Cohen and Massaro's method involved training HMMs on a dataset of audio-visual recordings, where the audio features were extracted from the speech signal, and the visual features were derived from video recordings of the corresponding facial movements. The HMMs learned to associate specific audio patterns with particular facial movements, allowing the system to generate animations that were more closely aligned with the speech input. However, like the rule-based systems, HMMs also had their limitations. One major challenge was the difficulty in accurately modeling the complex dynamics of facial movements using a finite set of hidden states. This often

led to oversimplified animations that failed to capture the subtleties and expressiveness of natural speech. Additionally, the training process for HMMs required a large amount of labeled data, which was not always readily available.

In their work, Cohen and Massaro [8] focused on modeling coarticulation in synthetic visual speech. They used HMMs to create more fluid and natural transitions between phonemes, aiming to overcome some of the stiffness observed in rule-based systems. By learning from a dataset of audio-visual recordings, their approach could adapt to variations in speech, although it still faced challenges in capturing the full complexity of human facial movements.

These early methods of audio-driven facial animation synthesis, including rule-based systems and HMMs, were groundbreaking at the time but had inherent limitations. The reliance on pre-defined rules and finite state models meant that these systems struggled to adapt to the variability and expressiveness of natural human speech. Despite these challenges, the foundational work of researchers like Lewis, Bregler, Covell, Slaney, Cohen, and Massaro provided crucial insights and set the stage for the development of more advanced techniques capable of achieving more realistic and expressive facial animations.

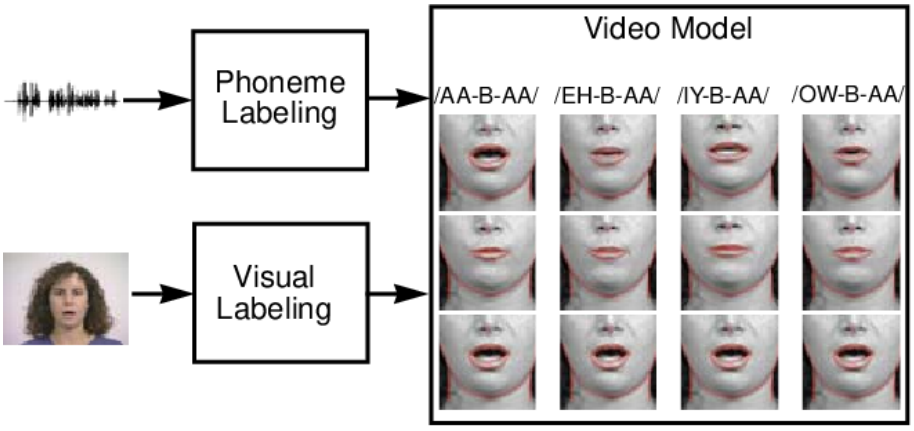


Figure 2.3: Overview of analysis stage. Video Rewrite[6] uses the audio track to segment the video into triphones. Vision techniques find the orientation of the head, and the shape and position of the mouth and chin in each image. In the synthesis stage, Video Rewrite selects from this video model to synchronize new lip videos to any given audio.

2.1.3 Enhanced Parametric and Neural Network Approaches

The seminal paper by Waters and Terzopoulos [28] leveraged Time-Delayed Neural Networks (TDNNs) to map audio signals to facial movements. TDNNs excel in capturing temporal patterns

in sequential data, making them ideal for processing speech. Their architecture, with each neuron connected to a local time window of neurons from the previous layer, allows the network to learn complex temporal relationships between phonetic inputs and facial expressions, enabling synchronized lip movements with audio. An illustrative example of this architecture is shown in figure 2.4.

Blanz and Vetter's work [4] introduced Multidimensional Morphable Models (MMM), which revolutionized facial animation by using statistical models to capture variations in facial shapes and textures from 3D scans. MMMs provide extensive control over facial expressions through the manipulation of shape and texture parameters, using Principal Component Analysis (PCA) to distill primary modes of variation in facial features. The flexibility of MMMs made them suitable for audio-driven animation.

Integrating TDNNs with MMMs offers a robust solution for synchronized facial animations from audio inputs. TDNNs process audio to predict visemes—the visual counterparts of phonemes. These visemes control the MMM, adjusting shape and texture components to generate facial expressions. Training TDNNs on datasets with synchronized audio and facial movement data ensures accurate mapping from sound to visual expression, producing animations that are both temporally and spatially coherent.

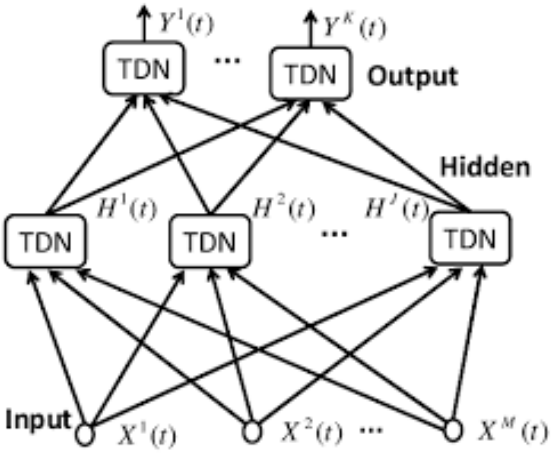


Figure 2.4: Architecture of a TDNN

This integration of TDNNs and MMMs represents a significant advancement in facial animation technology, combining the temporal resolution capabilities of TDNNs with the detailed spatial modeling of MMMs. This results in lifelike, synchronized facial animations, enhancing the real-

ism of digital avatars and opening new possibilities for applications in virtual reality, gaming, and digital communication.

2.2 Machine Learning-Based Approaches

The evolution of audio-driven facial animation has seen significant advancements with the introduction of machine learning-based approaches. These methods leverage sophisticated neural networks and large datasets to model the intricate relationships between audio features and facial movements. Key contributions in this domain include Wav2Vec 2.0, the MSP-IMPROV dataset, OpenSMILE, the Transformer architecture, and the end-to-end learning approach for audio-driven facial animation.

2.2.1 Self-Supervised Learning and Feature Extraction

Wav2Vec 2.0, introduced by Baevski et al. in 2020, is a framework for self-supervised learning of speech representations [3]. This model pre-trains on unlabeled audio data by masking portions of the input and predicting the missing parts, enabling it to learn rich, contextual representations of speech. After pre-training, the model is fine-tuned on labeled data for specific tasks. The ability of Wav2Vec 2.0 to capture nuanced speech features makes it highly effective for applications in speech recognition and audio-driven facial animation. However, the framework requires substantial computational resources and a large amount of unlabeled data for effective pre-training. Furthermore, the model's success in generating realistic facial animations is contingent upon the quality of the subsequent processing stages.

The MSP-IMPROV dataset, compiled by Busso et al. in 2017, is designed to study emotion perception in speech [7]. It includes audio-visual recordings of dyadic interactions featuring a range of emotional expressions, making it particularly useful for training models to recognize and replicate emotional nuances in facial animations. This enhances the expressiveness and realism of the animations. Nonetheless, the dataset's primary limitation is its focus on acted emotions, which may not fully capture the complexity of natural emotional expressions. Additionally, its size and diversity may be insufficient for training models that need to generalize across a broad spectrum of speakers and emotional states.

OpenSMILE, developed by Eyben et al. in 2010, is a versatile and fast audio feature extractor widely used in machine learning models [10]. It extracts a comprehensive range of audio features, including pitch, energy, and spectral properties, which are essential for modeling the relationship

between speech and facial movements. However, despite its effectiveness, OpenSMILE's reliance on predefined feature sets can limit its adaptability to new tasks or domains. Moreover, it does not inherently capture the temporal dynamics of speech, which are crucial for generating realistic facial animations.

2.2.2 Transformer Architectures and End-to-End Learning

The Transformer architecture, introduced by Vaswani et al. in 2017, revolutionized sequence modeling with its self-attention mechanism [26]. This architecture excels in handling long-range dependencies in data, making it particularly suitable for tasks like audio-driven facial animation where temporal coherence is vital. However, Transformers are computationally intensive and require large datasets for effective training. Their performance can be hampered by insufficient training data or computational resources. Additionally, while they excel at capturing dependencies, they may struggle with the finer nuances of facial expressions without specialized tuning.

In 2017, Karras et al. proposed an end-to-end learning framework for audio-driven facial animation that jointly models pose and emotion from audio inputs [16]. This model uses audio features to drive both the facial pose and emotional expression, aiming to produce more cohesive and expressive animations. The primary challenge with this approach is the need for large amounts of high-quality labeled data for effective training. Additionally, the joint modeling of pose and emotion is complex, requiring sophisticated architectures and significant computational resources.

In 2019, Oh et al. introduced Speech2Face, a model that learns to generate faces from voice input alone [20]. This model leverages the relationship between vocal characteristics and facial features, producing plausible face animations from audio data. However, the model's effectiveness is influenced by the diversity and quality of the training data, and it may not fully capture the emotional and expressive aspects of speech. Recent advancements also include AniPortrait by Wei et al. in 2024, which focuses on the audio-driven synthesis of photorealistic portrait animations [29]. This model employs advanced generative techniques to create lifelike animations, though it requires substantial computational resources and high-quality datasets to achieve optimal results.

Another significant contribution is NOFA by Yu et al. in 2023, which uses Neural Radiance Fields (NeRF) for one-shot facial avatar reconstruction [31]. This model excels in generating detailed 3D facial animations from a single image and audio input, though its performance heavily depends on the complexity of the scenes and the quality of the input data. Flow-guided One-shot

Talking Face Generation by Zhang et al. in 2021 introduces a method for generating talking face animations with high-resolution audio-visual datasets [33]. This approach leverages flow-guided mechanisms to ensure smooth and coherent animations, though it requires extensive computational resources for training and real-time application.

Our proposed method builds on the strengths and addresses the limitations of these approaches. We present a comprehensive survey of research papers and technologies, ranging from traditional feature extraction methods such as MFCC to state-of-the-art models for audio feature extraction. We recorded a 1-hour video dataset with a steady head pose, which was preprocessed to extract audio and segment it at 30fps. A neural network model was trained on this dataset to produce 3D mesh offsets using audio input. Emotional features were extracted from the audio using a pretrained model on the MSP-IMPROV dataset, enhancing the expressiveness of the animation [7]. The effectiveness of our approach was demonstrated through metrics such as R2 for lip synchronization, showing promising results. Additionally, 3D meshes were converted to 2D using orthogonal projection, proposed for use as target meshes in training a diffusion model for lifelike animation. Although full results have not yet been achieved, the evidence and logic suggest that this method could be successful with a larger, well-processed dataset and sufficient computational power.

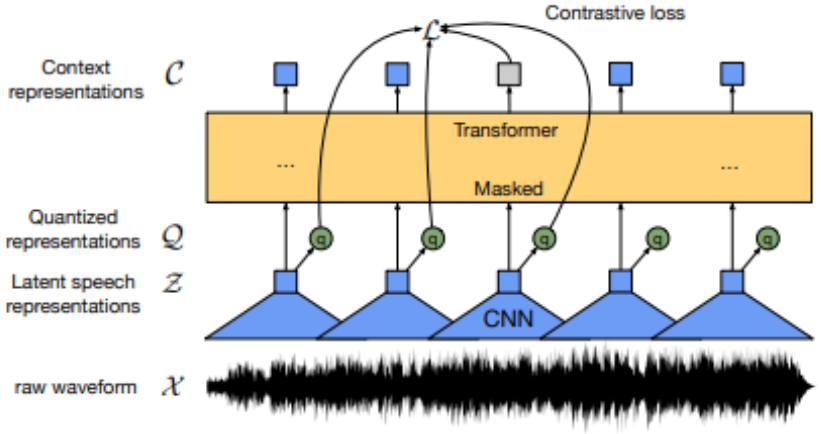


Figure 2.5: Illustration of framework by wave2vec 2.0[3] which jointly learns contextualized speech representations and an inventory of discretized speech units.

2.3 Hybrid and Novel Approaches

In the field of audio-driven facial animation, hybrid and novel approaches have emerged as powerful techniques, combining elements from various models and methodologies to push the bound-

aries of what is possible. These approaches leverage the strengths of traditional, deep learning, and generative models to achieve more realistic and expressive facial animations.

2.3.1 Diffusion Models

AnimateAnyone, introduced in 2023, utilizes pre-trained diffusion models to animate from a single image [2]. This method employs diffusion processes to progressively transform a static image into an animated sequence, capturing the subtle movements and expressions necessary for realistic facial animation. The primary advantage of this approach is its ability to generate animations from minimal input data. However, the reliance on pre-trained models means that the quality of the animation heavily depends on the diversity and representativeness of the training data.

High-Resolution Image Synthesis with Latent Diffusion Models, proposed in 2021, focuses on generating high-resolution images using latent diffusion processes [21]. This technique involves modeling the diffusion of latent variables through a deep neural network, allowing for the creation of detailed and high-fidelity images. While effective for static image generation, applying this method to dynamic facial animation poses challenges in maintaining temporal coherence and realism.

Diffused Heads, introduced in 2023, applies diffusion models specifically for talking-face generation [19]. This approach captures the dynamic aspects of facial movements by modeling the diffusion process across frames. Although promising, the computational complexity and the need for extensive training data remain significant barriers. In contrast, leveraging neural network models trained on specific datasets can sometimes offer more tailored solutions for temporal coherence.

Adding Conditional Control to Text-to-Image Diffusion Models, proposed in 2023, enhances diffusion models by introducing conditional controls for image synthesis [32]. This method allows for more precise manipulation of the generated images based on specific input conditions. While this adds flexibility and control, the approach requires robust mechanisms to integrate audio-driven cues seamlessly.

2.3.2 GAN-Based Models

The use of Generative Adversarial Networks (GANs) for video-realistic speech animation was pioneered in 2002 by Ezzat et al., with their Trainable Videorealistic Speech Animation [11]. This approach leverages GANs to generate realistic lip movements synchronized with speech. GANs

are effective in producing high-quality animations, but they often require extensive labeled data and careful tuning to avoid issues such as mode collapse and artifacts in the generated images. Comparatively, diffusion models can offer stability in generating high-fidelity details without some of the common pitfalls of GANs.

In 2012, Taylor et al. introduced Dynamic Units of Visual Speech, which uses GAN-based dynamic units for visual speech synthesis [24]. This method decomposes speech into dynamic units that are then synthesized using GANs to produce smooth and coherent animations. Despite its effectiveness, this approach can struggle with capturing the full range of emotional expressions and may require significant computational resources for training.

2.3.3 Hybrid Models

Speech2Face, introduced by Oh et al. in 2019, combines deep learning with rule-based techniques to generate faces from audio input [20]. This hybrid approach leverages the strengths of both methodologies to produce plausible facial animations. The model utilizes deep learning to extract features from audio input, capturing essential speech characteristics such as pitch, timbre, and rhythm. These features are then mapped to facial attributes using a rule-based system that ensures the generated faces adhere to basic phonetic rules and visual coherence. However, the model's performance is heavily influenced by the diversity and quality of the training data, and it may not fully capture the emotional and expressive aspects of speech. This limitation is often addressed in approaches that integrate emotional features extracted from audio, enhancing the expressiveness of the animation. For instance, incorporating emotional cues from datasets like MSP-IMPROV can significantly improve the emotional realism of the generated animations.

The paper "NOFA: NeRF-based One-shot Facial Avatar Reconstruction" by Yu et al. [31] presents a novel approach to facial animation by leveraging Neural Radiance Fields (NeRFs). The primary innovation lies in integrating NeRFs for volumetric rendering with a one-shot learning mechanism, allowing the model to create a detailed and accurate 3D facial avatar from a single image input. The core of this approach is illustrated in figure 2.6 taken from [31], which demonstrates the deformation field mechanism consisting of the Deformation Network (D-Net) and the Weighting Network (W-Net). The D-Net computes coordinate offsets from the deformed space to the canonical space, mapping facial features' movements to match different expressions and speech. The W-Net predicts the weights for these offsets, ensuring the significance of each defor-

mation in the final output. This combination generates weighted offsets that define the deformation field, enabling volumetric consistency and high-detail animations.

While the NOFA approach [31] offers significant advancements, such as one-shot learning for accessible avatar creation and maintaining high volumetric detail, it also presents challenges. The reliance on NeRFs may introduce computational overhead, potentially affecting real-time application scalability. Additionally, the current focus on geometric deformation could benefit from incorporating a more comprehensive emotional representation model. Evaluating the model's effectiveness across diverse facial structures and skin tones is crucial to ensure consistent performance and avoid biases. Despite these challenges, the NOFA method [31] marks a significant technical contribution to audio-driven facial animation, promising exciting developments in virtual reality, gaming, and beyond.

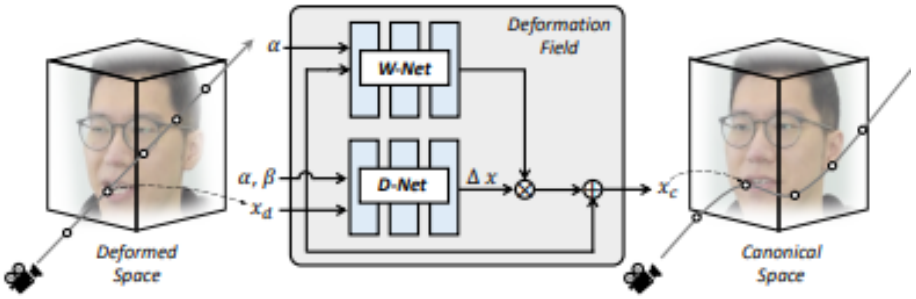


Figure 2.6: Illustration of the deformation field. It consists of a deformation network (D-Net) and a weighting network (W-Net). D-Net regress the coordinates offsets from the deformed space to the canonical space. W-Net predicts the per-location weight scalars to multiply with the offsets. With the weighted offsets, we can query volume features defined in the canonical space for volumetric rendering

Flow-guided One-shot Talking Face Generation by Zhang et al. in 2021 introduces a flow-guided approach that incorporates traditional and deep learning techniques to generate talking face animations with high-resolution audio-visual datasets [33]. This method leverages optical flow to capture the motion dynamics between consecutive frames, providing a more accurate and coherent representation of facial movements. By guiding the generation process with flow information, the model ensures that the animations are smooth and natural. The integration of traditional optical flow techniques with modern deep learning models allows for the preservation of fine details and temporal consistency. However, it requires extensive computational resources for training and real-time application, a challenge also faced by approaches integrating high-dimensional data pro-

cessing. The high computational demand is due to the need for real-time optical flow computation and the subsequent synthesis of facial animations.

AniPortrait, introduced by Wei et al. in 2024, combines various audio-driven synthesis techniques for photorealistic portrait animation [29]. This method uses advanced generative techniques to create lifelike animations. AniPortrait employs a combination of generative adversarial networks (GANs) and diffusion models to achieve high fidelity and photorealism. The GAN component is responsible for generating realistic textures and details, while the diffusion model ensures smooth transitions and temporal coherence. Though it demands substantial computational resources and high-quality datasets to achieve optimal results, the hybrid nature of AniPortrait allows it to leverage the strengths of both models. Similar to our approach, which combines different models for feature extraction and synthesis, AniPortrait exemplifies the potential of hybrid methodologies to enhance realism in facial animations. Controlled models like UNet-based and fused models also play a significant role in hybrid approaches. These models integrate different feature extraction and generative techniques to produce more controlled and precise outputs. For instance, UNet-based architectures are known for their effectiveness in image-to-image translation tasks due to their ability to capture both local and global features through skip connections. Fused models combine the strengths of multiple architectures to handle the complexity and variability of facial animations more effectively, ensuring that the generated animations are both detailed and dynamically accurate.

2.4 Comprehensive Review of Datasets for Audio-Driven Facial Animation

In the domain of audio-driven facial animation, datasets play a pivotal role in training and evaluating models to ensure they produce realistic and expressive animations. These datasets provide the necessary data to teach models how to correlate audio input with facial movements, expressions, and emotions. Here, we review several key datasets and their contributions to the field.

The MSP-IMPROV dataset, introduced by Busso et al. in 2017, is a well-regarded resource designed to study emotion perception in speech [7]. It contains audio-visual recordings of dyadic interactions, capturing a wide range of emotional expressions. This dataset is particularly valuable for training models to recognize and replicate emotional nuances in facial animations. The comprehensive nature of MSP-IMPROV, with its rich emotional context, makes it indispensable for projects aiming to enhance the expressiveness of animations. Its main strength lies in its ability to

provide varied emotional data, which is crucial for creating animations that can accurately convey different emotions.

The VFHQ dataset, presented by Xie et al. in 2022, is a high-quality dataset designed for video face super-resolution [30]. It offers high-fidelity video faces, which are essential for training models that require detailed and high-resolution facial features. VFHQ is instrumental in enhancing the quality and realism of facial animations by providing high-resolution data that helps models generate more precise and lifelike animations. The dataset's ability to support super-resolution tasks ensures that the facial animations produced are not only realistic but also visually appealing.

CelebV-HQ, introduced by Zhu et al. in 2022, is a large-scale video facial attributes dataset used for training and benchmarking [34]. This dataset contains extensive video facial attribute data, which is crucial for developing robust facial animation models. CelebV-HQ's large scale and diversity make it an excellent resource for training models that need to generalize across a wide range of facial attributes and expressions. Its extensive nature ensures that the models trained on it can handle a variety of facial features, enhancing their robustness and versatility.

LRW (Lip Reading in the Wild) is a widely used dataset for lip-reading tasks, containing over 500,000 spoken sentences from various speakers [14]. This dataset is invaluable for training models to understand and animate lip movements accurately. The diversity of speakers and the extensive number of sentences make LRW a robust dataset for improving lip synchronization in animations.

LRS-2 (Lip Reading Sentences 2) is another large-scale dataset for lip-reading, providing thousands of spoken sentences [23]. It offers a diverse range of speakers and speaking styles, which enhances the robustness of lip synchronization models. LRS-2's extensive dataset ensures that models can generalize well to different speakers and speaking conditions, making it an essential resource for lip-reading tasks.

MV-LRS (Multiview Lip Reading Sentences) includes multi-view recordings of lip-reading sentences, allowing for training models that can handle different viewing angles [13]. This capability is crucial for improving the accuracy of lip-reading animations, as it ensures that the models can generate accurate lip movements from various perspectives.

VoxCeleb2 is a large-scale dataset containing over a million utterances from thousands of speakers [15]. It is widely used for speaker recognition and audio-visual speech recognition tasks. The dataset's extensive nature and diversity make it a valuable resource for training models to recognize and animate a wide range of speaker voices and expressions.

HDTF (High-Definition Talking Faces) provides high-definition video recordings of talking faces, enabling the development of high-quality talking head generation models [33]. The high resolution of the videos in HDTF ensures that the generated animations are detailed and realistic, making it a critical dataset for high-quality facial animation tasks.

GRID is a multi-speaker audio-visual sentence dataset containing high-quality video and audio recordings of 34 speakers uttering a series of command-like sentences [18]. It is commonly used for lip-reading and audio-visual speech recognition tasks. The dataset's multi-speaker nature and command-like sentences make it ideal for training models to handle diverse speech patterns and speaker variations.

TCD-TIMIT is an audio-visual dataset derived from the TIMIT audio corpus, featuring video recordings of volunteers speaking sentences [12]. This dataset includes both frontal and profile views of speakers, providing valuable data for developing models that can handle different perspectives. TCD-TIMIT's inclusion of multiple views makes it a versatile dataset for training models to generate accurate facial animations from various angles.

Choosing the right sentence libraries is also crucial for developing effective audio-driven facial animation models. Harvard Sentences, a list of phonetically balanced sentences commonly used for testing audio equipment and speech synthesis systems, cover a wide range of phonetic contexts, making them useful for training and evaluating speech-related models [22]. The CMU Arctic Database is designed for speech synthesis research, containing high-quality recordings of sentences read by professional speakers [9]. LibriSpeech is a large-scale corpus of read English speech, commonly used for training and benchmarking speech recognition systems. It includes thousands of hours of speech from audiobooks, providing a diverse and extensive dataset for audio-driven animation models [25]. Choosing the correct sentences is crucial because it ensures the phonetic coverage and variability needed to train robust models. The selection of sentences can significantly impact the model's ability to generalize across different speech patterns and accents, thus improving the overall quality and realism of the generated animations.

These datasets collectively form the backbone of modern audio-driven facial animation research. They provide diverse, high-quality data that allows models to learn complex correlations between audio input and facial movements, ensuring that the generated animations are realistic and expressive. Each dataset offers unique strengths, such as emotional diversity, high resolution, or extensive lip-reading data, which contribute to the robustness and versatility of the models trained on them.

However, the effectiveness of these datasets also depends on their specific use cases. For example, while MSP-IMPROV excels in emotional expression, it may not fully capture the subtleties of natural emotions due to its acted nature [7]. Similarly, VFHQ and CelebV-HQ provide high-resolution visual data but might overlook audio synchronization nuances [30, 34]. On the other hand, datasets like LRW, LRS-2, and MV-LRS focus on lip synchronization but may require additional emotional context to enhance expressiveness [14, 23, 13]. Combining these diverse datasets can help mitigate individual limitations and create more comprehensive training data for robust and versatile models.

Our proposed project leverages this intuition by integrating traditional simple feature extraction methods or machine learning-based feature extraction with generative models to produce high-quality and realistic facial animations. By utilizing these datasets, we can develop models capable of handling various scenarios, ensuring that the animations are both detailed and expressive.

Table 2.1: Comparison of Datasets for Audio-Driven Facial Animation

Dataset Name	**Citation**	**Speakers**	**Duration (Hours)**	**Emotions**	**Phonetic Coverage**
MSP-IMPROV	[7]	12	9.4	Anger,Sadness, Happiness, Neutral	Comprehensive
CelebV-HQ	[34]	1000+	50+	Yes	Comprehensive
VFHQ	[30]	85	20	No	High
LRS2	[23]	1,083	224	No	High
HDTF	[33]	59	16	No	High
GRID	[18]	34	27	No	High
TCD-TIMIT	[12]	59	6.9	No	High

2.5 Summary

In this comprehensive literature review, we traverse the evolution of audio-driven facial animation, from foundational parametric models to advanced hybrid approaches. Beginning with the early efforts in parametric representation and rule-based modeling, such as the work on coarticulation and the Video Rewrite system, we see the groundwork laid for more sophisticated techniques. The advent of machine learning brought significant advancements, with frameworks for self-supervised learning and datasets for emotion perception enhancing model training. Transformer architectures revolutionized sequence modeling, leading to improved pose and emotion modeling in facial animations. Diffusion models further elevated the field, achieving remarkable realism and surpassing previous GAN-based methods in generating lifelike animations. Hybrid approaches, combining traditional techniques with deep learning innovations, pushed the boundaries even further. Essential datasets provided the rich, diverse data necessary for training and benchmarking these models. Overall, this review highlights the transformative impact of each technological leap, culminating in increasingly lifelike and expressive audio-driven facial animations.

3 METHODOLOGY

The proposed model architecture is designed to effectively translate audio inputs into high-fidelity facial animations by leveraging various neural network components. This approach captures the nuanced interplay between speech and facial movements, resulting in realistic and expressive animations. The model enhances emotional expressions by integrating different feature embeddings, focusing on accurately conveying complex facial muscle movements to ensure lifelike animations.

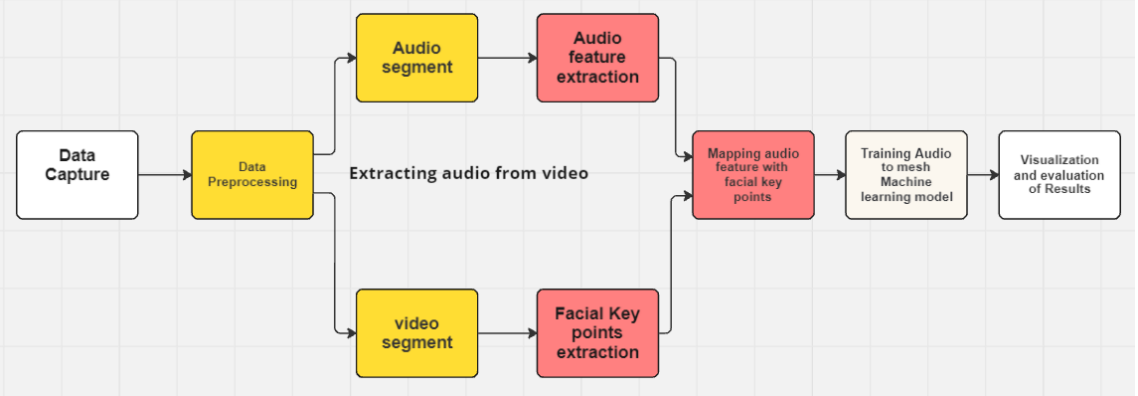


Figure 3.1: Overview of the proposed Methodology

The methodology for this project is a structured and systematic approach to develop an Audio-to-Mesh machine learning model, which aims to translate audio inputs into corresponding 3D facial key points. The process begins with Data Capture, where both video and audio data are collected. This data then undergoes Data Preprocessing to ensure synchronization and proper segmentation into audio and video components. As depicted in the methodology 3.1, the audio and video segments are subsequently processed through Audio Feature Extraction and Facial Key Points Extraction stages. These stages involve isolating the necessary features from the audio and identifying key points on the face from the video, which are essential for generating accurate facial meshes.

After feature extraction, the next step involves Mapping Audio Features with Facial Key Points. This mapping is crucial to maintain temporal coherence and accurately reflect the audio's impact on facial movements. The mapped data is then used in the Training Audio-to-Mesh Machine Learning Model phase, where a neural network is trained to predict the facial key points

based on the input audio features. Finally, the methodology concludes with Visualization and Evaluation of Results, where the model's performance is assessed, and the generated facial meshes are evaluated against the original data. This comprehensive approach ensures a well-rounded development process, capturing the intricate relationship between audio inputs and facial expressions, as illustrated in 3.1.

3.1 Architecture Overview

Initially, the audio input is processed through two instances of the Wave2Vec 2.0 model: one pre-trained on generic audio data to extract features for lip synchronization and another fine-tuned on the MSP-IMPROV dataset to capture facial expression features. The embeddings from these models are concatenated to form a comprehensive feature vector, which is then fed into the Audio2Mesh model. This model comprises a fusion module and a UNet decoder to generate 3D mesh offsets. Corresponding video segments of the audio inputs are processed using MediaPipe to extract facial keypoints, serving as ground truth data for training. The model learns to map audio to mesh offsets accurately, and the generated 3D mesh is combined with a neutral mesh and converted into a 2D mesh using orthogonal projection for final display.

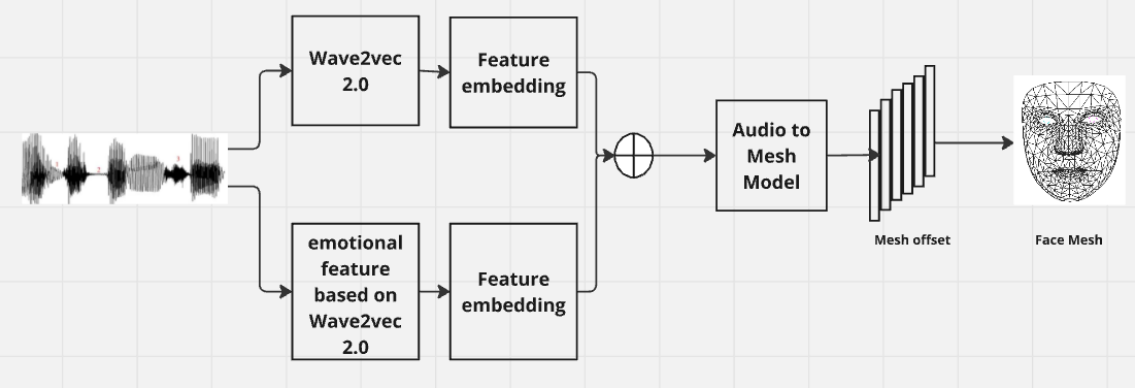


Figure 3.2: Overview of the proposed Model Architecture

- "Audio Feature Extraction Using Wave2Vec 2.0"

    1. **Lip Sync Features**: The first stage involves passing the audio input through the Wave2Vec 2.0 model, which is pre-trained and fine-tuned on 960 hours of Libri-Light and Librispeech on 16kHz sampled speech audio. This model was trained with a Self-Training objective. Our model is composed of a multi-layer convolutional feature encoder $f : X \rightarrow Z$ which takes as input raw audio X and outputs latent speech representations $z_1, \ldots, z_T$ for T time-steps. These representations are then fed to a

Transformer $g : Z \rightarrow C$ to build representations $c_1, \ldots, c_T$ capturing information from the entire sequence. The output of the feature encoder is discretized to q_t with a quantization module $Z \rightarrow Q$ to represent the targets in the self-supervised objective. Compared to vq-wav2vec, our model builds context representations over continuous speech representations and self-attention captures dependencies over the entire sequence of latent representations end-to-end. This detailed processing ensures that the generated facial animations correspond closely to the speech sounds.

	dev PER	test PER
CNN + TD-filterbanks [59]	15.6	18.0
PASE+ [47]	-	17.2
Li-GRU + fMLLR [46]	–	14.9
wav2vec [49]	12.9	14.7
vq-wav2vec [5]	9.6	11.6
This work (no LM)		
LARGE (LS-960)	7.4	8.3

Figure 3.3: An image from wave2vec 2.0 [3] showing TIMIT phoneme recognition accuracy in terms of phoneme error rate (PER). speech representations and an inventory of discretized speech units.

2. **Facial Expression Features**: At the same time, another instance of the Wave2Vec 2.0 model, takes raw audio as input and predicts arousal, dominance, and valence, ranging from 0 to 1. It also provides the pooled states of the last transformer layer. The model was fine-tuned on the MSP-Podcast (v1.7) dataset using the wav2vec2-large-robust model released by Facebook under Apache 2.0. Before fine-tuning, this model was pruned from 24 to 12 transformer layers. This setup helps the model generate facial animations that reflect the speaker's emotions and expressions while maintaining synchronization with the speech.

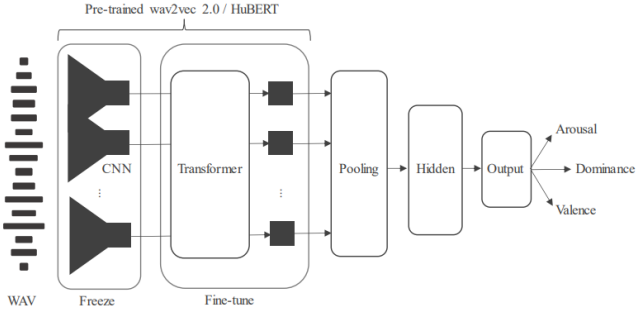


Figure 3.4: An image from [27] paper illustrating their proposed architecture built on wav2vec 2.0 / HuBERT

- ”Concatenation of Feature Embeddings” The embeddings obtained from both instances of Wave2Vec 2.0 are concatenated to form a unified feature vector. This concatenation integrates the detailed phonetic information necessary for lip sync with the expressive features needed for facial animations. By combining these features, the model can generate animations that are both temporally precise and emotionally expressive, addressing both key aspects of realistic facial animation.

- ”Audio2Mesh Model”

    1. **Fusion Module**:The concatenated audio features are then fed into a fusion module. This module integrates the multi-dimensional embeddings into a cohesive representation suitable for mesh generation. The fusion module typically consists of fully connected layers that refine and transform the combined features, ensuring that they are effectively prepared for the next stage of the process.
    2. **UNet Decoder**: Following the fusion module, the refined features are passed through a UNet decoder. The UNet architecture, known for its encoder-decoder structure with skip connections, is adapted here for generating 3D mesh offsets. The skip connections in the UNet help retain fine-grained details and maintain spatial coherence in the generated meshes, ensuring that the facial animations are both detailed and realistic.

- ”Ground Truth Data Extraction Using MediaPipe” To train the Audio2Mesh model, the corresponding video segments of the audio inputs are processed using MediaPipe. MediaPipe is a framework developed by Google that provides robust and accurate facial keypoint extraction. These keypoints serve as the ground truth data for training the model. By using MediaPipe, the model is trained with precise and reliable facial landmark information, which is essential for learning the correct mapping from audio inputs to 3D mesh outputs.

- ”Training the Audio2Mesh Model” The training process involves optimizing the Audio2Mesh model to minimize the error between the predicted mesh offsets and the ground truth keypoints extracted by MediaPipe. This optimization ensures that the model learns to generate 3D facial meshes that accurately correspond to the audio inputs. The training process focuses on both lip synchronization and expressive facial movements, resulting in high-fidelity animations.

- ”3D to 2D Mesh Conversion” The final stage involves converting the 3D mesh offsets generated by the model into a 2D mesh suitable for display. The generated 3D mesh offsets

are combined with a neutral mesh obtained from a reference image. This combined mesh is then projected into 2D using orthogonal projection. Orthogonal projection is chosen because it preserves the spatial structure and depth information of the 3D mesh while making it suitable for 2D display applications. This ensures that the final output maintains a realistic and visually coherent appearance, suitable for various applications such as virtual avatars and video conferencing.

3.1.1 Reasoning for Choosing the Architecture

The choice to use Wave2Vec 2.0 for feature extraction in our architecture is grounded in its demonstrated ability to capture rich and detailed audio features through a self-supervised learning approach. This capability is critical for extracting both phonetic details and emotional cues from audio inputs. Research has shown that Wave2Vec 2.0 outperforms traditional feature extraction methods in various speech-related tasks, including emotion recognition and speech recognition, due to its ability to learn from vast amounts of unlabeled audio data and fine-tune on specific tasks[3][27]. By employing two instances of Wave2Vec 2.0—one pre-trained on a broad dataset and another fine-tuned on the MSP-podcast dataset—we ensure a comprehensive representation that enhances both lip synchronization and the expressiveness of the generated facial animations.

The decision to concatenate these feature embeddings allows the model to leverage the strengths of both phonetic and emotional features, creating a robust input for the Audio2Mesh model. The UNet architecture is selected for its proven effectiveness in maintaining spatial details through its encoder-decoder structure with skip connections, which is ideal for generating detailed 3D meshes from complex feature embeddings. MediaPipe's facial keypoint extraction provides accurate and reliable ground truth data, which is essential for training the model to map audio inputs to facial mesh outputs accurately. Finally, the use of orthogonal projection for converting the 3D mesh to a 2D format ensures that the final output maintains depth and spatial structure, crucial for visual fidelity in 2D displays[5].

3.1.2 Potential Challenges and Limitations

While the architecture is robust and built on cutting-edge research, there are potential challenges and limitations. One significant challenge is the variability in audio quality and recording conditions, which can affect the performance of Wave2Vec 2.0 models. Inconsistent audio inputs may lead to inaccuracies in feature extraction, impacting the quality of the generated animations. Ad-

ditionally, the complexity of the UNet architecture requires substantial computational resources, which might limit its scalability and real-time application. Another potential hindrance is the alignment of audio features with facial keypoints; any mismatch could degrade the animation quality. Moreover, while orthogonal projection preserves spatial structure, it may not capture all nuances of 3D facial dynamics when viewed in 2D, potentially resulting in less realistic animations.

Addressing these challenges requires continuous refinement of the models and robust preprocessing techniques to ensure consistent and high-quality inputs. Optimizing the architecture for better performance and scalability, as well as exploring alternative methods for projection and feature alignment, will be critical for overcoming these limitations and achieving the desired outcomes.

3.2 Data Capture & Preprocessing

For this project, a one-hour video recording was made of an individual with a steady head pose and consistent lighting conditions. The individual uttered sentences from the Harvard Sentences collection and CMU Arctic Sentence Prompt, which are both designed for speech processing research purposes. These recordings were then segmented and processed to train an audio-to-mesh model aimed at generating high-fidelity facial animations from audio inputs.

- ”Harvard Sentences”

  The Harvard Sentences are a collection of phonetically balanced sentences developed at the Harvard University Psycho-Acoustic Laboratory for testing various aspects of speech and hearing. Each list contains ten sentences, carefully crafted to cover a wide range of phonetic contexts, making them ideal for evaluating speech intelligibility and quality in speech processing systems.

  Examples of Harvard Sentences:

  - ”The birch canoe slid on the smooth planks.”
  - ”Glue the sheet to the dark blue background.”
  - ”It’s easy to tell the depth of a well.”
  - ”A large size in stockings is hard to sell.”
  - ”The boy was there when the sun rose.”

- ”CMU Arctic Sentences”

  The CMU Arctic databases were created to support research in speech synthesis, particularly unit selection synthesis. These databases contain approximately 1,150 utterances from each of several speakers. The sentences are designed to be phonetically balanced, covering a wide range of phonetic contexts and providing a robust dataset for training and evaluating speech synthesis models.

  Examples of CMU Arctic Sentences:

  - ”The boy was there when the sun rose.”
  - ”A rod is used to catch pink salmon.”
  - ”The source of the huge river is the clear spring.”

  - ”Kick the ball straight and follow through.”
  - ”Help the woman get back to her feet.”
  - ”Author of the danger trail, Philip Steels, etc.”
  - ”Not at this particular case, Tom, apologized Whittemore.”
  - ”For the twentieth time that evening the two men shook hands.”
  - ”Lord, but I’m glad to see you again, Phil.”

  Using phonetically balanced sentences from the Harvard and CMU Arctic collections ensures comprehensive coverage of different phonetic contexts, which is essential for training robust speech processing models. These sentences are widely recognized and used in the research community for evaluating and developing speech synthesis and recognition systems. Moreover, these sentences encapsulate a variety of emotional expressions, from declarative statements to exclamations and questions, helping to elicit a wide range of facial muscle movements. This variety is crucial for creating realistic facial animations that reflect both the phonetic and emotional content of the speech.

  By leveraging these high-quality datasets and advanced models, this project aims to push the boundaries of audio-driven facial animation, making significant contributions to fields such as virtual avatars, video conferencing, gaming, and animation. The inclusion of sentences with different emotional tones ensures that the resulting animations are not only phonetically accurate but also expressively rich, enhancing the realism and engagement of the animations. This holistic approach to sentence selection and model training positions the project to set new standards in the fidelity and expressiveness of synthesized facial movements.

- ”Examples of Expressive Sentences”

  - ”The birch canoe slid on the smooth planks.” (Neutral, Declarative)
  - ”Can you imagine the depth of a well?” (Neutral, Interrogative)
  - ”What a huge surprise this has been!” (Happy, Exclamatory)
  - ”God bless ’em, I hope I’ll go on seeing them forever!” (Happy, Exclamatory)
  - ”Will we ever forget it?” (Sad, Interrogative)

3.2.1 Audio Processing

The audio processing part of the provided script is essential for preparing the data needed to train an audio-to-mesh model. This process ensures that the audio data is clean, appropriately segmented, and rich in features, which is critical for generating high-fidelity facial animations. The process begins with the separation of audio from the video. The video, which was recorded with a steady head pose and consistent lighting conditions, is first processed to extract its audio track. This extraction is achieved using the **moviepy.editor** library, which isolates the audio from the video file and saves it as a separate audio file in .wav format.

- Detection and Removal of Noise The script begins by detecting claps or other high-frequency noises within the audio track. This step is crucial because such noises can interfere with the training process, leading to less accurate models. The **`torchaudio`** library is used to load the audio file into a waveform and sample rate. If the audio is in stereo format, it is converted to mono by averaging the channels, which simplifies the processing by reducing the data to a single channel. The **`scipy.signal.find_peaks`** function is then employed to detect peaks in the audio waveform that exceed a specified threshold, representing potential clap sounds. These peaks are converted to time indices to identify the exact moments when claps occur.

  Once claps are detected, they are removed from both the audio and video files to maintain clean data. For the audio, silent segments are created using the **`pydub`** library to replace the detected clap segments. This involves generating 0.2-second segments of silence and inserting them in place of the claps. The video is also cleaned to ensure synchronization with the modified audio. This is achieved using **`moviepy.editor`**, where video segments around the detected claps are removed, and the remaining parts are concatenated to form a continuous video clip without interruptions.

3.2.2 Wave2Vec2.0 Features

The Wave2Vec 2.0 model processed the audio input and produced feature embeddings with a shape of **torch.Size([1, 1503, 1024])**. This shape indicates that the model generated embeddings of size 1024 for each of the 1503 time steps, with the first dimension (1) representing the batch size, signifying that only one audio sample was processed. These embeddings capture detailed phonetic features and temporal patterns from the raw audio signal.

```
Wave2Vec2.0 Features:
 [[ 0.02247964 -0.12937984  0.17878239 ...  0.10077487 -0.04571248
   0.17854832]
 [ 0.02238623 -0.12938085  0.17871034 ...  0.10084237 -0.04578343
   0.17855024]
 [ 0.02234292 -0.12942898  0.17856409 ...  0.10087609 -0.04577798
   0.17855996]
```

Figure 3.5: the values in the image reflect the model's interpretation of the audio signal at different time steps in 1 audio segment sample, essential for tasks like speech recognition and lip synchronization.

3.2.3 Emotion Model Features

Similarly, the emotion recognition model processed the same audio input and produced feature embeddings with a shape of **torch.Size([1, 1503, 1024])**. This shape indicates that the model generated embeddings of size 1024 for each of the 1503 time steps, with the first dimension (1) representing the batch size. These features capture emotional cues from the audio, such as arousal, valence, and dominance, which are crucial for generating expressive animations that align with the speaker's emotional state.

```
Emotion Model Features:
 [[-0.00731105  0.00970994 -0.00416438 ...  0.00721763  0.00930074
   0.00946413]
 [-0.00741599  0.00743567 -0.00120916 ...  0.00673495  0.00887502
   0.00568176]
 [-0.00731367  0.00754828 -0.00533749 ...  0.0072732   0.00812348
   0.0074364 ]
```

Figure 3.6: the values in the image reflect the model's interpretation of the emotional content in 1 sample audio segment.

3.2.4 Concatenated Features

The concatenated feature vector has a shape of **torch.Size([1, 1503, 2048])**, achieved by concatenating the feature embeddings from the Wave2Vec 2.0 model and the emotion recognition model along the last dimension. This horizontal concatenation, resulting in a combined size of 2048 for each of the 1503 time steps, ensures that the temporal alignment of features from both models is maintained.

```
Concatenated Features:
 [[ 0.02247964 -0.12937984  0.17878239 ...  0.00721763  0.00930074
   0.00946413]
 [ 0.02238623 -0.12938085  0.17871034 ...  0.00673495  0.00887502
   0.00568176]
 [ 0.02234292 -0.12942898  0.17856409 ...  0.0072732   0.00812348
   0.0074364 ]
```

Figure 3.7: the values in the image reflects concatenated feature embeddings of 1 sample audio segment.

Concatenating the features horizontally is beneficial as it maintains the temporal alignment of the phonetic and emotional information from both models. This alignment is critical for applications like speech-driven facial animation, ensuring that both the content and emotional tone of the speech are accurately reflected in the visuals. By integrating these features, the system can produce more realistic and expressive animations, improving user experience in virtual assistants, animated films, and video games. This approach leverages the strengths of both the Wave2Vec 2.0 model and the fine-tuned emotion recognition model, providing a comprehensive representation of the audio input that enhances accuracy and realism in various multimedia contexts.

3.2.5 Video Processing

Video processing is crucial for preparing data for tasks like emotion recognition and facial animation. It involves trimming unnecessary segments to reduce data size, detecting and removing claps to eliminate noise artifacts, splitting videos into manageable chunks for efficient processing, and generating 3D facial meshes to capture detailed facial expressions. These steps ensure the video data is clean, synchronized, and well-prepared for feature extraction. Proper video processing enhances the quality and reliability of the data, enabling models to learn and interpret temporal relationships effectively, thereby improving the performance of machine learning applications in producing accurate and expressive outcomes.

- Detecting and removing noise Claps or high-frequency noises in the audio track can interfere with the training process, leading to inaccurate models. The `detect_claps` function identifies these claps by analyzing the audio waveform and detecting peaks that exceed a specified threshold. Once detected, these segments are marked for removal. The `remove_claps_from_audio_and_video` function then utilizes these time indices to replace the claps in the audio with silent segments and removes the corresponding video

segments to maintain synchronization. This step is crucial to ensure that the training data remains clean, free of sudden noise artifacts, which could degrade model performance.

- Splitting the Video into Chunks

  The function `split_video` divides the input video into 30-second chunks, storing them in a specified directory. This segmentation is essential for handling long videos, as it breaks them down into manageable pieces for individual processing. Each chunk represents a smaller segment of the video, which can be processed in parallel or sequentially, thereby optimizing computational resources and making the feature extraction process more efficient. This chunking process ensures that each segment can be thoroughly analyzed without overwhelming the processing system.

- Generating 3D Meshes from Video Frames

  MediaPipe's[17] face mesh solution detects 468 facial landmarks, offering a high-resolution mesh that captures detailed facial expressions, making it ideal for tasks like emotion recognition and facial animation. The framework is optimized for real-time performance, crucial for processing video streams with low latency, and supports multiple platforms including Android, iOS, and web, allowing for broad deployment.

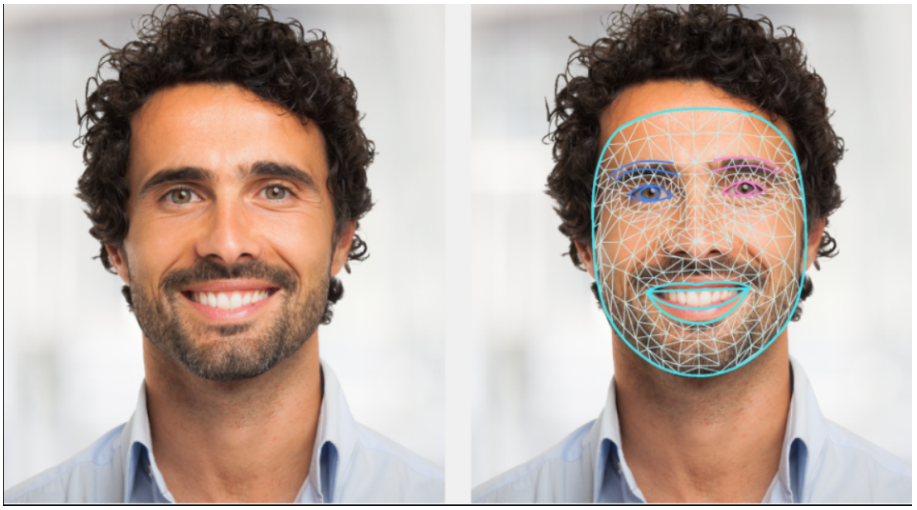


Figure 3.8: The face landmark detector in MediaPipe identifying facial expressions and landmarks.

In the code, the **`generate_3d_meshes_for_chunk`** function processes each video chunk frame by frame to extract these facial landmarks. Initially, a face detection model identifies the presence of faces in the video frames, reducing the search space for the landmark detection. The subsequent regression model predicts the precise coordinates of the 468 land-

marks, trained on a large dataset of annotated images to ensure robustness across various conditions. Post-processing steps then refine these landmarks to improve spatial consistency and smoothness, enhancing the visual quality of the detected mesh.

This detailed facial mapping is stored in **`video_meshes.npy`** files within **`output_chunk`** folders, alongside audio features extracted from the video chunks. These 3D meshes, visualized using the **`visualize_and_save_mesh`** function, provide a visual representation of the facial landmarks, ensuring accuracy and facilitating further analysis.

Compared to other solutions like OpenCV and Dlib, MediaPipe offers superior accuracy with its detailed 468 landmarks versus the typical 68 landmarks provided by these alternatives. Its real-time processing capabilities and ease of use through simple APIs make it an ideal choice for integrating into the data preprocessing pipeline.

Figure 3.9: The image shows a face with numerous facial landmarks highlighted by green dots. These landmarks represent key points detected by MediaPipe's face mesh solution. Each green dot represents a (x, y, z) coordinate corresponding to specific facial features such as the eyes, nose, mouth, and facial contours.

- Handling Frame Count and Sequence Alignment

  The extracted frames are padded or truncated to ensure a consistent sequence length using the **`pad_or_truncate_sequence`** function. This step ensures that all video chunks have the same number of frames, necessary for batch processing and alignment with corresponding audio features. Consistent frame count is crucial for synchronizing video and audio data, enabling the model to learn accurate temporal relationships between the two modal-

ities. This alignment is critical for applications that require comprehensive understanding and synchronization, such as speech-driven facial animation.

- Adding the Neutral Mesh To enhance the quality of the 3D meshes, the code also includes a function to generate a neutral mesh from a reference image. This neutral mesh serves as a baseline for applying mesh offsets derived from audio features. The **`get_neutral_mesh`** function processes a reference image to extract facial landmarks, ensuring the resulting mesh has the correct size and structure. This neutral mesh is then visualized for verification purposes.

- Saving and Visualizing Mesh Data

  Once the 3D meshes are generated, they are saved using the **`save_mesh_data`** function. Additionally, these meshes can be visualized and saved for further analysis using the **`visualize_and_save_mesh`** function. These steps help create a visual representation of the 3D facial landmarks, which is useful for verifying the accuracy of detection and ensuring that the data is correctly processed. Visualizing the data provides an additional layer of verification, ensuring that the processed data aligns with expected outcomes.

4 RESULTS AND EVALUATION

The Results and Evaluation section rigorously assesses the AudioConditionedNetwork's ability to transform audio inputs into precise 3D facial meshes. Supported by a detailed flowchart 4.2 and visualization in Figure 5.3, this evaluation highlights the model's performance, identifies key limitations, and suggests potential areas for improvement based on ablation studies and evaluation metrics.

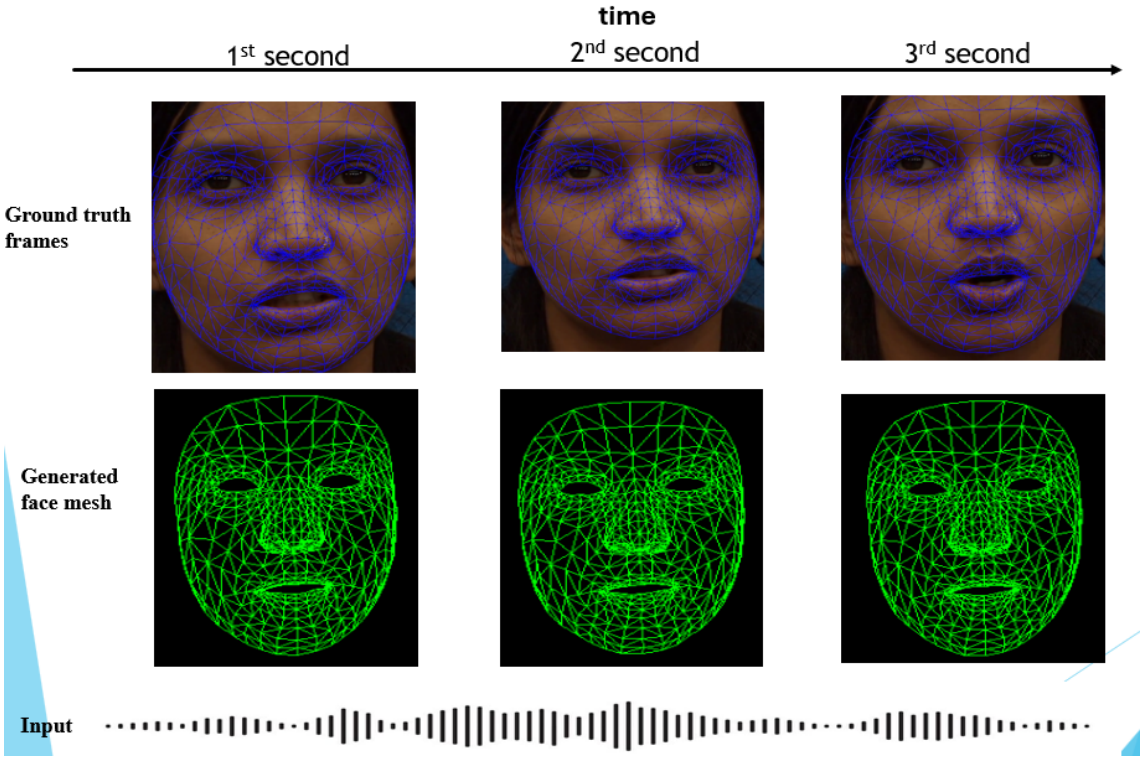


Figure 4.1: Comparison of ground truth facial mesh frames and generated face mesh over time, with corresponding audio input.

4.1 Implementation and Strategy

The implementation of this project focuses on the development of an Audio-to-Mesh machine learning model that translates audio features into corresponding 3D facial key points. The process begins with data capture, where both video and audio data are acquired, setting the foundation for the entire workflow. This raw data then undergoes Data Preprocessing, a crucial step that involves splitting the data into synchronized audio and video segments. Notably, the preprocessing of just one hour of captured video took approximately 20 hours, even when utilizing an A100 GPU in

Google Colab, highlighting the computational intensity of this phase. Following preprocessing, the Audio Feature Extraction phase isolates and processes audio segments to extract meaningful features that serve as input to the model. Concurrently, facial key points extraction from video segments captures the dynamic expressions of facial movements, which are vital for generating accurate facial meshes.

Once the extraction phases are complete, the next step, Mapping Audio features with facial key points, aligns the processed audio features with the corresponding facial key points. This mapping is essential to ensure temporal coherence and contextual relevance, allowing the model to accurately reflect audio-driven facial movements. Following the mapping, the project enters the core phase: Training the Audio-to-Mesh machine learning model. Here, the mapped data is used to train a neural network designed to predict facial key points based on the input audio features. The model was trained with a learning rate of 0.0001, and while the training process is relatively fast, completing 50 epochs in just 10 minutes using an L4 GPU, it's clear that the model's ability to capture finer details, such as lip movements and subtle facial expressions, could be improved. To achieve this, the loss value backpropagated during training should be increased, allowing for more hours of training and better focus on these intricate facial movements; however, this approach requires further experimentation and could be considered as future work. Finally, the project concludes with visualization and evaluation of Results, where the trained model's performance is assessed by comparing generated facial meshes with the original data. This stage includes a thorough analysis of the model's accuracy, temporal coherence, and effectiveness in capturing the subtle nuances of facial expressions driven by audio input. The flow chart in the figure 4.2 below provides an overall view of the data flow and the architecture employed in this project. Following this visual overview, detailed information on the different layers and components involved in the architecture is presented, offering insights into the implementation and design choices.

The implementation involves several key components, each of which contributes to the model's ability to accurately generate facial meshes from audio signals. Below, we explain each layer and module of the network, highlighting the rationale behind its design and how it integrates with the overall architecture, as illustrated in the flow chart.

4.1.1 Pretrained Audio Feature Extractor

The network begins with the Pretrained Audio Feature Extractor. This component leverages the ResNet-18 model, a well-established convolutional neural network (CNN) known for its robust

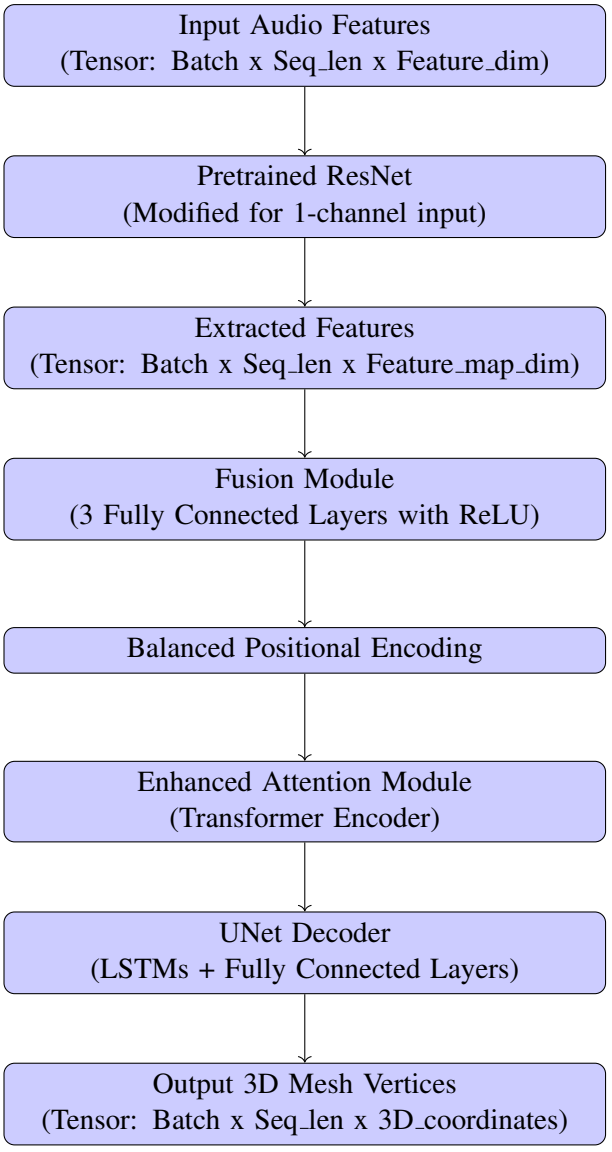


Figure 4.2: Flowchart of the AudioConditionedNetwork Architecture. This diagram shows the step-by-step data flow through the model, from input audio features to the final 3D mesh vertices output.

feature extraction capabilities. Originally designed for image processing, ResNet-18 is adapted here by modifying its first convolutional layer to accept a single-channel input, specifically designed to process audio spectrograms rather than RGB images. The final fully connected layer of ResNet-18 is removed to directly obtain the high-level features necessary for the next stages of the model. Using a pre-trained model like ResNet-18 significantly accelerates the training process and enhances the quality of the extracted features due to the model's prior exposure to a large, diverse dataset.

4.1.2 Balanced Positional Encoding

Following feature extraction, the network employs Balanced Positional Encoding. This layer introduces positional information into the sequence of extracted features, which is crucial for maintaining the temporal order of the audio frames as they progress through the model. The positional encoding is initialized in a way that prevents any inherent bias towards certain frames, ensuring that each frame contributes equally to the model's learning process. This approach is essential because it allows the model to effectively capture and understand the sequential nature of speech, thereby aiding in the synchronization of facial movements with the audio input.

4.1.3 Enhanced Attention Module

The features, now enriched with positional information, are passed through the Enhanced Attention Module. This module is composed of multiple Transformer Encoder layers, each designed to implement a self-attention mechanism. The self-attention mechanism allows the model to dynamically focus on different parts of the input sequence, making it capable of capturing long-range dependencies and contextual information. The attention mechanism is crucial for this task because it enables the model to understand which frames in the sequence are most relevant at any given time, ensuring that the generated facial movements are both realistic and well-timed. To enhance stability during training, Layer Normalization is applied after the Transformer Encoder, further refining the output of the attention layers.

4.1.4 Fusion Module

The next step in the architecture is the Fusion Module, which is a sequence of fully connected layers. The features processed by the attention module are passed through this module, which consists of three linear layers, each followed by a ReLU activation function. The purpose of the Fusion Module is to combine and transform the refined features into a more suitable format for decoding. This stage integrates information from various frames, enabling the network to produce a coherent representation of the input sequence that encapsulates both the temporal and contextual elements needed for accurate facial key point generation.

4.1.5 UNet Decoder

Finally, the processed features are decoded into the output format through the UNet Decoder. The decoder begins with a linear layer that adjusts the feature dimensions before passing them through two LSTM layers. These LSTM layers are crucial as they are specifically designed to handle

sequences, making them ideal for modeling the temporal dynamics inherent in facial expressions and speech. The output from the LSTM layers is then passed through another linear layer that maps the hidden states to the final 3D facial key points. This architecture ensures that the model can generate continuous and smooth facial movements that are temporally coherent with the audio input, accurately capturing the nuances of speech-driven facial dynamics.

The architecture chosen for this project is inspired by state-of-the-art approaches that have proven effective in related domains. Specifically, this architecture has been selected because it mirrors strategies employed in cutting-edge technologies and research, such as in the Wav2Lip model. Wav2Lip, which is detailed in the paper "Wav2Lip: Accurately Lip-syncing Videos In The Wild," demonstrates the efficacy of using a similar architecture for accurate lip-syncing and facial animation tasks. For more details, you can refer to the paper available on arXiv: `https://arxiv.org/abs/2008.10010`.

4.2 Metrics and Implementation

Metrics play a crucial role in evaluating the performance of machine learning models. They provide quantitative measures to assess how well a model performs in tasks such as prediction, classification, or reconstruction. In the context of generative models like the Audio-to-Mesh network described here, metrics are used to quantify the accuracy of the predicted outputs compared to the ground truth data. Specifically, reconstruction accuracy and temporal coherence are vital, as they ensure that the generated 3D meshes are not only accurate per frame but also consistent over time, which is essential for generating smooth, realistic animations.

Loss Functions

The training of the model involves minimizing specific loss functions, which are mathematical representations of the error between the predicted and actual outputs. The primary loss function used here is the **Mean Squared Error (MSE)**:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

This loss measures the average squared difference between the predicted mesh offsets and the ground truth offsets, encouraging the model to produce accurate 3D reconstructions. Additionally,

the **Temporal Coherence Loss**, defined as:

$$\text{L1 Loss} = \frac{1}{N} \sum_{i=1}^{N} |(y_{i+1} - y_i) - (\hat{y}_{i+1} - \hat{y}_i)|$$

is used to penalize abrupt changes between consecutive frames, ensuring that the generated mesh sequences are temporally consistent.

Weighted Bias

In the total loss function, the temporal coherence loss is weighted by a bias factor, typically set to 0.1, resulting in the following total loss:

$$\text{Total Loss} = \text{Reconstruction Loss} + 0.1 \times \text{Temporal Coherence Loss}$$

The purpose of this weighted bias is to balance the influence of the temporal coherence term relative to the reconstruction accuracy. Since accurate mesh reconstruction is generally more critical, it is given a higher weight, while the temporal coherence is added with a smaller bias to subtly guide the model towards producing smoother transitions without overwhelming the primary objective. This ensures that the model remains focused on accurately predicting each frame while still maintaining a reasonable degree of temporal smoothness across frames.

The results presented in the figure 4.3 offer a comprehensive look at how different hyper-parameters affect the performance of an audio-to-mesh model. Across the seven experiments, variations in learning rate, batch size, number of epochs, and attention heads were meticulously tested to evaluate their impact on the model's ability to predict 3D mesh representations from audio inputs. A critical analysis of these experiments reveals both promising trends and areas that warrant further exploration.

Experiment	Learning Rate	Batch Size	Epochs	Attention Heads	Training Loss	Validation Loss	Training R^2	Validation R^2	Test loss	Test R^2
1	0.0001	1	50	2	0.0063	**0.0053**	0.9580	0.9667	**0.0043**	0.9718
2	0.0001	1	100	2	**0.0058**	0.0073	0.9635	0.9516	**0.0057**	0.9619
3	0.005	1	50	2	0.0138	0.0057	0.8968	0.9694	0.0046	0.9765
4	0.0005	1	50	2	0.0108	0.0075	0.9124	0.9361	0.0074	0.9391
5	0.0001	4	50	2	0.0133	0.0056	0.9052	0.9655	0.0059	0.9646
6	0.0001	1	50	4	0.0063	0.0047	0.9596	0.9697	0.0047	0.9725
7	0.0001	1	50	8	0.0061	0.0041	0.9600	0.9778	0.9778	0.9820

Figure 4.3: Results of Hyperparameter Experiments on Audio-to-Mesh Model: Impact on Training, Validation, and Test Performance Metrics

- **Training Loss:** Measures how well the model is performing on the training data.
- **Validation Loss:** Assesses how well the model generalizes to unseen validation data.
- **Training R^2:** Indicates the proportion of variance in the training data explained by the model.
- **Validation R^2:** Reflects how well the model's predictions match the validation data.
- **Test Loss:** Evaluates the model's performance on completely unseen test data.
- **Test R^2:** Measures the accuracy of the model's predictions on the test dataset.

The learning rate, one of the most pivotal hyperparameters, shows clear evidence that lower values lead to more stable and effective training. For instance, the majority of the experiments (1, 2, 5, 6, and 7) used a learning rate of 0.0001 and consistently achieved strong performance metrics, particularly in terms of test R^2 values. Experiment 1, which serves as a baseline, demonstrates this with a solid test R^2 of 0.9718. On the other hand, Experiment 3, which experimented with a higher learning rate of 0.005, resulted in significantly higher training loss and lower training R^2, although the test R^2 remained relatively high. This suggests that while a higher learning rate might accelerate convergence, it can also lead to instability, potentially causing the model to overshoot optimal solutions. This underscores the importance of carefully selecting a learning rate that balances convergence speed and model accuracy.

Batch size is another critical factor, and its impact is nuanced. Experiments with a batch size of 1 generally produced strong results, indicating that for this task, smaller batch sizes may be more effective. For example, Experiment 1 achieved a training R^2 of 0.9580 and a test R^2 of 0.9718. In contrast, Experiment 5, which used a batch size of 4, did not significantly outperform the smaller batch sizes, suggesting that the model benefits from the more granular updates provided by smaller batches. This might be particularly important in capturing the subtle nuances in the audio-to-mesh mapping process, where the model needs to adapt to highly specific patterns within the data.

The number of epochs also played a significant role, but the benefits of increasing the epoch count appear to be limited. Experiment 2, which doubled the epochs to 100, showed slight improvements in training loss but did not lead to substantial gains in validation or test metrics compared to Experiment 1. This indicates that while longer training allows the model to better fit the training data, it does not necessarily enhance its generalization to new data. This could suggest that the model quickly reaches a point of diminishing returns where additional epochs contribute little to performance improvements, possibly due to the risk of overfitting.

The experiments also explored the effect of varying the number of attention heads in the model's architecture. Interestingly, increasing the number of attention heads generally improved the model's performance. Experiment 7, with 8 attention heads, achieved the highest test R^2 of 0.9820, indicating that a more complex attention mechanism allows the model to better capture the relevant features in the audio inputs. However, this increase in performance must be weighed against the additional computational complexity that comes with a higher number of attention heads. It's clear that while more attention heads can enhance the model's ability to focus on important parts of the input, it also demands more resources and longer training times, which may not always be justified depending on the application.

Overall, the experiments provide valuable insights into how different hyperparameters influence the performance of an audio-to-mesh model. The results consistently highlight the effectiveness of lower learning rates, smaller batch sizes, and more attention heads, while also illustrating the limited benefits of extending training epochs. These findings set a strong foundation for future work, suggesting that while the model is performing well, there is still room for optimization. Further experiments could delve into other aspects of the model, such as exploring different architectures or integrating additional regularization techniques, to continue improving the model's ability to generalize and accurately predict mesh outputs from audio inputs.

4.3 Ablation Study

Experiment No	Component Modified	Description	Training Loss	Validation Loss	Observations
1	Baseline	All components included	0.0127	0.0059	Best generalization; balanced performance.
2	Attention Heads	Increased Attention head to 4	0.0128	0.0070	Increased complexity; slight overfitting
3	**Positional Encoding**	**Removed Positional Encoding**	**0.0058**	**0.0055**	**Reduced complexity; no sequence info needed.**
4	Temporal Coherence	Removed temporal coherence loss	0.0064	0.0077	Reduced smoothness; overfitting risk.
6	Layer Normalization	Removing Layer norm	0.0062	0.0044	Simplified model; best generalization

Figure 4.4: Ablation study results highlighting the impact of different model components on training and validation loss for the audio to mesh model

The ablation study conducted on the audio-to-mesh model, as depicted in the provided table in figure 4.4, offers a comprehensive exploration into the intricacies of deep learning architecture and the significance of each component within a model. By systematically altering specific parts of the model, the study provides critical insights into how different elements contribute to both the training process and the model's generalization ability. The table outlines the results of various experiments (Experiment 1 through Experiment 6), each focusing on a particular modification or removal of a component within the baseline model. These experiments reveal the delicate balance maintained by the baseline configuration and underscore the nuanced roles that each component plays in either enhancing or potentially hindering the model's effectiveness.

One of the intriguing findings from this study is the role of attention mechanisms, particularly as explored in Experiment 2, where the number of attention heads was increased. While multi-head attention is often celebrated for its ability to capture complex patterns by focusing on various parts of the input simultaneously, the study shows that this modification resulted in a slight increase in both training and validation losses. This suggests that adding complexity in this area could introduce overfitting, where the model begins to focus too narrowly on specific patterns that do not generalize well to new inputs. This insight challenges the common assumption that more attention heads inherently lead to better performance, highlighting the importance of carefully considering the architecture's capacity relative to the task.

The removal of positional encoding in Experiment 3 brings another dimension to the analysis. Positional encoding is typically crucial in sequence models, especially for tasks involving temporal or sequential data like audio processing. However, in this model, its removal led to a reduction in both training and validation losses, suggesting that the sequence information might already be sufficiently captured by other mechanisms within the model. This raises important questions about the necessity of positional encoding in certain contexts, indicating that simpler models can sometimes perform better by avoiding redundant components.

Temporal coherence, as examined in Experiment 4, also reveals significant insights. Temporal coherence ensures that consecutive frames in a sequence are consistent and smooth, which is vital for generating sequences like 3D meshes. Removing this component led to a reduction in training loss but a notable increase in validation loss, indicating that while the model might perform better on training data without this constraint, it struggles with generalization. The lack of temporal coherence likely results in less smooth transitions between frames, which can be problematic in applications where consistency is key.

One of the most surprising outcomes of the study comes from Experiment 6, where layer normalization was removed. Layer normalization is typically used to stabilize the training process by normalizing inputs across layers, yet its removal led to the lowest validation loss observed across all experiments. This suggests that in this particular model, layer normalization might have introduced unnecessary constraints or complexities that hindered optimal learning. The improved performance without this component challenges conventional wisdom, emphasizing the importance of empirical testing even for widely accepted techniques.

Overall, this ablation study provides valuable lessons on the delicate balance required in deep learning model design. The findings suggest that more complex architectures, which include multiple commonly accepted components, do not necessarily equate to better performance. On the contrary, simplifying the model by removing certain elements, as demonstrated in Experiments 3 and 6, can lead to better generalization and more efficient learning. The careful empirical examination of each component's contribution, as detailed in the table, is crucial for identifying elements that may be redundant or even detrimental to the model's performance.

In conclusion, the study highlights the importance of ablation experiments in refining model architectures. It delivers a clear message that in deep learning, more is not always better. Instead,

a targeted approach that considers the specific needs of the task at hand and is willing to challenge and test standard practices, can lead to more effective and efficient models. This study serves as a reminder that in the quest for optimal model performance, simplicity and specificity often outweigh sheer complexity, as clearly illustrated by the results in the table provided above.

4.4 Limitations of the Model

The audio-to-mesh model you've developed has shown promising results in generating facial meshes based on audio input. However, several limitations affect its ability to consistently produce detailed and accurate outputs across the entire sequence of frames.

4.4.1 Inconsistent Detail in Early Frames

The model's reliance on LSTM layers within the UNetDecoder and Transformer-based attention mechanisms results in a sequential processing limitation. Early frames often lack the necessary temporal context to produce detailed predictions, leading to outputs that are more global or "messy" in nature. The model gradually refines its predictions as more context becomes available, which explains the improved detail in later frames.

4.4.2 Over-Smoothing of Initial Frames

The use of temporal coherence loss, while essential for ensuring smooth transitions between frames, can lead to overly smoothed or generalized outputs in the initial frames. This smoothing effect prioritizes consistency across frames, which may cause the model to sacrifice detail in the early stages of the sequence, only becoming more specific as the sequence progresses.

4.4.3 Slow Adaptation to Detailed Features

The model's training process and initialization contribute to a slow adaptation to detailed features. During training, the model initially learns coarse global features before gradually refining its predictions. This behavior carries over into inference, where the early frames are less detailed until the model accumulates enough information to generate more refined outputs.

4.4.4 Generalization Issues Due to Limited Training Data

The model may suffer from generalization issues due to the limited amount of training data—specifically, the use of a dataset that contains only about one hour of data from a single person. This narrow scope may cause the model to struggle when applied to new or varied audio inputs, potentially

limiting its effectiveness across different speakers or audio conditions. To mitigate this issue, the model would benefit from further training with a more diverse and extensive dataset.

4.4.5 Challenges in Achieving Accurate Lip Sync

Accurate lip synchronization remains a challenge for the model, especially in the early frames. The complexity of coordinating precise facial movements with audio input requires high accuracy, which the model might not fully achieve in the initial frames. While the model improves in capturing lip sync as the sequence progresses, it still struggles with the intricacies of synchronization early on.

In summary, the model faces several limitations that impact its ability to generate consistent and detailed facial meshes throughout the sequence. These limitations, including inconsistent detail in early frames, over-smoothing, slow adaptation to detailed features, and challenges in accurate lip sync, highlight areas for further refinement. Addressing these issues through additional training, architectural adjustments, and experimentation will be crucial in enhancing the model's performance, particularly in the initial stages of the sequence.

5 CONCLUSIONS

This chapter presents a comprehensive summary of the work carried out throughout this project, highlighting the key achievements, the methodologies employed, and the insights gained. The primary objective of this project was to explore and develop an audio-driven face animation model with a specific focus on accurately representing emotions through 3D mesh sequences. To this end, an extensive literature review was conducted to understand the current landscape of audio-driven face animation models, with particular emphasis on how these models handle the expression of emotions. This foundation informed the subsequent stages of the project, including data collection, preprocessing, model development, and experimental evaluation.

5.1 Evaluation

Reflecting on the objectives outlined in Section 1.2, this project successfully navigated through all the critical stages required to develop an audio-to-mesh model, with a particular focus on animating facial expressions driven by audio inputs. The primary objective was to create a model capable of translating audio into a sequence of 3D facial meshes that accurately represent the emotional content of speech. While this objective was met in terms of developing a functional model and conducting a thorough series of experiments, the results reveal that there is still significant room for improvement in accurately capturing and representing emotional nuances within the generated mesh sequences.

The extensive literature review, one of the key objectives, was executed effectively, providing a strong theoretical foundation for the project. This review informed the design and development of the model by identifying state-of-the-art techniques and challenges in the field of audio-driven facial animation, particularly in the context of emotion expression. This knowledge was instrumental in shaping the model's architecture and the subsequent experiments conducted.

Data collection and preprocessing were also carried out with diligence, ensuring that high-quality, relevant data was used for training the model. The preprocessing steps, which included extracting audio features and aligning them with corresponding facial mesh data, were performed to a high standard, setting a solid groundwork for the model training phase.

During the model development and experimentation phase, the project aimed to explore various hyperparameters and architectural configurations to optimize performance. This objective was achieved through a comprehensive series of experiments, including an ablation study that systematically tested the impact of different model components. These experiments provided valuable insights into the model's behavior, revealing both strengths and limitations. However, the outcomes, while informative, did not fully meet expectations, particularly in terms of the model's ability to accurately grasp and reproduce emotional features in the generated meshes. The results, visualized through the training and validation loss curves (as shown in Figure 5.1), suggest that while the model was able to achieve low loss values—0.0058 for training loss and 0.0036 for validation loss—there were challenges in capturing the finer emotional nuances in the facial expressions. Despite the low loss values indicating good overall accuracy, further refinement is needed to enhance the model's sensitivity to subtle emotional variations in the audio input.

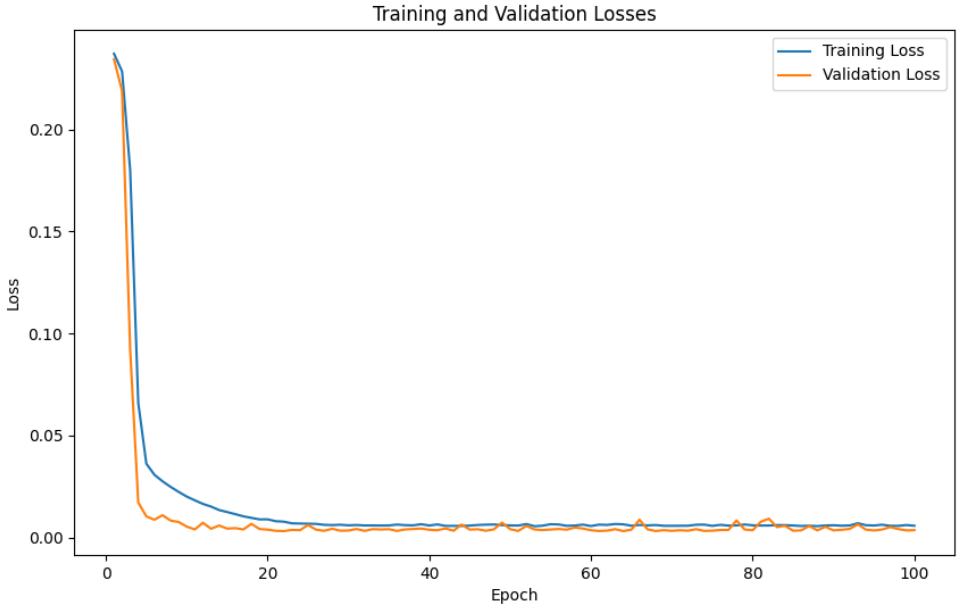


Figure 5.1: Loss plot of the best performing model

The final objective involved the visualization and evaluation of the model's performance, including the comparison of generated mesh sequences to the original data and the assessment of the model's generalization ability. This was successfully accomplished through the use of mesh offset comparison graphs and the analysis of training and validation losses. These visualizations

provided a clear indication of the model's current performance level, highlighting areas where the model excels as well as areas requiring further refinement.

Despite the successful completion of all project components in a professional and methodical manner, it is clear that the results indicate limitations in the current model. These limitations—particularly in capturing the emotional nuances in facial expressions—serve as a baseline for future work. The extensive experiments conducted offer a foundation upon which future research can build, exploring new techniques and approaches to improve the model's ability to accurately reflect the emotional content of audio inputs.

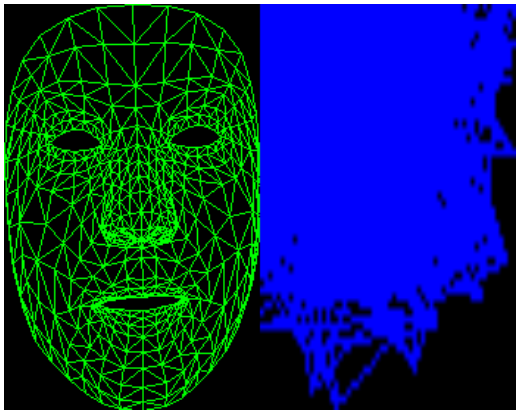
In conclusion, while the project met its objectives by successfully completing all planned tasks, the results demonstrate that there is still much work to be done to achieve the desired accuracy and effectiveness in emotional feature representation. The findings from this project, including the identified limitations and the observations from extensive experiments, provide a valuable benchmark for future efforts to enhance the performance and applicability of audio-driven facial animation models.

5.1.1 Visualization of Resiults

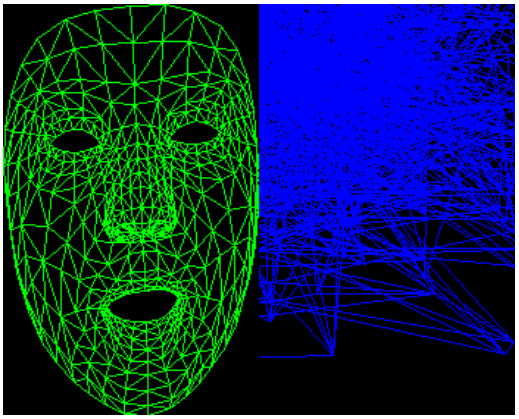
Visualizing the generated face meshes and mesh offsets at different training epochs provides critical insights into how well the audio-to-mesh model learns to translate audio features into realistic facial geometries. This step allows us to observe the model's progression and improvement in capturing the subtle details of facial expressions over time, thereby validating the effectiveness of the training process.

The face mesh comparisons at various epochs highlight the model's learning trajectory. In the early epochs (3 and 5), the generated meshes are rough approximations, capturing basic facial structures but missing finer details. By the 25th epoch, the model produces meshes with better alignment and detail, showing significant improvement. By the 50th epoch, the generated meshes closely resemble the original, indicating that the model has effectively learned to map audio inputs to detailed facial features. This progression underscores the importance of extended training for refining the model's ability to generate accurate and realistic facial meshes.

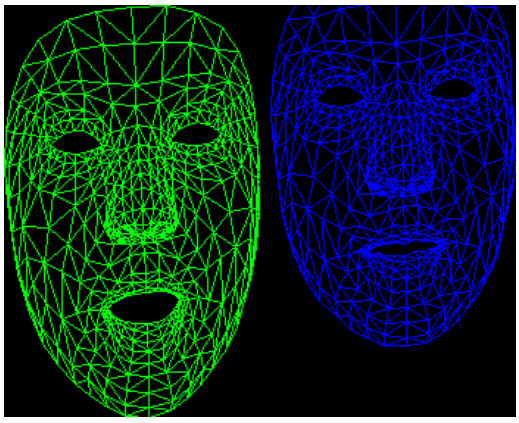
Mesh offset plots, as illustrated in the figures 5.3, offer a distinct method of evaluating the performance of the audio-to-mesh model, providing insights that differ from the direct visualization



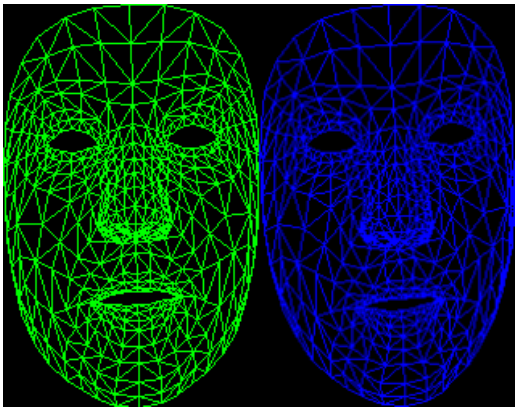
After 3 Epochs



After 5 Epochs

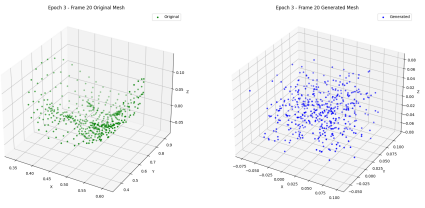


After 25 Epochs

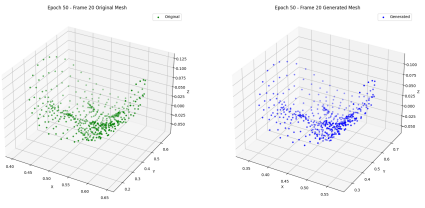


After 50 Epochs

Figure 5.2: Comparison of the original and generated face meshes from the audio-to-mesh model for a specific frame of test data, showing the progression in the model's ability to capture facial features over different stages of training.



After 3 Epochs



After 50 Epochs

Figure 5.3: Comparison of the original and generated mesh offset plots from the audio-to-mesh model for a specific frame of test data, showing the progression in the model's ability to capture facial features over different stages of training.

of face meshes. While face meshes visually depict the generated 3D facial structures, helping to assess the accuracy and realism of the model's output, mesh offset plots focus on the discrepancies in vertex positions between the original and generated meshes. These plots, particularly as seen in the comparisons after 3 epochs and 50 epochs, reveal how each vertex has shifted, enabling a more detailed analysis of the model's capacity to capture subtle facial movements and expressions over time. By examining these offsets, we can pinpoint specific areas where the model either excels or needs improvement, information that might not be as apparent from face mesh visualizations alone. Thus, the combined use of face mesh and mesh offset visualizations provides a comprehensive understanding of the model's strengths and limitations, facilitating more effective fine-tuning for enhanced performance

5.2 Future Work

Building upon the current model's capabilities and addressing its limitations, several avenues for future work could significantly enhance the performance and applicability of the audio-to-mesh generation system:

- **Improving Detail and Consistency Across Frames:** One of the key areas for future work is enhancing the model's ability to produce consistent and detailed outputs across all frames, particularly in the early stages of the sequence. This could involve experimenting with more advanced architectural techniques, such as incorporating bidirectional LSTMs or attention mechanisms that can better capture context from both past and future frames simultaneously. Additionally, exploring multi-scale feature extraction methods might help the model better capture both global and fine-grained details from the audio input, resulting in more accurate mesh predictions from the start of the sequence.

- **Addressing Over-Smoothing and Enhancing Temporal Coherence:** To balance the need for smooth transitions with the requirement for detailed frame-by-frame accuracy, future work could focus on refining the temporal coherence loss function or introducing new loss functions that specifically target early frame accuracy. For example, implementing a curriculum learning approach, where the model is gradually exposed to more complex sequences, could help it learn to generate detailed outputs earlier in the sequence without compromising smoothness.

- **Expanding Training Data for Better Generalization:** To overcome the model's current generalization issues, future research should involve training on a more diverse and extensive dataset. By incorporating data from multiple speakers, varying audio conditions, and different languages, the model could become more robust and versatile. This expanded dataset would allow the model to better generalize to new and unseen audio inputs, improving its applicability across different contexts and making it more reliable for real-world use.

- **Enhancing Audio Feature Representation:** Given the limitations of using a modified ResNet for audio feature extraction, future work could explore alternative approaches that are more tailored to audio data. For instance, employing models specifically designed for audio processing, such as Wav2Vec 2.0 or Transformer-based audio encoders, could lead to

more accurate and nuanced feature extraction. These improved audio features could help the model generate more detailed and expressive meshes throughout the sequence.

- **Leveraging Generated Meshes in Generative and Diffusion Models:** The meshes generated by this model, which are produced directly from audio without any reference images, present a unique opportunity for use in generative models or diffusion models. These meshes could serve as target reference meshes for more advanced generative approaches, potentially leading to the creation of highly realistic and expressive facial animations. By integrating these generated meshes with generative adversarial networks (GANs), diffusion models, or other generative techniques, future research could explore the possibility of creating high-quality animations that faithfully represent the input audio, with improved expression and movement accuracy.

- **Incorporating Headpose Models and Control Nets:** To further enhance the model's performance, particularly in capturing head movements and overall facial dynamics, future work could integrate headpose models or control nets into the audio-to-mesh generation pipeline. These models could provide additional control and guidance for generating realistic head and facial movements, ensuring that the generated meshes not only capture fine details but also adhere to natural movement patterns. This integration could be particularly beneficial for applications requiring synchronized audio-visual output, such as virtual avatars or realistic speech animation.

- **Further Experimentation with Expression and Movement Capture:** Lastly, more experimentation is needed to analyze how well the model captures various facial expressions and movements. Future work could involve testing the model on a wider range of emotional expressions and dynamic movements to assess its performance and identify areas for improvement. This could also include exploring the impact of different audio characteristics, such as pitch, tone, and speed, on the generated meshes, and adjusting the model accordingly to better capture the nuances of human expression.

BIBLIOGRAPHY

[1] A Parametric Model for Human Faces. https://collections.lib.utah.edu/ark:/87278/s67q1sd6, 1974. Accessed: 14 April 2024.

[2] AnimateAnyone Group. AnimateAnyone: Animation from a Single Image using Pre-trained Diffusion Models. *arXiv preprint arXiv:2311.17117*, 2023.

[3] A Baevski, H Zhou, A Mohamed, and M Auli. Wav2Vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *arXiv preprint arXiv:2006.11477*, 2020.

[4] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999.

[5] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[6] C Bregler, M Covell, and M Slaney. Video Rewrite: Driving Visual Speech with Audio. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '97*, pages 353–360. ACM Press/Addison-Wesley Publishing Co., 1997. doi: 10.1145/258734.258880.

[7] C Busso, S Parthasarathy, A Burmania, M AbdelWahab, N Sadoughi, and E Mower Provost. MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception. *IEEE Transactions on Affective Computing*, 8(1):67–80, 2017. doi: 10.1109/TAFFC.2016.2515617.

[8] M Cohen and D W Massaro. Modeling Coarticulation in Synthetic Visual Speech. *Speech Communication*, 13(1), 1993.

[9] CMU Arctic Database. CMU Arctic Database: Speech Synthesis Database. `http://festvox.org/cmu_arctic/cmuarctic.data`.

[10] F Eyben, M Wöllmer, and B Schuller. OpenSMILE: the Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 1459–1462, 2010.

[11] T Ezzat, G Geiger, and T Poggio. Trainable Videorealistic Speech Animation. *ACM Transactions on Graphics*, 21(3):388–398, 2002. doi: https://doi.org/10.1145/566654.566594.

[12] N. Harte and E. Gillen. TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech. In *IEEE Transactions on Multimedia*, pages 569–582, 2015.

[13] A. Senior O. Vinyals J. Chung, J. S. Chung and A. Zisserman. LRW: Lip Reading in the Wild Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3444–3453, 2019.

[14] O. Vinyals J. Chung, A. Senior and A. Zisserman. Lip Reading Sentences in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3444–3453, 2016.

[15] O. Vinyals J. Chung, A. Senior and A. Zisserman. VoxCeleb2: Deep Speaker Recognition. In *Proceedings of the Interspeech*, pages 1086–1090, 2018.

[16] T Karras, T Aila, S Laine, A Herva, and J Lehtinen. Audio-driven Facial Animation by Joint End-to-End Learning of Pose and Emotion. *ACM Transactions on Graphics*, 36(4):1–12, 2017. doi: https://doi.org/10.1145/3072959.3073658.

[17] Camillo Lugaresi, Fan Tang, Luke Nash, Connor McClanahan, Lance Hertel, Justin Huddleston, Justin Schoenblum, Alex Jepson, David Fuller, Wayne Thies, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

[18] S. Cunningham M. Cooke, J. Barker and X. Shao. An Audio-Visual Corpus for Speech Perception and Automatic Speech Recognition. In *Journal of the Acoustical Society of America*, pages 2421–2424, 2006.

[19] S. He M. Zieba S. Petridis M. Stypułkowski, K. Vougioukas and M. Pantic. Diffused Heads: Diffusion Models Beat GANs on Talking-Face Generation. *arXiv preprint arXiv:2301.03396*, 2023. doi: 10.48550/arXiv.2301.03396.

[20] T.-H Oh, T Dekel, C Kim, I Mosseri, W.T Freeman, M Rubinstein, and W Matusik. Speech2Face: Learning the Face Behind a Voice. *arXiv preprint arXiv:1905.09773*, 2019. doi: https://doi.org/10.48550/arXiv.1905.09773.

[21] R Rombach, A Blattmann, D Lorenz, P Esser, and B Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv preprint arXiv:2112.10752*, 2021.

[22] Harvard Sentences. Harvard Sentences: Phonetically Balanced Sentences for Speech Testing. `https://www.cs.columbia.edu/~hgs/audio/harvard.html`.

[23] A. Senior O. Vinyals T. Afouras, J. S. Chung and A. Zisserman. Deep Audio-Visual Speech Recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 158–172, 2018.

[24] S.L Taylor, M Mahler, B.-J Theobald, and I Matthews. Dynamic Units of Visual Speech. In *Proceedings of the Conference on Computer Graphics and Interactive Techniques*, page 275–284, 2012. doi: https://doi.org/10.5555/2422356.2422395.

[25] D. Povey V. Panayotov, G. Chen and S. Khudanpur. LibriSpeech: An ASR Corpus Based on Public Domain Audio Books. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

[26] A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A N Gomez, and I Polosukhin. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30, 2017.

[27] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Bjorn W. Schuller. DAWN OF THE TRANSFORMER ERA IN SPEECH EMOTION RECOGNITION: CLOSING THE VALENCE GAP. In *Proceedings of the [Conference Name]*. audEERING GmbH, Gilching, Germany; EIHW, University of Augsburg, Augsburg, Germany; GLAM, Imperial College, London, UK, 2024. $*$.

[28] Kenji Waters and Demetri Terzopoulos. Realistic talking heads using multilayer neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 776–779. IEEE, 1993.

[29] H Wei, Z Yang, and Z Wang. AniPortrait: Audio-Driven Synthesis of Photorealistic Portrait Animation. *arXiv preprint arXiv:2403.17694*, 2024. Accessed: 14 April 2024.

[30] L Xie, X Wang, H Zhang, C Dong, and Y Shan. VFHQ: A High-Quality Dataset and Benchmark for Video Face Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 657–666, 2022.

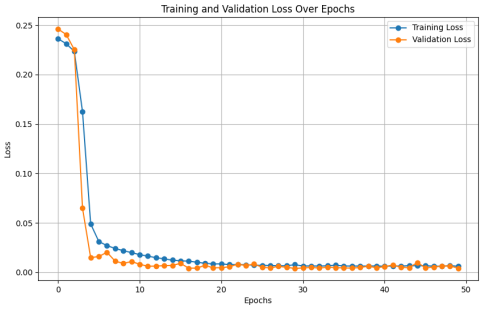
[31] W Yu, Y Fan, Y Zhang, X Wang, F Yin, Y Bai, Y Cao, S Ying, Y Wu, Z Sun, and B Wu. NOFA: NeRF-based One-shot Facial Avatar Reconstruction. *arXiv preprint arXiv:2301.03396*, 2023. doi: https://doi.org/10.48550/arXiv.2301.03396.

[32] L Zhang, A Rao, and M Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, page 3836–3847, 2023.

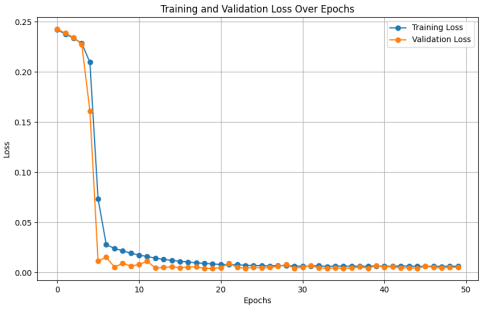
[33] Z Zhang, L Li, Y Ding, and C Fan. Flow-guided One-shot Talking Face Generation with a High-Resolution Audio-Visual Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 3661–3670, 2021.

[34] H Zhu, W Wu, W Zhu, L Jiang, S Tang, L Zhang, Z Liu, and C.C. Loy. CelebV-HQ: A Large-Scale Video Facial Attributes Dataset. In *European Conference on Computer Vision*, page 650–667. Springer, 2022.

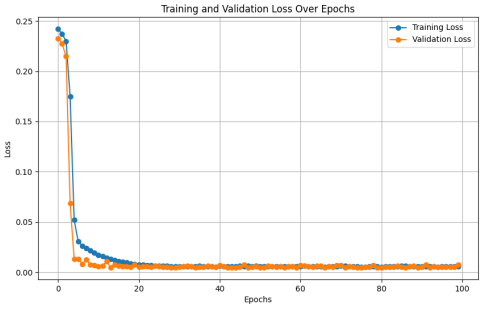
APPENDIX



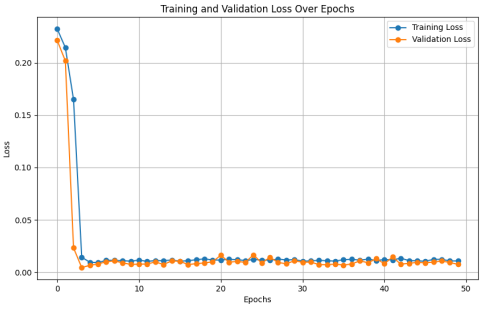
(a) Experiment 1



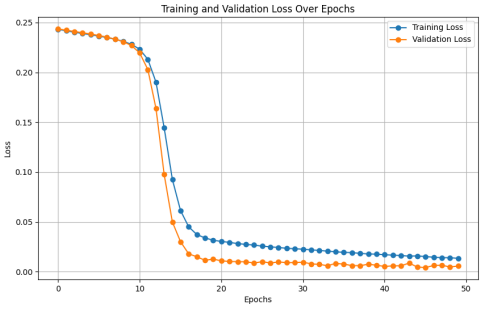
(b) Experiment 2



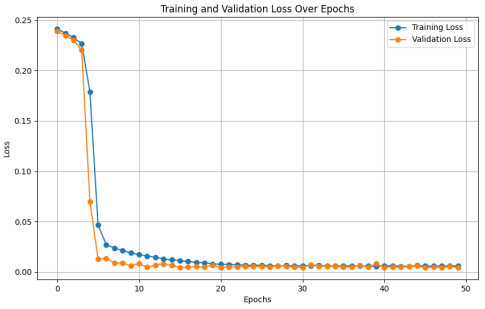
(c) Experiment 3



(d) Experiment 4

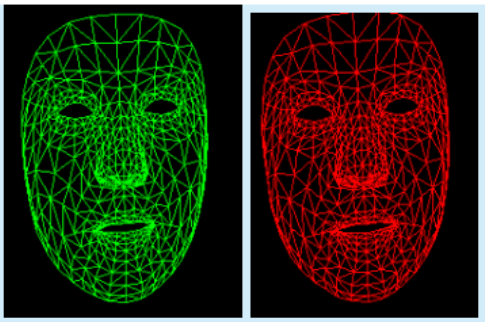
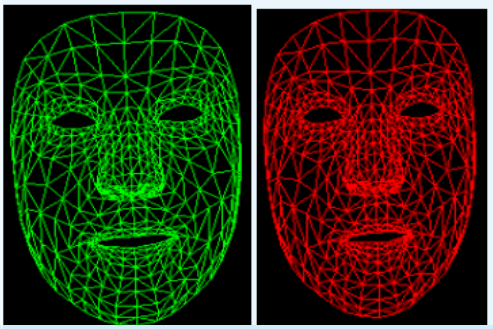
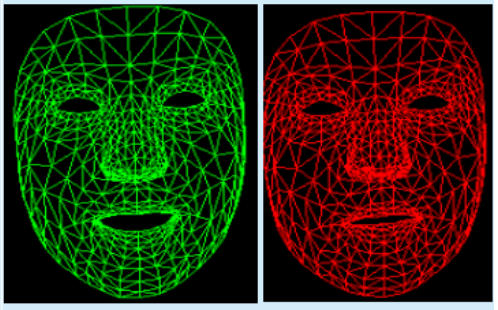
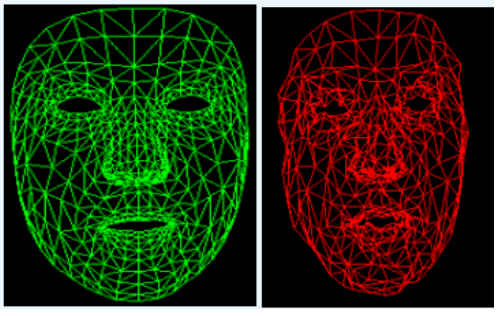


(e) Experiment 5



(f) Experiment 6

Figure 5.4: Training vs. Validation Loss for Different Models/Configurations.

Experiment	Learning Rate	Batch Size	Epochs	Attention Heads	Visualization of Frame 10
1	0.0001	1	50	2	
2	0.0001	1	100	2	
3	0.005	1	50	2	
4	0.0005	1	50	2	

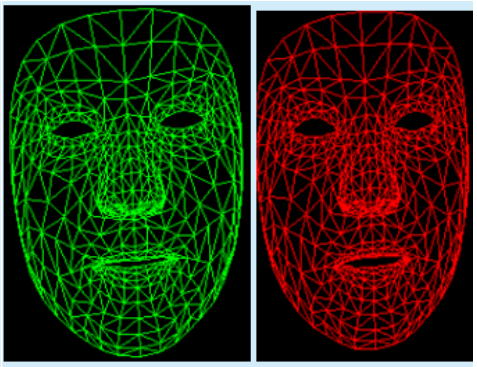
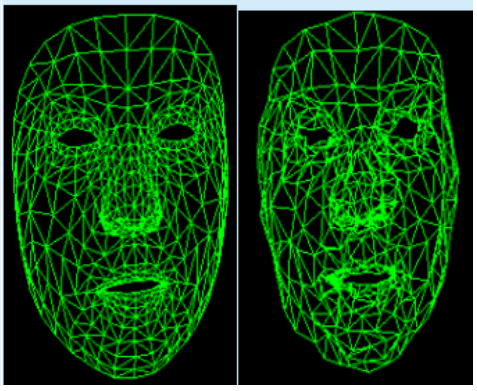
5	0.0001	4	50	2	
6	0.0001	1	50	8	

Figure 5.5: Visualization and experiments with different model parameters.

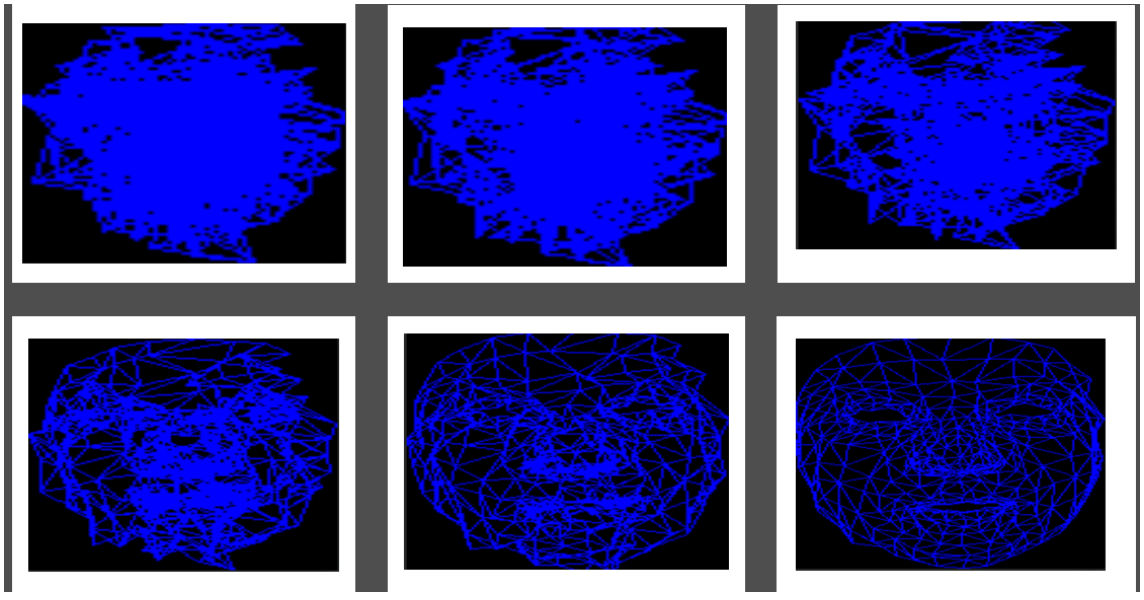


Figure 5.6: Generated 3D facial meshes from the AudioConditionedNetwork model using a completely new audio input of a different person that was not part of the training data. This figure demonstrates the model's ability to generalize and produce realistic facial expressions for unseen audio inputs

```
Vertices for batch 0:
[[ 0.49437982  0.66405731 -0.04074001]
 [ 0.49182197  0.59513557 -0.06716326]
 [ 0.49200445  0.62029803 -0.03726862]
 ...
 [ 0.50707453  0.48300251 -0.00164085]
 [ 0.57412052  0.46106011  0.02641658]
 [ 0.58011287  0.45413163  0.0272582 ]]
```

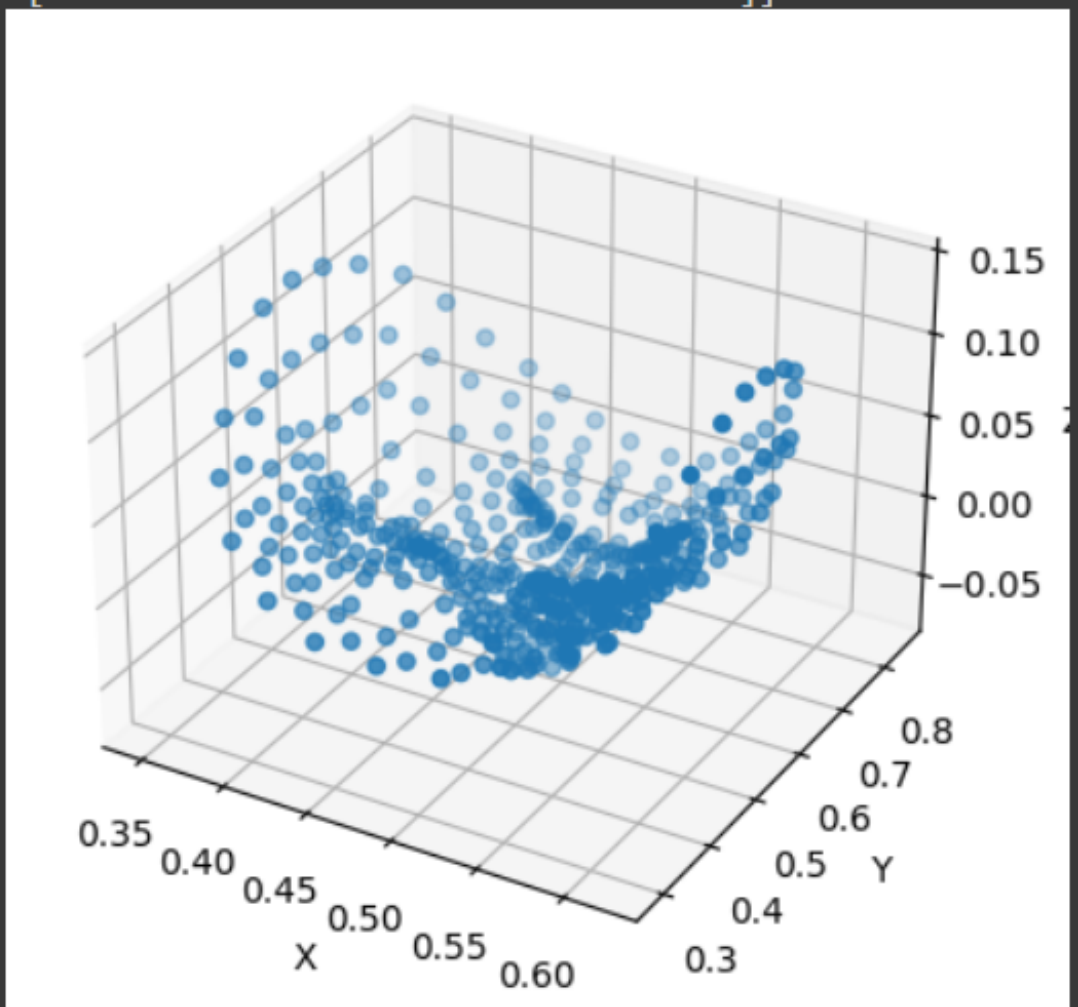


Figure 5.7: 3D plot visualizing mesh offsets in the first batch of test data, showing the displacement of each vertex (in X, Y, Z directions) where each point corresponds to a specific vertex's adjustment from its baseline position

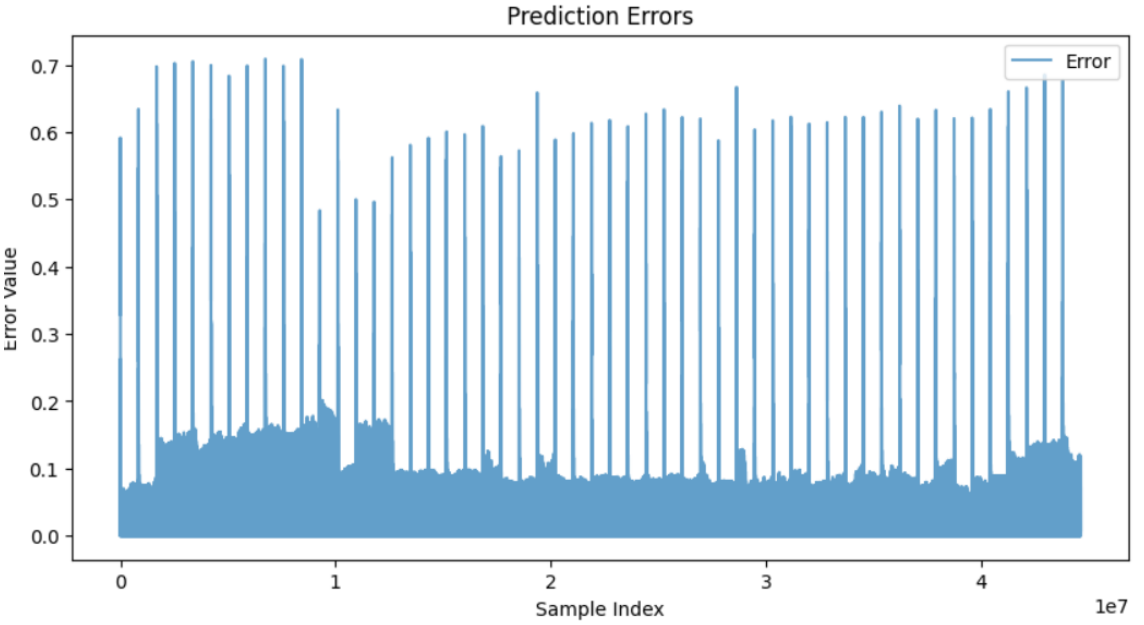


Figure 5.8: Prediction error plot showing the error values between actual and predicted outputs across test samples; the Sample Index represents the sequence of data points, while the Error Value indicates the discrepancy between predictions and ground truth