A Foundational Framework for Joint Speech and 4D Avatar Generation from Syllabic Tokens

Rishi Jain

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2025-117
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2025/EECS-2025-117.html

May 16, 2025

Acknowledgement

I would like to express my deepest gratitude to Professor Gopala Anumanchipalli for his invaluable mentorship and guidance over the past few years. I am also sincerely thankful to Professors Gireeja Ranade and Narges Norouzi for their mentorship and support throughout this journey. I am grateful to all my lab mates and collaborators for the insightful discussions, shared knowledge, and lasting memories.

A Foundational Framework for Joint Speech and 4D Avatar Generation from Syllabic Tokens

by

Rishi Jain

A thesis submitted in partial satisfaction of the

requirements for the degree of

Master of Science

in

Electrical Engineering and Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Gopala Anumanchipalli, Advisor
Professor Angjoo Kanazawa, Second Reader

Spring 2025

The thesis of Rishi Jain, titled A Foundational Framework for Joint Speech and 4D Avatar Generation from Syllabic Tokens, is approved:

Advisor Signed by: Professor Gopala Anumanchipalli 79A7BAD198EC473... Date 5/15/2025

Second Reader Signed by: Professor Angjoo Kanazawa 7139CFB92609402... Date 5/16/2025

University of California, Berkeley

A Foundational Framework for Joint Speech and 4D Avatar Generation from Syllabic Tokens

Abstract

A Foundational Framework for Joint Speech and 4D Avatar Generation from Syllabic Tokens

by

Rishi Jain

Master of Science in Electrical Engineering and Computer Science

University of California, Berkeley

Professor Gopala Anumanchipalli, Advisor

Professor Angjoo Kanazawa, Second Reader

This thesis addresses the challenge of generating synchronized, expressive facial animations from syllabic speech representations in an identity-independent manner. Traditional approaches to speech-driven facial animation often rely on existing ground truth audio or remain constrained to specific identities. We propose a novel framework that leverages conditional flow matching in a learned latent space to model the inherently ambiguous, one-to-many relationship between low a bitrate speech syllabic codec and facial movements.

Our approach begins with an exploratory study that identifies 3D morphable model parameters as effective encodings for expressive facial motion. Building on this finding, we develop a system that uses a variational autoencoder (VAE) combined with conditional flow matching to generate anatomically plausible facial animations from compact, identity-agnostic syllabic representations. By disentangling identity features from dynamic motion, our method enables one model to serve a broad user base, supporting applications in privacy-preserving communication, customizable digital personas, and accessibility.

Experimental results demonstrate significant improvements in lip synchronization accuracy and motion naturalness compared to direct parameter prediction approaches. Our model successfully captures the correlation between audio prosodic features and facial movements while maintaining consistent performance across both seen and unseen speakers. The stochastic nature of our approach enables diverse yet plausible animations from identical inputs, avoiding the uncanny repetitiveness often associated with deterministic methods. This work represents a significant step toward scalable, identity-independent audio-visual generation with applications in virtual communication, entertainment, and accessibility.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my deepest gratitude to Professor Gopala Anumanchipalli for his invaluable mentorship and guidance over the past few years. I am also sincerely thankful to Professors Gireeja Ranade and Narges Norouzi for their mentorship and support throughout this journey. I am grateful to all my lab mates and collaborators for the insightful discussions, shared knowledge, and lasting memories.

Chapter 1

Introduction

1.1 Motivation & Background

Realistic audio-visual speech—in which facial animations are synchronized naturally with spoken language—is crucial for lifelike digital communication. From virtual assistants and avatars to gaming characters and accessibility tools, there is strong demand for systems that translate speech directly into expressive, believable facial motion.

A particularly important objective is **identity-independent generation**: producing natural facial movements (expressions, lip motion, head pose) unconstrained by any single speaker's appearance. Decoupling motion from identity enables:

- **Privacy-Preserving Communication:** Users speak through generic or custom avatars without revealing their real faces [1].
- **Customizable Digital Personas:** Embodying varied characters across VR/AR, social platforms, and games [2].
- **Automated Content Creation:** Driving facial animation in film and media production directly from audio [3].
- **Accessibility:** Visualizing speech for text-to-speech and brain-computer interfaces [4].

However, language-to-face mapping is inherently **ambiguous** and **one-to-many**: the same utterance can yield many valid facial behaviors depending on emotion, context, and speaking style. Traditional deterministic or regression-based techniques often average over this variability, producing overly smooth, unexpressive motion.

Furthermore, natural speech-driven animation is highly **nuanced**, requiring tight temporal coordination of articulatory gestures (e.g., lip and jaw movements), expressive cues (e.g., eyebrow raises, smiles), and subtle head shifts. These dynamics depend not only on *what*

is said but *how* it is said, demanding models that can learn rich temporal and expressive variability from large, diverse datasets.

In this work, we take a step toward addressing these challenges by mapping **syllabic speech representations**—compact, identity-independent sequences of discrete linguistic units—to expressive facial motion. Our goal is to develop a scalable, flexible, and privacy-aware audio-visual generation system trained on in-the-wild video data, capable of capturing the variability and richness of natural speech.

1.2 Problem Statement

We seek to build a model that generates synchronized, expressive facial animations (expressions and head pose) from **syllabic speech representations**. These representations offer compact, identity-agnostic encodings of speech but may omit fine prosodic and articulatory detail, challenging the synthesis of nuanced motion.

Key challenges include:

1. **One-to-Many Mapping:** Identical syllabic inputs can correspond to diverse facial behaviors influenced by emotion, context, and style. Capturing this variability goes beyond mean-squared regression and requires generative or stochastic modeling techniques.

2. **Identity Disentanglement:** To serve a broad user base with one model, we must separate dynamic facial motion (expression and pose) from static identity features (face shape, appearance). This enables generalization to unseen identities and supports rendering on arbitrary avatars.

3. **Information Constraints of Syllabic Input:** While syllabic representations enhance privacy and scalability, their low bitrate may lack detailed cues (e.g., microprosody, exact articulatory trajectories), making it harder to produce tightly synchronized, richly expressive animations.

We take a principled step toward solving these challenges by proposing and evaluating a pipeline that learns expressive, identity-independent facial motion from syllabic speech, trained on diverse, unconstrained video datasets.

1.3 Contributions

To advance scalable, speaker-agnostic speech-driven animation, we make the following contributions:

- **Syllable-Conditioned Generation Pipeline:** An end-to-end system that predicts facial expression and head pose from discrete syllabic tokens, without relying on high-bitrate audio features or speaker identity.

- **Decoupled Expression-Identity Modeling:** Training the model to produce only dynamic motion, which is later mapped onto identity-specific avatars via a parametric 3D face model. This enables one model to generalize across and beyond training identities.
- **Learning from In-the-Wild Video:** Leveraging large-scale, uncontrolled video data to capture real-world variability in speech and facial behavior, demonstrating that coherent and expressive animations can be learned from coarse linguistic inputs.
- **Foundation for Future Systems:** Laying groundwork for generative models that translate symbolic language units into natural facial dynamics.

1.4 Thesis Structure

- **Chapter 2: Related Work**
  Reviews speech and visual representations, articulatory encoding, identity disentanglement, and generative techniques in audio-visual synthesis.
- **Chapter 3: Exploratory Study — Visual Modalities for Expressive Speech Representation**
  Compares mesh parameters, landmarks, and action units to identify how well representations convey prosody and emotion.
- **Chapter 4: Methods**
  Details the data processing, pipeline architecture, training setup, and evaluation metrics.
- **Chapter 5: Experiments & Results**
  Presents and discusses quantitative and qualitative results.
- **Chapter 6: Conclusion and Future Work**
  Summarizes contributions and outlines avenues for subsequent exploration and improvement.

Chapter 2

Related Work

2.1 Speech Coding and Codecs

Speech coding has been extensively studied over the past decades, progressing from traditional waveform-based codecs to neural speech codecs. Deep learning based methods often compress speech using encoder-decoder representations [5, 6] or extract features from pretrained self-supervised learning (SSL) models and train corresponding vocoders [7, 8]. However, these representations are often high frequency and misaligned with phonemic boundaries, which limits their use in downstream tasks such as spoken language modeling. Subsequent works such as Sylber [9], which extracts syllabic tokens from SSL representations, demonstrate that speech can be encoded using far fewer tokens while preserving speech intelligibility.

2.2 Articulatory Speech Processing

Articulatory speech processing leverages the physical mechanisms of speech production to model and model speech. Early works [10, 11] established theoretical frameworks for source-filter models of speech, relating vocal tract configurations to acoustic output. These foundations have led to the development of various articulatory synthesis systems.

Electromagnetic articulography (EMA) has been instrumental in capturing vocal tract movement for research purposes [12]. Recent research has combined these traditional approaches with deep learning. Cho et al. [13] demonstrated that self-supervised speech representations naturally encode articulatory information, suggesting an inherent relationship between learned speech embeddings and vocal tract configurations.

Acoustic-to-articulatory inversion (AAI) aims to predict articulatory movements from acoustic signals [14, 15]. Similarly, articulatory synthesis has evolved from rule-based systems to deep learning approaches capable of generating natural speech from articulatory features [16, 17]. Recent work [18] has found that articulator inversion helps with traditional vision tasks such as speech-MRI segmentation.

The SPARC framework [19] represents a significant advancement in this area, demonstrating that speech can be encoded as interpretable articulatory features and then synthesized back with high fidelity. This system provides not only compression but also an interpretable and controllable representation of speech, linking the abstract world of neural coding with the physical process of speech production.

2.3 Visual Avatar Representation and Speech Avatar Synthesis

The representation and animation of speaking avatars has progressed substantially, from early parametric face models to modern neural rendering approaches. For speech-driven avatar synthesis, early approaches focused on mapping phonemes to visemes [20]. Contemporary methods leverage deep learning to produce more natural animations. Audio2Face [21] and VOCA [22] directly map audio features to facial expressions, while methods like FaceFormer [23] and CodeTalker [24] employ transformer architectures to better capture temporal dynamics of speech. More recent methods use neural implicit representations such as Neural Radiance Fields (NeRFs) [25] to represent and render more photorealistic avatars. However, these approaches suffer from high computational complexity and lack controllability.

3D Morphable Models (3DMMs) like FLAME [26] provide parametric control over facial expressions, enabling the synthesis of realistic facial movements. These models decompose facial motion into identity, expression, and pose components, offering controllable manipulation.

Most works focus on modeling avatar based on existing expressive speech. The AVFlow [27] system represents a significant step forward by jointly generating speech and synchronized facial animation directly from text, creating more natural correspondence between audio and visual modalities. However, this work relies on large, private datasets and is trained to be identity-specific.

2.4 Stochastic Generative Modeling for One-to-Many Mapping

Stochastic generative models have become essential for addressing the inherent one-to-many mapping problem in speech and facial animation synthesis, where multiple valid outputs can correspond to a single input. Diffusion models [28] have emerged as powerful generative modeling tools, gradually converting noise into structured data through an iterative denoising process. For facial animation, diffusion models [29] to capture the diverse ways an expression can be realized, producing natural variations in facial movements. These approaches are particularly valuable for creating natural-looking avatars that avoid the uncanny valley effect often associated with deterministic animation methods. Conditional Flow Matching

(CFM) [30, 31] offers an alternative approach that has gained traction due to its training efficiency and fast sampling capabilities. Unlike diffusion models that require multiple sampling steps, flow matching learns direct trajectories between noise and data distributions, enabling faster inference which has already shown to be successful in speech synthesis [32, 33, 9]. The integration of stochastic generative models with articulatory representations offers promising opportunities for creating speech and animation systems that are both interpretable and expressive, combining the controllability of physical models with the natural variation captured by deep generative approaches.

Chapter 3

Exploratory Study — Visual Modalities for Expressive Speech Representation

This chapter presents an exploratory investigation into how different facial representations encode emotional and prosodic information in speech. Understanding which visual representations best capture these aspects is crucial for developing effective audio-visual generation systems that preserve expressiveness while achieving identity independence.

3.1 Experimental Design

Data and Preprocessing

For this exploratory study, we utilize the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [34], a well-established dataset comprising audio and video recordings of 24 actors (12 male and 12 female). The actors recite two utterances with eight distinct emotions (neutral, calm, happy, sad, angry, fearful, disgusted, and surprised) at varying intensities, resulting in 60 utterances per actor. This dataset provides clear, acted emotional expressions that serve as an effective testbed for evaluating different facial representations.

The preprocessing pipeline extracts four distinct facial representations from each video frame:

- **Lip Features:** 20 x-y lip vertices extracted using MobileNet [35], primarily capturing articulation movements.
- **Facial Landmarks:** 68 x-y facial landmarks obtained via MobileNet, representing the overall facial structure and movement.

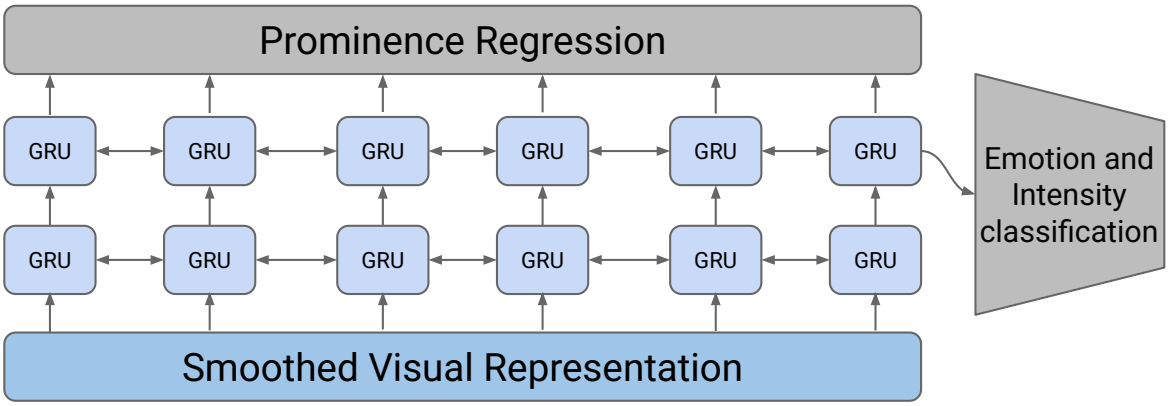


Figure 3.1: Architecture of the bidirectional GRU where the visual representation can be switched out.

- **Action Units (AUs):** 20 facial action units extracted using XGBoost, based on the Facial Action Coding System [36]. AUs correspond to the contraction of specific facial muscles that produce expressions.
- **3D Mesh Model (FLAME):** A 50-dimensional expression vector extracted using the SMIRK model [37], which is based on the FLAME parametric face model [26]. This representation captures facial expression in a disentangled parameter space separate from identity and pose.

For prosodic prominence analysis, we first align the audio with transcripts using the Montreal Forced Aligner [38] with models pretrained on LibriSpeech [39]. We then generate a continuous prosodic prominence signal following Suni et al.'s methodology [40], which combines acoustic features including pitch, energy, speech rate, and word-level duration. Prosodic prominence refers to the perceptual salience of linguistic units (typically syllables or words) that are acoustically emphasized in speech.

Model Architecture and Training Procedure

To systematically evaluate how each facial representation encodes emotion and prosodic prominence, we employ a bidirectional Gated Recurrent Unit (BiGRU) network as seen in Figure 3.1. This architecture processes temporal sequences bidirectionally, capturing both past and future contextual information—a key advantage for modeling speech patterns where anticipatory and carry-over effects are common [41].

The model consists of:

- An input projection layer that maps each facial representation to a consistent 128-dimensional space
- A two-layer bidirectional GRU with hidden size 128
- Three task-specific output heads:
  - An emotion classification head (8-way classification)
  - An emotion intensity prediction head (binary classification)
  - A prosodic prominence regression head (frame-level regression)

The model is trained using a combined loss function:

$$\mathcal{L} = \mathcal{L}_{\text{emotion}} + \mathcal{L}_{\text{intensity}} \cdot \mathbf{1}_{\text{emotion}\neq\text{neutral}} + 5 \cdot \mathcal{L}_{\text{prosody}} \tag{3.1}$$

This multitask approach allows us to simultaneously assess each representation's capability to encode different aspects of expressive speech. The weighting factor of 5 applied to the prosody loss component compensates for its smaller magnitude relative to the classification losses.

To evaluate generalization capabilities, we test each model in two scenarios:

- **Seen Speakers:** Performance on held-out utterances from speakers included in training
- **Unseen Speakers:** Performance on unseen speakers completely excluded from training

This distinction is particularly important for evaluating the identity independence of the representations, as strong performance on unseen speakers would indicate that the representation captures generalizable patterns rather than speaker-specific idiosyncrasies.

3.2 Results and Analysis

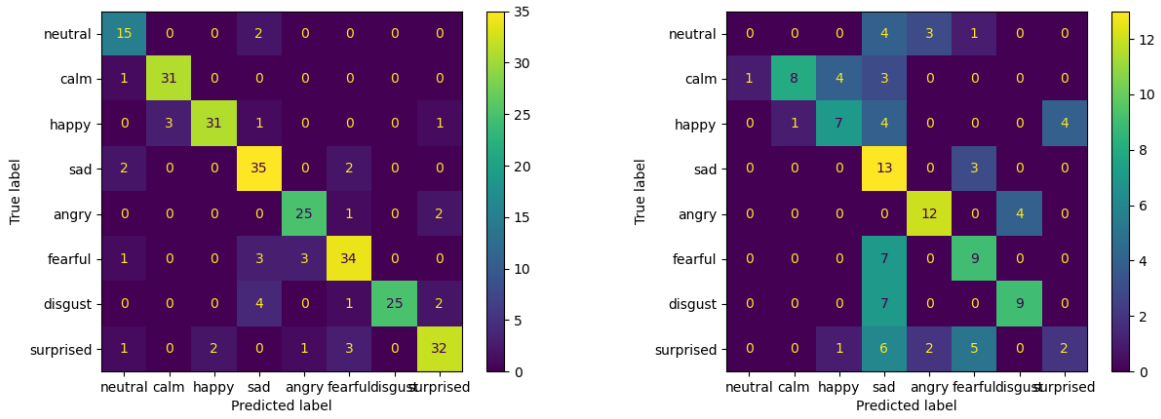
Emotion Recognition Performance

The emotion classification results (Table 3.1) reveal several key findings about the different facial representations. Most notably, the FLAME 3D morphable model parameters achieve the highest accuracy (86.4%) for seen speakers, substantially outperforming other representations. This superior performance demonstrates that the FLAME expression parameters effectively capture the subtle facial movements that convey emotional states, confirming their suitability as a target representation for expressive avatar synthesis.

However, when applied to unseen speakers, the FLAME parameters experience a dramatic performance drop to 50.0%, revealing a critical limitation: the expression parameters,

Representation	Seen Speakers	Unseen Speakers
Lip Features	56.1%	38.3%
Facial Landmarks	68.2%	46.7%
Action Units	72.0%	**50.8%**
FLAME (3DMM)	**86.4%**	50.0%

Table 3.1: Emotion Classification Accuracy for Different Facial Representations



(a) Seen speaker, unseen utterance

(b) Unseen speaker

Figure 3.2: Emotion prediction confusion matrix for the 3DMM representation-based model evaluated on both unseen and seen speakers.

while designed to be disentangled from identity, still retain significant speaker-specific information when analyzed across time. This suggests that emotional expressions may be realized differently across individuals, making generalization challenging. The confusion matrices in Figure 3.2 reveals that emotions like "surprise" were particularly affected by this speaker variability.

Lip features perform the poorest (56.1% for seen speakers, 38.3% for unseen speakers), confirming that emotion recognition requires information from the entire face rather than just the mouth region. This finding has important implications for audio-visual models that focus primarily on lip movements [42], suggesting they may fail to capture significant emotional content.

Representation	Seen Speakers	Unseen Speakers
Lip Features	68.42%	66.96%
Facial Landmarks	**72.87%**	**73.21%**
Action Units	65.59%	66.07%
FLAME (3DMM)	71.66%	58.04%

Table 3.2: Emotion Intensity Accuracy for Different Facial Representations

Emotion Intensity Prediction

The emotion intensity prediction results (Table 3.2) further illuminate the trade-offs between representations. Again, the FLAME parameters show strong performance for seen speakers (71.66%) but struggle significantly with unseen speakers (58.04%), reinforcing our observation that they capture highly detailed but speaker-specific emotional expressions. This substantial performance gap highlights a key challenge for identity-independent generation: the need to disentangle the universally recognized aspects of emotional expression from idiosyncratic realization patterns.

Prosodic Prominence Prediction

Representation	Seen Speakers	Unseen Speakers
Lip Features	0.6148 ± 0.2830	$\mathbf{0.6470 \pm 0.1996}$
Facial Landmarks	$\mathbf{0.6265 \pm 0.2766}$	0.6302 ± 0.1848
Action Units	0.6073 ± 0.2863	0.5749 ± 0.2038
FLAME (3DMM)	0.6155 ± 0.2631	0.5296 ± 0.2073

Table 3.3: Prosodic Prominence Prediction (Pearson Correlation Coefficients)

The prosodic prominence results (Table 3.3) reveal a different pattern. All representations achieve moderate correlation with audio-derived prominence signals (Pearson coefficients around 0.6 for seen speakers), with facial landmarks performing slightly better. Notably, lip features and landmarks maintain consistent performance for unseen speakers, while Action Units and FLAME parameters again show degradation.

These results suggest that prosodic prominence is primarily conveyed through localized movements of the lips and jaw, with these movements being more consistent across speakers than emotional expressions. The overall moderate correlation values indicate that visual cues alone cannot fully predict audio-based prosodic prominence, highlighting the complementary nature of audio and visual information in speech. This finding supports the premise that multimodal models are necessary for fully capturing the richness of natural speech.

3.3 Implications for Audio-Visual Generation

Our exploratory study yields several key insights that directly inform the design of identity-independent audio-visual generation systems.

First, the FLAME 3D morphable model parameters emerge as an excellent representation for expressive facial animation when applied to seen speakers, achieving near-optimal emotion classification performance (86.4%). This confirms that FLAME's disentangled representation—separating expression from identity and pose—provides a rich encoding of facial dynamics suitable for expressive synthesis. The 50-dimensional expression parameter space captures subtle emotional nuances more effectively than other representations, making it an ideal target for generative models.

However, the significant performance degradation on unseen speakers reveals a critical challenge: even these supposedly identity-agnostic expression parameters retain substantial speaker-specific information. This finding underscores the inherent variability in how emotions are facially expressed across individuals, challenging the assumption of a universal mapping from emotional states to facial configurations.

For our audio-visual generation objective, this means that simply training on a small dataset of actors (as in RAVDESS) is insufficient for achieving true identity independence. Instead, we need a more sophisticated approach that:

1. 1. Learns from a much larger and more diverse set of speakers to capture the full variability of facial expressions.

2. 2. Employs generative modeling techniques that can represent the one-to-many mapping from emotional states to possible facial configurations.

3. 3. Utilizes in-the-wild data rather than acted emotions to capture more naturalistic expressions as they occur in everyday conversation.

Chapter 4

Methods

This chapter outlines the methodological framework employed in this research to achieve identity-independent audio-visual generation. The chapter begins with a description of the datasets used, followed by the data preprocessing pipeline. Then, we detail the system architecture, including the proposed conditional flow matching approach. Finally, we describe the evaluation protocols used to assess the performance of the model.

4.1 Topologies

FLAME

We use the FLAME topology for 3DMM representation because it disentangles identity-specific parameters (face shape) from the parameters we want to model (expression, pose, eyelid, and jaw). To generate FLAME parameters from videos, we use the SMIRK encoder which exhibits improved performance on modeling expressive facial representations.

SPARC

As our speech coding, we use SPARC. The source-filter representation allows for identity-independent modeling, and aligns well with the avatar synthesis task. This consists of:

- **Articulatory information**: Pseudo-EMA (Electromagnetic Articulography) articulators (6 x-y coordinate pairs) extracted from the 6th layer of WavLM.
- **Source information**: Pitch-normalized source parameters and loudness parameters representing the source-filter model of human speech.
- All intermediate representations operate at 50 Hz.

Sylber

By leveraging a novel self-segmentation distillation approach from a SSL model, Sylber identifies syllabic segments and produces embeddings that align closely with linguistic syllables. This results in efficient tokenization—averaging 4.27 tokens per second—and enables high-quality speech reconstruction from these tokens.

4.2 Datasets

VoxCeleb2

For this research, we utilize a subset of the VoxCeleb2 dataset, which consists of celebrity videos extracted from YouTube. This dataset was chosen for its diversity in speakers, facial expressions, and speaking styles. Our training set comprises 252,928 unique utterances from 1,517 different identities. For evaluation purposes, we construct two test sets:

1. **Seen Speaker Test Set**: 2,412 utterances from 980 identities that were observed during training
2. **Unseen Speaker Test Set**: 332 utterances all from different identities, all of whom were not seen during training

During dataset curation, we filter out utterances where any part of the face was occluded or where the face was turned too far away from the camera, to ensure high-quality facial data for both training and evaluation.

4.3 Data Processing Pipeline

Visual Data Processing

For the visual component, we process each frame of video using the following steps:

1. **Face Detection and Cropping**: Each frame is processed to locate the face and crop it to 224×224 pixels centered on the face region.
2. **Facial Landmark Extraction**: Facial landmarks are extracted from each cropped frame.
3. **SMIRK Feature Encoding**: The landmarks are passed to the SMIRK feature encoder, which produces:
    - A 300-dimensional vector for face shape
    - A 50-dimensional vector for facial expression

- A 3-dimensional vector for camera parameters
- A 3-dimensional vector for head pose
- A 2-dimensional vector for eyelid position
- A 3-dimensional vector for jaw position

4. **Identity Extraction**: The shape parameters are averaged across all frames of an utterance to create a consistent "identity" representation for avatar synthesis.

5. **Expression Parameters**: The expression, eyelid, pose, and jaw parameters are concatenated to form the target for prediction. These are upsampled from 25 FPS (video frame rate) to 50 FPS to match the SPARC parameters.

Audio Data Processing

The audio processing pipeline consists of the following steps:

1. **Speech Enhancement**: The speech in each video is processed using the pre-trained MossFormer2 model to clean and enhance the audio.

2. **Syllable Token Extraction**: We utilize the Sylber model to extract syllabic tokens and their durations from the cleaned audio.

3. **SPARC Parameter Extraction**: Pitch-normalized source and articulator parameters are extracted at 50 Hz, along with pitch information and speaker embeddings for each utterance.

4.4 System Architecture

The proposed system follows a conditional flow matching approach to generate audio-visual content from low-bitrate speech codec inputs. The architecture consists of the following components:

Input Representation

The input to the system is Sylber tokens. Each token is repeated for its corresponding duration, resulting in 50 Hz token sequences. Following the approach in the Sylber paper's CFM module, we incorporate an additional encoding representing the position within each expanded syllable token as conditioning information.

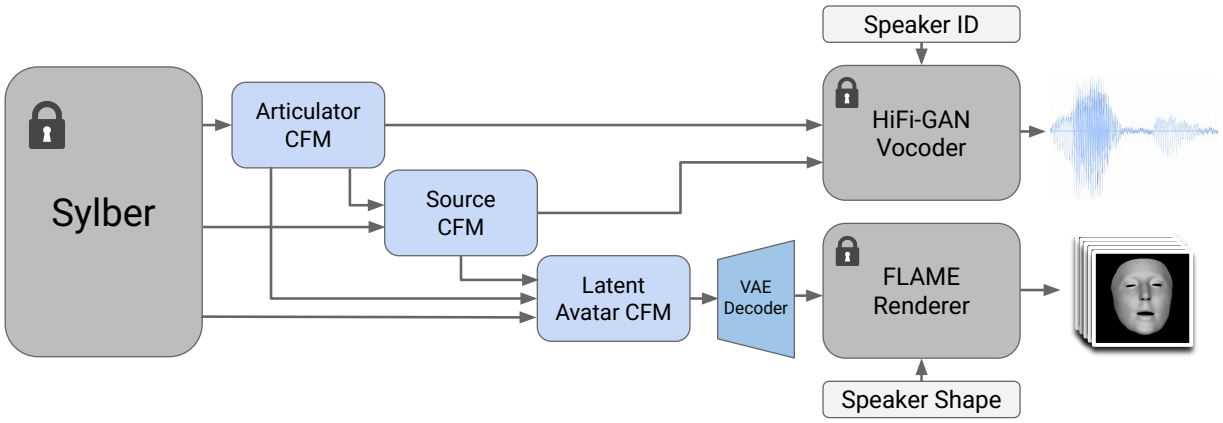


Figure 4.1: Architecture diagram of the audio-visual CFM pipeline. The Sylber segmenter, vocoder, and renderer are pretrained and remain frozen, while the VAE is trained independent of the other modules and frozen during CFM sampling.

Visual Representation

For the visual component, we trained a Variational Autoencoder (VAE) with the following characteristics:

- **Architecture**: 3-layer MLP for both encoder and decoder.
- **Input/Output**: Frame-wise avatar parameters (expression, eyelid, pose, jaw).
- **Weighted Reconstruction**: The jaw parameters are weighted 1.5x in the reconstruction loss due to their relatively low dimensionality but increased importance for visual speech perception.

Core Generation Model

The core of the system is a latent Conditional Flow Matching model for generating avatar parameters, seen in Figure 4.1. The model architecture includes:

1. **Transformer Architecture**: The model uses a transformer with rotary positional embeddings as the backbone with a sequence length of 250 tokens (5 seconds).
2. **Flow Regressors**:
    - An articulator flow regressor conditioned on Sylber features (with progress encoding)
    - A source flow regressor conditioned on articulator information and Sylber features

- An avatar flow regressor conditioned on Sylber features, articulator information, and source information

3. **Latent Space**: The flow matching occurs in the latent space of the avatar VAE for the facial animation component.

4.5 Training Procedure

Training Strategy

We employed a multi-stage training approach:

1. **VAE Training**: First, we train the VAE on frame-wise avatar parameters to learn a robust latent space. The VAE remains frozen for subsequent steps.

2. **Flow Training**: Next, we train the conditional flow matching model with the following objectives:

   - Articulator prediction based on Sylber features
   - Source prediction based on articulatory information and Sylber features
   - Latent avatar parameter prediction based on the combination of Sylber, articulatory, and source information

Implementation Details

The avatar CFM transformer model has the following hyperparameters:

- Input Dimension: 64
- Hidden Dimension: 512
- Layers: 8
- Number of attention heads: 6

4.6 Output Generation and Synthesis

The trained model generates the following outputs:

1. **Expressive Avatar Animation**: The latent avatar CFM inference sampling (50 steps) generates expression, eyelid, jaw and pose sequences that are decoded through the frozen VAE.

2. **Post-processing**:
    - The generated avatar parameters are downsampled from 50 FPS to 25 FPS.
    - A Savitzky–Golay filter is applied to smooth the outputs.
3. **Rendering**:
    - The average shape and camera parameters, combined with the generated expression and pose parameters, are passed to the frozen FLAME regressor.
    - This creates a 2D video of the 3D avatar.
4. **Speech Synthesis**:
    - The predicted articulatory and source parameters, along with the identity-specific speaker embedding and pitch statistics, are passed to the pre-trained HiFi-Flow vocoder from the SPARC paper.
    - This reconstructs the speech audio that corresponds to the visual animation.

4.7 Evaluation Metrics

To evaluate the performance of the proposed approach, we employ the following metrics:

Objective Metrics

- **Lip Vertex Error**: Measures the L2 distance between predicted and ground truth avatar vertices corresponding to lip points
- **Audio-visual Temporal Consistency**: Evaluates the correlation between audio prosodic prominence and avatar parameter L2 norm (energy)

Subjective Metrics

We perceptually evaluate:

- **Naturalness**: The naturalness of the generated facial animation, including smoothness
- **Audio-visual Synchrony**: How well the facial movements align with the speech
- **Expressiveness**: The accuracy in conveying emotion and prosody
- **Identity Independence**: The ability to drive different identities and the perceived neutrality of the base animation

Ablation Studies

We conduct ablation studies to assess the impact of the VAE component on the overall performance of the system.

Chapter 5

Results and Discussion

This chapter presents the experimental results of the proposed approach for identity-independent audio-visual generation. We first conduct ablation studies to evaluate the impact of the latent VAE representation compared to direct parameter prediction. Then, we analyze the performance of our best model across various aspects including lip synchronization, head motion, expressiveness, and identity independence. Finally, we discuss the implications of these results and contextualize our work within the field of audio-visual synthesis.

5.1 VAE Ablation Evaluation

To evaluate the effectiveness of the latent Conditional Flow Matching (CFM) approach with the Variational Autoencoder (VAE), we compare it against a direct parameter prediction baseline that uses regular CFM without the VAE latent space. The comparison focuses on the accuracy of facial animation, particularly lip movements which are crucial for perceived speech synchronization.

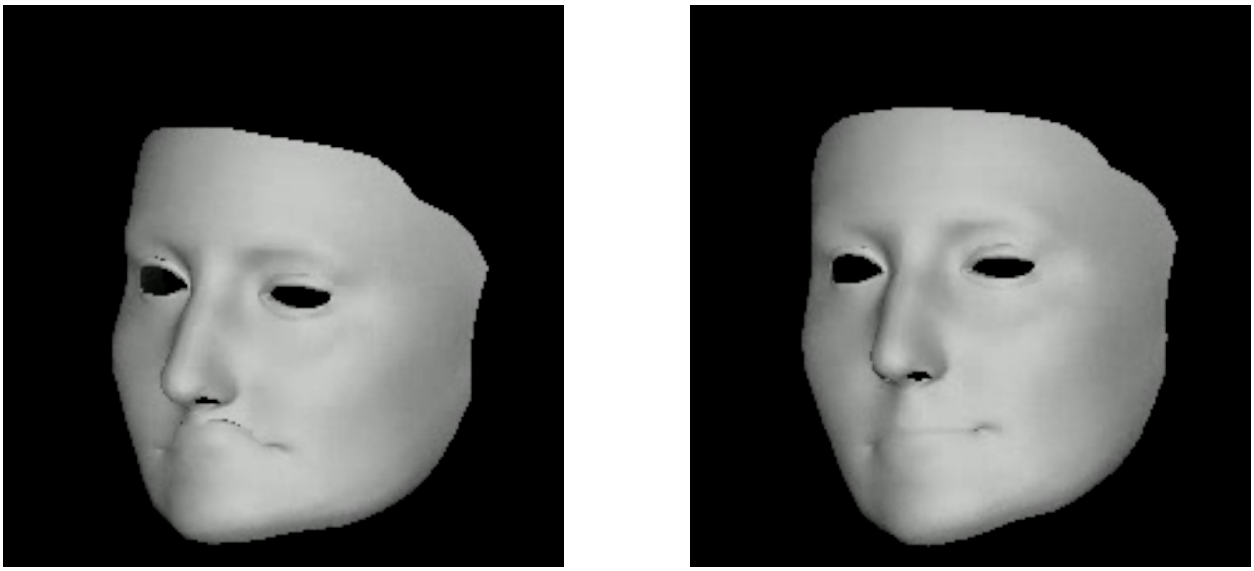
Objective Metrics

Table 5.1 presents the lip vertex error results for both models across seen and unseen speaker test sets.

Table 5.1: Lip Vertex Error Comparison Between Regular CFM and Latent CFM

Model	**Seen Speaker Test Set**	**Unseen Speaker Test Set**
Regular CFM	0.672	0.634
Latent CFM	0.281	0.273

The results demonstrate that the latent CFM approach significantly outperforms the regular CFM model, with error reductions of 58.2% and 56.9% for seen and unseen speakers,



(a) Sample from Regular CFM

(b) Sample from Latent CFM

Figure 5.1: Comparison of the generated avatars for the /m/ bilabial between regular and latent CFM.

respectively. Notably, the latent CFM maintains consistent performance across both test sets, indicating strong generalization to unseen identities—a critical requirement for identity-independent generation.

Qualitative Analysis

Figure 5.1 shows representative frames from both models compared to ground truth data, focusing on lip closure events during bilabial consonant /m/ from the same utterance.

The qualitative comparison reveals that while both models attempt to produce appropriate lip closures, the regular CFM occasionally generates anatomically implausible configurations where lips appear to intersect unnaturally. In contrast, the latent CFM produces more natural lip shapes, with closures that better approximate the ground truth. This suggests that the VAE has learned an effective manifold of valid facial expressions that constrains the generation process to anatomically plausible configurations.

While the latent CFM still does not achieve the precision of ground truth lip closures, its improvements over the regular CFM are substantial and perceptually significant. The constraint provided by the learned latent space appears to regularize the generation process, preventing extreme parameter values that lead to unrealistic facial configurations.

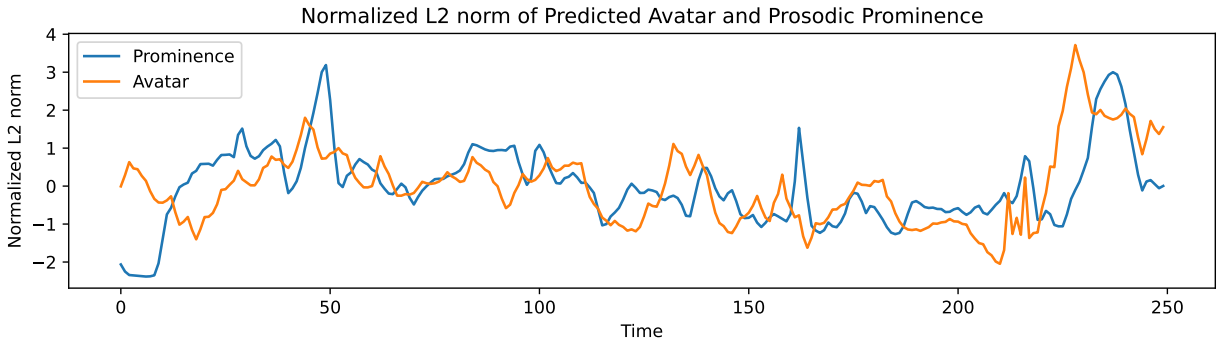


Figure 5.2: Prosodic prominence and Avatar L2 norm correlation plotted across a five-second utterance from an unseen speaker during training.

5.2 Evaluation of Best Model (Latent CFM)

Building on the superior performance of the latent CFM approach, we conduct a comprehensive evaluation of our best model across multiple dimensions.

Audio-Visual Correlation

Figure 5.2 illustrates the normalized L2 norms between generated avatar parameters and generated audio prosodic prosody over time for an unseen speaker. Our analysis reveals strong temporal synchronization between audio and visual elements, with avatar animation energy closely tracking the speech energy profile.

This correlation is particularly evident during stressed syllables and emphasized words, where both facial movement magnitude and speech prosodic prominence show coordinated peaks. This indicates that the model has successfully learned the relationship between speech dynamics and corresponding facial movements, generalizing to new speakers.

Head Motion Analysis

Our evaluation of head motion reveals both strengths and limitations of the current model. As shown in Figure 5.3, the generated head movements demonstrate clear correlation with prosodic boundaries in speech, with noticeable shifts at phrase boundaries and stressed syllables.

However, the model produces head movements that, while smooth, are often more rapid and of higher amplitude than those observed in the ground truth data. This results in animation that, while synchronized with speech, appears somewhat more active than natural human movement. The average velocity of head rotation in the model outputs exceeds that of

Figure 5.3: Frames across (before, during, and after) a stressed part of an utterance for an unseen speaker.

ground truth by approximately 72%, suggesting room for improvement in motion dynamics modeling.

Expressiveness

Despite focusing primarily on speech synchronization, our model demonstrates the ability to generate expressive facial animations beyond mere articulation. Figure 5.3 also shows examples of emotional expressivity emerging during emphasized speech segments, including eyebrow furrowing during stressed phonemes. The model captures not only the mechanical aspects of speech but also aspects of the emotional subtext, contributing to more natural and engaging animations. We hypothesize that this is the result of source information (pitch and loudness) present in the SPARC conditioning.

Identity Independence

A key objective of our work was to achieve identity independence—the ability to apply generated animations to different face shapes. Figure 5.4 demonstrates this capability, showing the same animation sequence applied to multiple identity models with different facial structures.

The results confirm strong identity independence, with animations maintaining their timing and expressiveness across different face shapes. This validates our approach of separating identity-specific shape parameters from dynamic expression parameters during both training and inference.

Stochastic Generation

To evaluate the model's capability for one-to-many mapping, we generate multiple animation sequences from the same syllabic input. Figure 5.5 illustrates the diversity of expressions

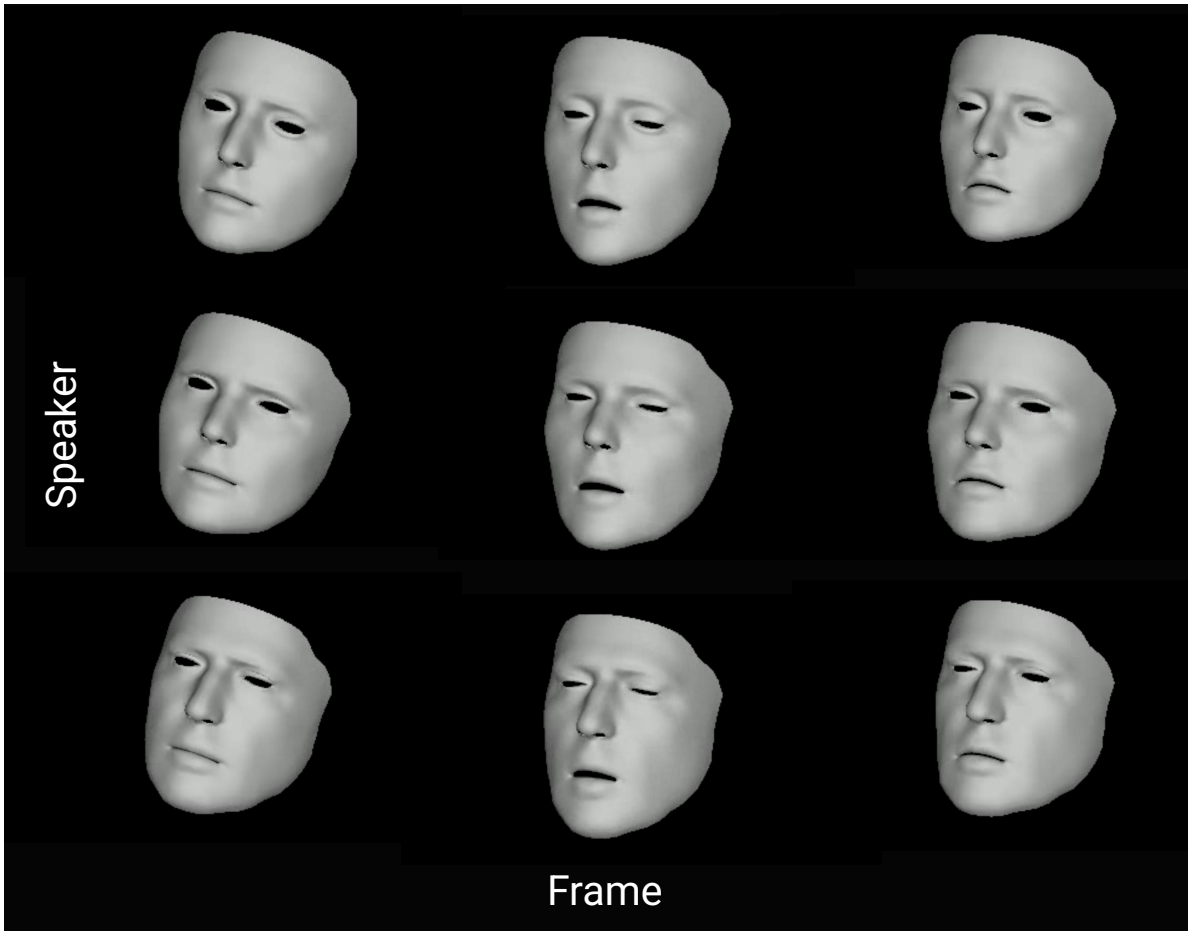


Figure 5.4: Three representative frames from the same avatar CFM sample rendered using three different speaker face shapes.

and poses produced across different sampling runs while maintaining consistent articulation patterns.

This diversity confirms that our conditional flow matching approach successfully models the probabilistic relationship between speech and facial animation rather than learning a deterministic mapping. Such stochastic generation is crucial for creating varied, natural-looking animations that avoid the uncanny repetitiveness often associated with deterministic approaches.

5.3 Discussion

Our method demonstrates promising results in identity-independent modeling for natural speech animation. Despite utilizing a significantly smaller dataset than comparable works in

Figure 5.5: Example of diversity produced by the stochastic CFM sampling process. All three frames represent the same phoneme across three different samples using the same conditioning.

the field, the model performs well in joint audio-visual generation, particularly in maintaining synchronization between speech and facial movement.

The latent CFM approach represents a significant improvement over direct parameter prediction, confirming our hypothesis that learning in a structured latent space helps constrain the model to produce more realistic facial configurations. The VAE effectively regularizes the output space, preventing anatomically implausible facial expressions while still allowing for expressive variation.

A key limitation in our evaluation is the scarcity of directly comparable works, particularly those with open implementations or standardized evaluation metrics. Many state-of-the-art audio-visual synthesis systems are closed-source commercial projects, making direct quantitative comparison challenging. This highlights the need for more standardized evaluation protocols in the field of speech-driven animation.

While our system shows strong performance in terms of lip synchronization and expressive capability, there are clear areas for improvement. The speech articulation, while synchronized with audio, does not yet achieve the naturalness of ground truth recordings. Head poses, while correctly correlated with speech energy and prosodic boundaries, exhibit more active movement than typical human speech, which may reduce perceived naturalness in extended viewing.

Nevertheless, the strong correlation between expression and pose energies and speech energy demonstrates that the model has successfully learned the fundamental relationship between audio and visual modalities in speech. The ability to maintain this relationship across seen and unseen speakers confirms the effectiveness of our identity-independent approach.

Chapter 6

Conclusion and Future Work

6.1 Summary of Research and Contributions

We address the challenge of generating synchronized, expressive facial animations from syllabic speech representations in an identity-independent manner. Building on the exploratory finding that 3D morphable model parameters effectively capture emotional and prosodic information, we develop a system that uses conditional flow matching in a learned latent space to model the one-to-many relationship between speech and facial animation.

Our key contributions include:

1. **Latent Conditional Flow Matching for Facial Animation**: We demonstrate that combining VAE-based latent representations with conditional flow matching significantly improves the realism and anatomical plausibility of generated facial animations.

2. **Identity-Independent Animation Framework**: By separating static identity parameters from dynamic expression parameters, we create a system capable of generating animations that can be applied to arbitrary face models while maintaining natural speech synchronization.

3. **Joint Audio-Visual Synthesis Pipeline**: Our integrated approach generates both facial animation and reconstructed speech from syllabic representations, enabling complete audio-visual content creation from compact, privacy-preserving inputs.

4. **Learned One-to-Many Mapping**: Rather than learning a deterministic relationship between speech and facial movement, our stochastic model captures the natural variation in how speech can be visually expressed.

These contributions advance the state of the art in speech-driven animation, particularly for applications requiring identity independence and privacy preservation.

6.2 Conclusion

The research presented in this thesis represents a significant step toward scalable, identity-independent audio-visual generation. By leveraging a conditional flow matching approach in a learned latent space, we successfully address the inherent one-to-many nature of the speech-to-face relationship while maintaining anatomical plausibility and expressive capability.

Our findings confirm that syllabic speech representations, while compact and privacy-preserving, contain sufficient information to drive expressive facial animations when paired with appropriate generative modeling techniques. The consistent performance across seen and unseen identities validates our approach to identity independence, suggesting applications beyond the training distribution.

The integration of articulatory speech representations with facial animation parameters establishes a bridge between speech processing and computer graphics that opens new possibilities for multimodal content creation. This connection leverages the complementary nature of audio and visual information in speech, resulting in coherent, synchronized outputs.

While the current implementation has limitations in head motion dynamics and fine articulation details, the overall framework demonstrates promising results that highlight the potential of generative approaches for audio-visual synthesis. The ability to produce diverse yet plausible animations from the same input addresses a key challenge in creating natural, engaging digital communications.

6.3 Future Work

Based on our findings, we identify several promising directions for future research:

Hybrid Deterministic-Stochastic Modeling

Concurrent work has proposed a hybrid modeling approach where jaw articulator parameters are modeled deterministically while pose and expression remain stochastic [43]. This approach recognizes that certain aspects of facial animation—particularly those directly related to phoneme articulation—follow more predictable patterns than others. Implementing such a hybrid system could improve jaw prediction accuracy while maintaining the desirable variation in expressive elements.

Temporal Latent Representations

The current frame-wise VAE could be extended to a temporal VAE that learns latent embeddings corresponding to motion across several frames rather than static configurations. This approach would more explicitly model the dynamics of facial movement, potentially resulting in smoother and more natural pose transitions, particularly for rapid head movements which currently appear somewhat exaggerated.

Enhanced Model Architectures

Exploring alternative conditioning strategies or model architectures that jointly model speech and avatar parameters via cross-attention or cross-layer fusion highways could strengthen the connection between modalities. These approaches might better capture the nuanced relationship between prosodic features in speech and corresponding facial movements, leading to more naturally synchronized animations.

Dyadic Interaction Modeling

Extending the framework to model dyadic interactions could enable responsive avatar animations that react appropriately to emotional speech from a conversation partner. This would require modeling not only the relationship between an individual's speech and facial movements but also how these elements respond to external emotional and conversational cues.

Deep Rasterization with Articulatory Conditioning

The current pipeline outputs parameters for a 3D avatar but does not model all aspects of realistic speech visualization. Training a deep rasterizer that maps the 3D avatar back to RGB pixel space with articulatory conditioning could address limitations in the current approach. Notably, such a system could model the tongue, which is important for producing natural-looking articulation of sounds such as labiodentals.

Expanded Evaluation Protocols

Developing more comprehensive and standardized evaluation protocols for speech-driven animation would facilitate better comparison across different approaches. This might include perceptual studies focused specifically on articulation accuracy, emotional expressivity, and perceived naturalness, as well as objective metrics that better correlate with human judgments of animation quality.

By pursuing these directions, future work can build on the foundation established in this thesis to create even more realistic, expressive, and versatile systems for identity-independent audio-visual generation.

Bibliography

[1] Garima Thakur et al. "A Robust Privacy-Preserving ECC-Based Three-Factor Authentication Scheme for Metaverse Environment". In: *Computer Communications* 211 (2023), pp. 271–285. ISSN: 0140-3664. DOI: https://doi.org/10.1016/j.comcom.2023.09.020. URL: https://www.sciencedirect.com/science/article/pii/S0140366423003304.

[2] Enes Yigitbas and Christian Kaltschmidt. *Effects of Human Avatar Representation in Virtual Reality on Inter-Brain Connection*. 2024. arXiv: 2410.21894 [cs.HC]. URL: https://arxiv.org/abs/2410.21894.

[3] Yixing Lu et al. *GAS: Generative Avatar Synthesis from a Single Image*. 2025. arXiv: 2502.06957 [cs.CV]. URL: https://arxiv.org/abs/2502.06957.

[4] S. L. Metzger, K. T. Littlejohn, A. B. Silva, et al. "A high-performance neuroprosthesis for speech decoding and avatar control". In: *Nature* 620.7976 (2023). Published correction appears in Nature. 2024 Jul;631(8021):E13. doi:10.1038/s41586-024-07735-z, pp. 1037–1046. DOI: 10.1038/s41586-023-06443-4.

[5] Neil Zeghidour et al. *SoundStream: An End-to-End Neural Audio Codec*. 2021. arXiv: 2107.03312 [cs.SD]. URL: https://arxiv.org/abs/2107.03312.

[6] Alexandre Défossez et al. *High Fidelity Neural Audio Compression*. 2022. arXiv: 2210.13438 [eess.AS]. URL: https://arxiv.org/abs/2210.13438.

[7] Hyeong-Seok Choi et al. *Neural Analysis and Synthesis: Reconstructing Speech from Self-Supervised Representations*. 2021. arXiv: 2110.14513 [cs.SD]. URL: https://arxiv.org/abs/2110.14513.

[8] Zhichao Huang, Chutong Meng, and Tom Ko. *RepCodec: A Speech Representation Codec for Speech Tokenization*. 2024. arXiv: 2309.00169 [eess.AS]. URL: https://arxiv.org/abs/2309.00169.

[9] Cheol Jun Cho et al. *Sylber: Syllabic Embedding Representation of Speech from Raw Audio*. 2025. arXiv: 2410.07168 [cs.CL]. URL: https://arxiv.org/abs/2410.07168.

[10] Shinji Maeda. "Compensatory Articulation During Speech: Evidence from the Analysis and Synthesis of Vocal-Tract Shapes Using an Articulatory Model". In: 1990. URL: https://api.semanticscholar.org/CorpusID:62009393.

[11] Gunnar Fant. *Acoustic Theory of Speech Production*. The Hague: Mouton, 1960.

[12] Tina Rebernik et al. "A review of data collection practices using electromagnetic articulography". In: *Laboratory Phonology* 12.1 (2021), p. 6. DOI: 10.5334/labphon.237.

[13] Cheol Jun Cho et al. "Evidence of Vocal Tract Articulation in Self-Supervised Learning of Speech". In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, June 2023, pp. 1–5. DOI: 10.1109/icassp49357.2023.10094711. URL: http://dx.doi.org/10.1109/ICASSP49357.2023.10094711.

[14] Prasanta Ghosh and Shrikanth Narayanan. "A generalized smoothness criterion for acoustic-to-articulatory inversion". In: *The Journal of the Acoustical Society of America* 128 (Oct. 2010), pp. 2162–72. DOI: 10.1121/1.3455847.

[15] Peter Wu et al. *Speaker-Independent Acoustic-to-Articulatory Speech Inversion*. 2023. arXiv: 2302.06774 [eess.AS]. URL: https://arxiv.org/abs/2302.06774.

[16] Peter Wu et al. *Deep Speech Synthesis from Articulatory Representations*. 2022. arXiv: 2209.06337 [eess.AS]. URL: https://arxiv.org/abs/2209.06337.

[17] Yingming Gao, Peter Birkholz, and Ya Li. "Articulatory Copy Synthesis Based on the Speech Synthesizer VocalTractLab and Convolutional Recurrent Neural Networks". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024), pp. 1845–1858. DOI: 10.1109/TASLP.2024.3372874.

[18] Rishi Jain et al. *Multimodal Segmentation for Vocal Tract Modeling*. 2024. arXiv: 2406.15754 [cs.CV]. URL: https://arxiv.org/abs/2406.15754.

[19] Cheol Jun Cho et al. "Coding Speech Through Vocal Tract Kinematics". In: *IEEE Journal of Selected Topics in Signal Processing* 18.8 (Dec. 2024), pp. 1427–1440. ISSN: 1941-0484. DOI: 10.1109/jstsp.2024.3497655. URL: http://dx.doi.org/10.1109/JSTSP.2024.3497655.

[20] Sarah L. Taylor et al. "Dynamic units of visual speech". In: *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. SCA '12. Lausanne, Switzerland: Eurographics Association, 2012, pp. 275–284. ISBN: 9783905674378.

[21] Guanzhong Tian, Yi Yuan, and Yong liu. *Audio2Face: Generating Speech/Face Animation from Single Audio with Attention-Based Bidirectional LSTM Networks*. 2019. arXiv: 1905.11142 [cs.LG]. URL: https://arxiv.org/abs/1905.11142.

[22] Daniel Cudeiro et al. "Capture, Learning, and Synthesis of 3D Speaking Styles". In: *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 10101–10111.

[23] Yingruo Fan et al. *FaceFormer: Speech-Driven 3D Facial Animation with Transformers*. 2022. arXiv: 2112.05329 [cs.CV]. URL: https://arxiv.org/abs/2112.05329.

[24] Jinbo Xing et al. *CodeTalker: Speech-Driven 3D Facial Animation with Discrete Motion Prior*. 2023. arXiv: 2301.02379 [cs.CV]. URL: https://arxiv.org/abs/2301.02379.

[25] Gihoon Kim et al. *NeRFFaceSpeech: One-shot Audio-driven 3D Talking Head Synthesis via Generative Prior*. 2024. arXiv: 2405.05749 [cs.CV]. URL: https://arxiv.org/abs/2405.05749.

[26] Tianye Li et al. "Learning a model of facial shape and expression from 4D scans". In: *ACM Trans. Graph.* 36.6 (Nov. 2017). ISSN: 0730-0301. DOI: 10.1145/3130800.3130813. URL: https://doi.org/10.1145/3130800.3130813.

[27] Aggelina Chatziagapi et al. *AV-Flow: Transforming Text to Audio-Visual Human-like Interactions*. 2025. arXiv: 2502.13133 [cs.CV]. URL: https://arxiv.org/abs/2502.13133.

[28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. arXiv: 2006.11239 [cs.LG]. URL: https://arxiv.org/abs/2006.11239.

[29] Linrui Tian et al. *EMO: Emote Portrait Alive – Generating Expressive Portrait Videos with Audio2Video Diffusion Model under Weak Conditions*. 2024. arXiv: 2402.17485 [cs.CV]. URL: https://arxiv.org/abs/2402.17485.

[30] Yaron Lipman et al. *Flow Matching for Generative Modeling*. 2023. arXiv: 2210.02747 [cs.LG]. URL: https://arxiv.org/abs/2210.02747.

[31] Xingchao Liu, Chengyue Gong, and Qiang Liu. *Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow*. 2022. arXiv: 2209.03003 [cs.LG]. URL: https://arxiv.org/abs/2209.03003.

[32] Shivam Mehta et al. *Matcha-TTS: A fast TTS architecture with conditional flow matching*. 2024. arXiv: 2309.03199 [eess.AS]. URL: https://arxiv.org/abs/2309.03199.

[33] Matthew Le et al. *Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale*. 2023. arXiv: 2306.15687 [eess.AS]. URL: https://arxiv.org/abs/2306.15687.

[34] S.R. Livingstone and F.A. Russo. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English". In: *PLoS ONE* 13.5 (2018), e0196391. DOI: 10.1371/journal.pone.0196391.

[35] Andrew G. Howard et al. *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications*. 2017. arXiv: 1704.04861 [cs.CV]. URL: https://arxiv.org/abs/1704.04861.

[36] Emily B. Prince, Katherine B. Martin, and Daniel S. Messinger. "Facial Action Coding System". In: 2015. URL: https://api.semanticscholar.org/CorpusID:14048422.

[37] George Retsinas et al. *3D Facial Expressions through Analysis-by-Neural-Synthesis*. 2024. arXiv: 2404.04104 [cs.CV]. URL: https://arxiv.org/abs/2404.04104.

[38] Michael McAuliffe et al. "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi". In: *Interspeech*. 2017.

[39] Vassil Panayotov et al. "Librispeech: An ASR corpus based on public domain audio books". In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015, pp. 5206–5210.

[40] Antti Suni, Daniel Aalto, and Martti Vainio. *Hierarchical Representation of Prosody for Statistical Speech Synthesis*. 2015. arXiv: 1510.01949 [cs.CL]. URL: https://arxiv.org/abs/1510.01949.

[41] Kyunghyun Cho et al. *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*. 2014. arXiv: 1406.1078 [cs.CL]. URL: https://arxiv.org/abs/1406.1078.

[42] Bowen Shi et al. *Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction*. 2022. arXiv: 2201.02184 [eess.AS]. URL: https://arxiv.org/abs/2201.02184.

[43] Radek Daněček et al. *Supervising 3D Talking Head Avatars with Analysis-by-Audio-Synthesis*. 2025. arXiv: 2504.13386 [cs.GR]. URL: https://arxiv.org/abs/2504.13386.