Lip-to-Speech Synthesis

A thesis submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy
in
Computer Science and Engineering

by

Rudrabha Mukhopadhyay
2018801004
radrabha.m@research.iiit.ac.in

Advisor: Dr. C.V. Jawahar

International Institute of Information Technology Hyderabad
500 032, India

June 2025

Copyright © Rudrabha Mukhopadhyay, 2025
All Rights Reserved

International Institute of Information Technology Hyderabad
Hyderabad, India

CERTIFICATE

This is to certify that work presented in this thesis titled ***Lip-to-Speech Synthesis*** by *Rudrabha* has been carried out under our supervision and is not submitted elsewhere for a degree.

_____ _____

Date Advisor: Dr. C.V. Jawahar

Acknowledgements

The completion of my Ph.D. journey has been influenced by many - it was not a path taken in isolation. This thesis is the culmination of persistent support and encouragement from a network of individuals and institutions. As I reach the end of this significant academic endeavor, I want to express my sincere gratitude to everyone who played a part in this journey. My family, teachers, friends, colleagues, and well-wishers have all contributed to the success of this study, making it a memorable and enriching experience.

I would like to express my deepest gratitude to my supervisors, Prof. C.V. Jawahar and Prof. Vinay Namboodiri, for their invaluable guidance and mentorship over the past six years. Working under their tutelage has been an immense honor, and their wisdom has been instrumental in helping me navigate through challenging research problems and life lessons.

I am profoundly grateful to my parents, Prof. Jayanta Mukhopadhyay and Dr. Jhuma Mukhopadhyay, for being my constant source of strength and support through all the highs and lows of my life.

A special thanks to Prof. Santanu Chowdhury and Dr. Manoj Singh for introducing me to the fascinating world of deep learning and research during a summer internship at CEERI Pilani in 2017. This experience was pivotal in inspiring me to pursue a PhD.

I extend my heartfelt gratitude to K R Prajwal, who has been the most significant collaborator throughout my Ph.D. journey. Since the onset of our collaborative efforts in early 2019, our partnership has been both incredibly fruitful and immensely enjoyable. The synergy between us has been remarkable, and it continues to amaze me how seamlessly and effectively we have worked together. Our combined efforts have led to numerous successes and breakthroughs, which I cherish deeply. As I look forward to future ventures, I am excited about the prospect of continuing this journey with Prajwal in our joint venture, building upon the strong foundation we have established and exploring new horizons in research. His influence on my academic growth has been profound, and I consider myself fortunate to have had such a productive and harmonious collaboration during my Ph.D.

My sincere thanks to Sindhu, whose exceptional implementation skills and attention to detail have greatly eased my work and made our collaboration a pleasure. I also extend my gratitude to Aditya Agarwal and Bipasha Sen for their assistance in tackling intriguing research challenges during my PhD.

I am grateful to Anchit Gupta, Faizan Farooq Khan, Madhav Agarwal, and R Sai Niranjan, who aided in honing my supervisory skills, allowing me to grow as a senior PhD scholar. Lately, Aparna Agarwal, Vansh Garg, and Akshat Sanghvi have joined this list, continuing to work with me at this point in time. I want to extend my thanks to all these individuals for their patience with my demands

and for gracefully handling my occasional tantrums. A special thanks to Souvik Ghosh for helping me completing experiments on one of the contributory chapters in this thesis.

My journey at IIIT Hyderabad would have been more challenging without the support of my first friend at the institute, Sangeeth Reddy. His friendship and support were crucial in navigating the coursework. I also cherish the interactions with Shubh Maheshwari, who helped me manage the academic pressures at IIITH. My appreciation also extends to Raghava, Ashish, Deepak, and Kiran for their assistance with coursework, enabling me to focus on research.

Lastly, I would like to acknowledge all the faculty, staff, and fellow students at CVIT and IIITH. Their support and assistance have been invaluable, and my PhD journey would have been incomplete without their contributions.

Rudrabha Mukhopadhyay
PhD in CSE,
Centre of Visual Information Technology (CVIT)
2018801004

Abstract

This thesis explores the development and advancement of lip-to-speech synthesis techniques, addressing the challenge of generating speech directly from visual lip movements. Unlike text-to-speech systems that rely on explicit linguistic information in the form of text tokens, lip-to-speech synthesis, aims to interpret ambiguous visual cues, presenting unique challenges in mapping similar lip shapes that can produce different sounds. Inspired by the chronological advancements in text-to-speech synthesis the research goals are broken into single-speaker lip-to-speech where a specific model is trained for each speaker with a large amount of speaker-specific data followed by multi-speaker approaches which aims to train a single model which can work for any speaker in-the-wild.

The first work presented in this thesis deals with lip-to-speech generation problem in large vocabulary in unconstrained settings albeit with a model trained for a particular speaker. In this work, a novel sequence-to-sequence model was introduced that leveraged spatio-temporal convolutional architectures to effectively capture the fine-grained temporal dynamics of lip movements and implemented a monotonic attention mechanism that more accurately aligned the visual features with corresponding speech parameters. Testing on the LRS2 dataset showed a 24% improvement in intelligibility metrics over baseline methods. In this work, a new dataset was released providing sufficient speaker-specific data with a diverse vocabulary of around $5,000$ words to support the development of accurate, speaker-specific models. While this approach showed promise, it was obviously limited to single-speaker scenarios and failed to scale effectively to sentence-level multi-speaker tasks, necessitating further research.

To address these limitations, a Variational Autoencoder-Generative Adversarial Network (VAE-GAN) architecture was developed for multi-speaker synthesis in unconstrained settings with a vocabulary exceeding $50,000$ words. This model was designed to overcome the inherent stochasticity in lip-to-speech mapping and handle multiple speaker identities without speaker-specific training, requiring only about 3 minutes of data per speaker compared to previous approach of over 600 minutes of data for a particular speaker. A key contribution of this work was the use of variational autoencoders to predict separate distributions for encoding speech content and lip movements and tieing them together with a KL-divergence loss. Additionally a Wasserstein GAN was also used to enhance the speech quality. Extensive ablation studies validated the architecture components, showing that the model produced more intelligible and realistic speech compared to existing approaches. However, significant quality limitations remained, with Automatic Speech Recognition tests revealing approximately 90% Word Error Rate, rendering it impractical for real-world applications.

Building upon these findings and acknowledging parallel advancements in lip-to-text technologies, a third approach was developed utilizing noisy text supervision. This method integrated a state-of-the-art lip-to-text network to generate intermediate text from lip movements, followed by a visual text-to-speech network that conditioned not only on the noisy text but also on the lip movements to produce speech synchronized with the original lip movements while following the text content. The key contribution in this case was a novel cross-attention mechanism in the visual TTS module that effectively aligned the visual features with the text tokens. By addressing the synchronization challenges that would arise from a simple lip-to-text followed by text-to-speech pipeline, this approach successfully maintained temporal alignment with the original visual input. Comprehensive experimentation demonstrated consistent superiority across multiple challenging benchmarks, including LRW, LRS2, and LRS3 datasets, with the model achieving notable improvements in all speech quality metrics (PESQ, STOI, ESTOI). Human evaluations further validated these findings, with particularly strong performance in intelligibility, content clarity, and synchronization accuracy. The approach scored $3.31/5$ in overall perceptual quality compared to $2.96/5$ for the baseline lip-to-text + TTS approach, and achieved a Word Error Rate of 26% compared to 36% for competing methods. Detailed analysis through various ablation studies provided deeper insights into the model's behavior. Phoneme error rate analysis revealed that the model primarily struggled with phonemes having minimal lip visibility (D, EH, K, N, and ER), an inherent limitation of visual-only approaches. Additional testing across different demographics (gender, age, race) and varied conditions (emotions, head poses) demonstrated the model's robustness while identifying specific areas for improvement. Most significantly, this approach was successfully demonstrated on an ALS patient who could mouth words but had limited vocal cord function, generating intelligible speech with a Word Error Rate of approximately 37%. This real-world application represents the first demonstration of automatic lip-to-speech synthesis for an unseen speaker in an entirely out-of-domain scenario, highlighting its transformative potential for assistive technology applications.

The research extended beyond basic lip-to-speech synthesis to explore practical applications, particularly in Audio-Visual Speech Enhancement. Two interconnected problems were investigated: audio-visual speech super-resolution and audio-visual speech denoising, both conceptualized as extensions of lip-to-speech synthesis incorporating additional noisy audio inputs. This exploration demonstrated how integrating lip movement data with traditional speech processing techniques could significantly improve speech signal quality and intelligibility in challenging environments where conventional audio-only methods fall short. The visual cues from lip movements were leveraged to reconstruct and augment low-quality audio data, achieving higher-resolution speech output and more effective noise reduction in heavily contaminated environments. Throughout all approaches, extensive experimentation was conducted to optimize model parameters. Resolution analysis revealed that 96×96 pixel inputs provided the optimal balance between performance and computational efficiency, with higher resolutions (256×256) showing decreased performance due to increased computational complexity and noise sensitivity. The studies demonstrated that temporal modeling capacity was more critical than spatial resolution for accurate lip-to-speech synthesis. All models were rigorously evaluated using multiple objective metrics

(PESQ, STOI, ESTOI, and WER) and subjective listening tests, with particular attention to their potential in assistive technology applications.

Overall, this thesis contributes significant advancements to the field of lip-to-speech synthesis across single-speaker and multi-speaker domains, progressively addressing limitations of each approach and establishing new benchmarks in this rapidly evolving field. The research demonstrates the potential for creating more accessible assistive technologies for individuals who retain lip mobility despite speech impairment, as well as applications in media enhancement, silent communication interfaces, and audio-visual processing systems.

Contents

Chapter Page

1 Introduction . . . 1
1.1 Audio-Visual Perception in Humans . . . 1
1.2 Multimodal Learning: Mirroring Human Perception . . . 2
1.2.1 Why using lip movements and speech for multimodal deep learning is beneficial? 2
1.2.1.0.1 Applications of lip-to-speech . . . 4
1.2.2 Overview of the thesis . . . 4
1.3 Audio-Visual Sync . . . 5
1.3.1 Motivation: . . . 5
1.4 Speech-to-Lip Synthesis . . . 7
1.5 Lip-to-Speech Synthesis: Contribution of this thesis . . . 9
1.5.1 Motivation for lip-to-speech synthesis . . . 10
1.6 Structure of the thesis . . . 12

2 Background . . . 14
2.1 Audio-Visual Datasets . . . 14
2.1.1 Self-supervision between Lip Movements and Speech . . . 14
2.1.2 Constrained Datasets . . . 15
2.1.3 Unconstrained Datasets . . . 15
2.2 How are faces and speech represented in our works? . . . 16
2.2.1 Representing faces . . . 16
2.2.2 Representing speech . . . 17
2.2.2.1 Capturing the content information in speech . . . 17
2.2.2.2 Capturing the style information in speech . . . 18
2.2.2.2.1 The GE2E loss used in SV2TTS . . . 19
2.3 Recognizing content solely from lip movements: Lip-to-Text . . . 19
2.3.1 The initial approaches for lip-to-text . . . 20
2.3.2 Word-level lip-to-text approaches . . . 20
2.3.3 Sentence-level lip-to-text . . . 20
2.3.4 Recent advances in lip-to-text . . . 20
2.3.5 Why is lip-to-speech important, when lip-to-text is accurate? . . . 21
2.4 Speech generation using text-to-speech algorithms . . . 21
2.4.1 Pre-deep learning TTS systems . . . 22
2.4.2 Neural Text-to-Speech Systems . . . 22
2.4.3 Vocoders in TTS . . . 23
2.5 Generating Speech Solely from Lip Movements: Lip-to-Speech Synthesis . . . 23

2.5.1 The initial works: Constrained Single Speaker Lip-to-Speech Synthesis 23
2.5.2 Unconstrained Single Speaker Lip-to-Speech Synthesis 24
2.5.3 Unconstrained Multi-Speaker Lip-to-Speech Synthesis 25
2.6 Lip-to-Speech synthesis conditioned on low-quality speech 28
2.6.1 Adding visual input for speech super-resolution and denoising 29

3 Unconstrained Single-speaker Lip-to-Speech Synthesis 31
3.1 Introduction 31
3.2 Speaker-specific Lip2Wav Dataset 33
3.2.1 Steps to collect the dataset 33
3.3 Lip2Wav: Speaker-specific lip-to-speech synthesis in unconstrained environments . . . 34
3.3.1 Problem Formulation 34
3.3.2 Speech Representation 35
3.3.3 Spatio-temporal Face Encoder 36
3.3.4 Attention-based Speech Decoder 37
3.3.5 Gradual Teacher Forcing Decay 41
3.3.6 Context Window Size 42
3.4 Benchmark Datasets and Training Details 42
3.4.1 Datasets 42
3.4.2 Training Methodology and Hyper-parameters 42
3.4.3 Speech Generation at Test Time 43
3.4.4 Metrics used to measure the quality of the generated speech from different methods 43
3.4.4.1 Short-Time Objective Intelligibility (STOI) 43
3.4.4.2 Extended Short-Time Objective Intelligibility (ESTOI) 43
3.4.4.3 Perceptual Evaluation of Speech Quality (PESQ) 44
3.4.5 Lip to Speech in Constrained Settings 44
3.4.6 Lip to Speech in Unconstrained Settings 45
3.4.7 Human Evaluation 46
3.4.7.1 Objective Human Evaluation 46
3.4.7.2 Subjective Human Evaluation 46
3.4.7.3 Qualitative results 47
3.5 Ablation Studies 48
3.5.1 Larger context window helps in disambiguation 48
3.5.2 Model is highly attentive to the mouth 49
3.5.3 Teacher Forcing vs Non-Teacher Forcing 49
3.5.4 Effect of different spatio-temporal encoders 49
3.5.5 Effect of resolution of the input face crops 50
3.6 Multi-speaker Word-level Lip-to-Speech 52
3.7 Summary 53

4 Towards Lip-to-Speech Synthesis for Arbitrary Identities in the Wild 55
4.1 Introduction 55
4.1.1 Lip2Wav fails to learn the attention alignment 55
4.1.2 Overcoming Challenges in Lip-to-Speech Synthesis 55
4.2 Essential Background 58
4.2.1 Variational Autoencoders (VAEs) 58

4.2.2 Generative Adversarial Networks (GANs) . . . 58
4.2.3 Wasserstein GAN (WGAN) . . . 58
4.2.4 Wasserstein GAN with Gradient Penalty (WGAN-GP) . . . 58
4.2.5 VAE-GAN . . . 59
4.2.6 KL Divergence for Multivariate Gaussians . . . 59
4.3 VAE-GAN architecture for multi-speaker lip-to-speech . . . 59
4.3.1 Formulating the task . . . 59
4.3.2 Fundamental issues in Previous Works . . . 59
4.3.2.1 Stochastic Nature in Lip-to-Speech Synthesis . . . 59
4.3.2.2 Scaling to Multi-speaker Lip-to-Speech . . . 60
4.3.2.3 What "Space" is Right to Learn these Ambiguous Audio-Visual Correspondences? . . . 60
4.3.3 The core idea of this chapter . . . 61
4.3.4 Key changes to melspectrogram formation . . . 61
4.3.4.1 Visual Encoder . . . 62
4.3.4.2 Speech Content Encoder . . . 62
4.3.4.3 Variational Auto Encoder Based Approach & Latent Distribution Matching . . . 63
4.3.4.4 Speaker Embedding . . . 64
4.3.4.5 Speech Decoder . . . 65
4.3.4.6 Enforcing Realism with a VAE + GAN . . . 66
4.3.4.7 Improving Voice of the Generated Speech . . . 67
4.3.5 Training Settings & Inference . . . 67
4.3.5.1 Datasets and Training Strategy . . . 68
4.3.5.2 Computation cost . . . 68
4.4 Experiments . . . 68
4.4.1 Evaluation in Constrained Settings . . . 68
4.4.1.1 Baselines . . . 68
4.4.1.2 Metrics . . . 68
4.4.1.3 Results . . . 69
4.4.2 Evaluation in Unconstrained Settings . . . 69
4.4.2.1 Baselines . . . 69
4.4.2.2 Additional metrics . . . 70
4.4.2.3 Results . . . 70
4.4.2.3.1 Lip-to-text + TTS baseline . . . 70
4.4.2.3.2 Qualitative results . . . 71
4.4.3 Human Evaluations . . . 71
4.5 Adapting to Single-Speaker Lip-to-Speech . . . 71
4.6 Ablation Studies . . . 73
4.6.1 Impact of each discriminator . . . 73
4.6.2 Importance of Local and Global Alignment . . . 74
4.6.3 Additional comparisons on LRS3 dataset . . . 74
4.6.4 Near Frontal vs. Non-frontal Videos . . . 75
4.6.5 What Kind of Visual Input is the best? . . . 75
4.6.6 Sampling strategy of VAE at train-time . . . 76
4.6.7 Auto-encoder vs. VAE . . . 76

4.6.8 Model's variation across speaker attributes . . . 77
4.6.9 Generative Strength of the proposed Lip-to-Speech Model . . . 77
4.6.10 Where does the model attend in the visual front-end? . . . 79
4.6.11 Plotting the distributions . . . 79
4.7 Summary . . . 79

5 Accurate Lip-to-Speech Synthesis for arbitrary identities in the wild . . . 82
5.1 Introduction . . . 82
5.1.1 Contributions . . . 83
5.2 Issues and challenges in existing works . . . 84
5.2.1 Learning language from speech . . . 84
5.2.2 The missing block of lip-to-speech: lip-to-text . . . 84
5.2.3 Achieving accurate lip-sync . . . 84
5.3 Essential Background . . . 84
5.3.1 Pre-trained Lip-to-Text . . . 85
5.3.2 A transformer-based TTS model: The FastSpeech2 . . . 85
5.4 The proposed approach . . . 86
5.4.1 Adopting VTP for the lip-to-speech task . . . 86
5.4.2 Visual Text-to-Speech . . . 88
5.4.2.1 Text Encoder . . . 89
5.4.2.2 Visual Encoder . . . 89
5.4.2.3 Visual-Text Attention . . . 89
5.4.2.4 Speaker Embedding . . . 90
5.4.2.5 Spectrogram Decoder . . . 90
5.4.3 Datasets and Training Settings . . . 90
5.4.3.1 Datasets . . . 90
5.4.3.2 Data pre-processing . . . 90
5.4.3.3 Model configuration and training . . . 91
5.5 Experiments . . . 91
5.5.1 Quantitative Evaluations . . . 91
5.5.1.1 Metrics . . . 91
5.5.1.2 Speech Synthesis in Constrained Settings . . . 91
5.5.1.2.1 Comparisons . . . 91
5.5.1.2.2 Results . . . 92
5.5.1.3 Speech Synthesis in Unconstrained Settings . . . 92
5.5.1.3.1 Comparisons . . . 92
5.5.1.3.2 Results . . . 93
5.5.1.3.3 Qualitative comparisons . . . 93
5.5.2 Human Evaluations . . . 93
5.6 Applications in Assistive Technology . . . 94
5.6.1 Generating Speech for a Patient suffering from ALS . . . 96
5.6.2 Ethical Considerations . . . 96
5.7 Ablation Studies . . . 97
5.7.1 Effect of different pre-trained lip-to-text models . . . 97
5.7.2 Effect of different visual representations . . . 97
5.7.3 Effect of using the speaker-embedding . . . 98

5.7.4 Using different vocoders for generating speech . . . 98
5.7.5 Word-error-rate comparison . . . 99
5.7.5.1 Comparison of phoneme-error-rates . . . 99
5.7.6 Comparison based on gender . . . 101
5.7.7 Comparison based on age . . . 101
5.7.8 Comparison based on race . . . 102
5.7.9 Comparison for Different Emotions . . . 102
5.7.10 Comparison for Different Head Poses . . . 104
5.7.11 Limitations . . . 105
5.8 Summary . . . 105

6 Adding Degraded Speech to Lip-to-Speech Synthesis setup . . . 106
6.1 Audio-Visual Speech Super-resolution . . . 106
6.1.1 Background on audio-only speech super-resolution . . . 107
6.1.2 Formally formulating the problem . . . 108
6.1.3 The Architecture . . . 108
6.1.3.1 Speech Encoder . . . 108
6.1.3.2 Visual Encoder . . . 109
6.1.3.3 Speech Decoder . . . 110
6.1.4 Speech Super-Resolution using Pseudo-Visual Stream . . . 111
6.1.4.1 Synthetic generation of frames from degraded speech . . . 111
6.1.4.2 Can the current lip synthesis models be readily used for noisy speech inputs? . . . 111
6.1.4.3 A single identity is all you need. . . . 112
6.1.4.4 Distilling lip motion knowledge for a single identity . . . 112
6.1.4.5 A Visual Noise Filter . . . 112
6.1.4.6 Training the Student Model . . . 112
6.1.4.7 Learning from a Lip Synthesis Teacher . . . 113
6.1.5 Experiments and Results . . . 114
6.1.5.1 Dataset and Training Settings . . . 114
6.1.5.1.1 Dataset . . . 114
6.1.5.1.2 Training Setup . . . 114
6.1.5.2 Results . . . 114
6.1.5.2.1 Comparison . . . 114
6.1.5.3 Quantitative Evaluation . . . 114
6.1.5.3.1 Evaluation Metrics . . . 115
6.1.5.3.2 How does the performance vary when the scale factor increases? . . . 116
6.1.5.3.3 Computation comparison . . . 117
6.1.5.4 Human Evaluation . . . 117
6.1.6 Ablation Studies . . . 118
6.1.6.1 What kind of Visual Input is the Best? . . . 118
6.1.6.2 Robustness to Noise . . . 119
6.1.6.3 Additive Mask v/s Multiplicative Mask . . . 119
6.1.6.4 Importance of the Student Network . . . 119
6.1.6.5 Model's Variation to Identity Attributes . . . 119

6.1.6.6 Comparison of Pseudo-Lip Identities . . . 120
6.1.7 Summary of the audio-visual super-resolution . . . 120
6.2 Audio-Visual Speech Denoising . . . 121
6.2.1 Audio-only Speech Enhancement . . . 122
6.2.2 Formally formulating the problem . . . 123
6.2.3 Architecture . . . 123
6.2.3.1 Audio representation . . . 123
6.2.3.2 Visual representation . . . 124
6.2.3.3 Network architecture . . . 124
6.2.3.3.1 Magnitude sub-network . . . 124
6.2.3.3.2 Phase sub-network . . . 124
6.2.4 Experiments and Results . . . 125
6.2.5 Dataset . . . 125
6.2.5.1 Experimental setup . . . 125
6.2.5.2 Evaluation . . . 125
6.2.5.3 Results . . . 125
6.2.5.3.1 Robustness to Various Noise Levels . . . 126
6.2.5.3.2 Robustness to Unseen Noises . . . 126
6.2.5.3.3 Robustness to Unseen Speakers . . . 127
6.2.6 Summary of Audio-Visual Speech Denoising . . . 127

7 Conclusion . . . 128
7.1 Limitations and future research directions . . . 131
7.1.1 Frontal faces dependency . . . 131
7.1.2 Language specificity . . . 131
7.1.3 Model Size and Efficiency . . . 132
7.1.4 Inherent Challenges in Lip Reading . . . 132
7.1.5 Voice Attributes Optimization . . . 133
7.1.6 Ethical concerns . . . 133

Bibliography . . . 135

List of Figures

Figure Page

1.1 The multimodal ecosystem of audiovisual speech processing. This diagram illustrates the bidirectional relationship between visual information (lip movements) and auditory signals (speech waveforms). At its center, audiovisual synchronization serves as the foundational mechanism connecting these modalities. The diagram showcases how information from one modality can generate or augment another: lip movements can be synthesized into speech with varying degrees of complexity (progressing from constrained single-speaker to unconstrained multi-speaker systems), while speech audio can drive visual representations through both speaker-specific and speaker-agnostic lip synchronization, as well as broader head movement generation. This circular architecture highlights how advancements in cross-modal synthesis enable increasingly natural human-computer interaction systems, with each approach building upon shared principles of temporal alignment and modality translation. 3

1.2 In this Figure, we show the different lip shapes (visemes) that are associated with particular phonemes. Researchers have long envisioned using this association to solve several challenging tasks. 5

1.3 SyncNet-like models are trained by creating positive and negative audio-visual pairs and are trained with a contrastive learning strategy. The three most popular models improved the loss function incrementally, improving the detection accuracy of sync between audio and video pairs. 6

1.4 Detecting active speakers in a video is a major application of SyncNet-like networks. . 6

1.5 The general idea for the speaker-specific talking head generation works are presented in this figure. 8

1.6 The different training strategies of seminal speaker-agnostic lip sync generation works are presented in this Figure. Each method made incremental improvements, leading to improved generation quality. 8

1.7 Networks designed to predict head motion from audio commonly employ intermediate representations to capture facial nuances. These models typically use a sequential architecture to forecast various forms of face representations from audio, such as face landmarks, learnable key points, and 3D parameters. Subsequently, a separate rendering network takes a single RGB frame and morphs it in accordance with the predicted intermediate representations, enabling more accurate and detailed simulations of head movements. 9

1.8 In recent years, the field of lip-to-speech synthesis has undergone remarkable advancements, evolving from its initial focus on constrained single-speaker scenarios to more complex and realistic settings. The journey commenced with studies centered on a single speaker under controlled conditions, laying the foundational framework for the domain. Subsequently, research efforts shifted to unconstrained single-speaker models, broadening the applicability and robustness of these systems by accounting for variations in speech and facial expressions. Most recently, the frontier has extended to unconstrained multi-speaker lip-to-speech synthesis, representing a quantum leap in complexity and real-world relevance. This evolution underscores the significant strides the field has made, moving closer to robust and versatile applications that can handle various scenarios. 11

2.1 In the figure, we present a comprehensive overview of the primary training strategies employed in lip-to-speech networks over time. Early models addressing constrained lip-to-speech synthesis typically utilized a standard encoder-decoder architecture. This has since been improved by sequence-to-sequence learning models, which often leverage transformer architectures or other advanced sequence-to-sequence techniques. Additionally, some recent works have ventured into mapping lip movements and speech to interrelated distributions, aiming to achieve a more accurate and nuanced correspondence between the two modalities. 26

2.2 A visual representation of the general idea utilized by all the audio-visual denoising works. We also depict the use of a synthetic visual stream generated from the noisy speech to replace the real visual stream, allowing the standard audio-visual models to extend to audio-only settings where a visual stream is not naturally present. 30

3.1 A sequence-to-sequence architecture named "Lip2Wav" is proposed for accurate speech generation from silent lip videos in unconstrained settings for the first time. The text in the bubble is manually transcribed and is shown for presentation purposes. 32

3.2 Lip2Wav model for lip-to-speech synthesis. The spatio-temporal encoder is a stack of 3D convolutions to extract the sequence of lip movements. This is followed by a decoder adapted from [51] for high-quality speech generation. The decoder is conditioned on the face image features from the encoder and generates the melspectrogram in an autoregressive fashion. 35

3.3 A video is presented in this link: `https://youtu.be/8VnEHcRa2l4`. This video contains qualitative results and comparisons from the proposed proposed Lip2Wav model. This image is presented as the thumbnail for this video. 48

3.4 The activations of the penultimate layer of the face encoder and the attention alignment from the decoder are plotted. It is observed that the face encoder is highly attentive towards the mouth region. 50

3.5 The decoder alignment curve illustrates that the model is generating speech by strongly conditioning on the corresponding lip movements. 51

3.6 Lip2Wav model for the multi-speaker lip-to-speech synthesis scenario. A speaker embedding is added to the network to provide additional target voice information to the network. 53

4.1 The problem of generating speech from silent lip videos for any speaker in the wild is addressed in this work. In previous works, training was conducted either on large amounts of data of isolated speakers or in laboratory settings with a limited vocabulary. Conversely, in this approach, speech can be generated for the lip movements of arbitrary identities in any voice without additional speaker-specific fine-tuning. A new VAE-GAN approach is introduced, which allows strong audio-visual associations to be learned despite the ambiguous nature of the task. 56

4.2 Attention alignment plots at various training stages indicating that Lip2Wav fails to learn temporal attention in the unconstrained multi-speaker setting. 56

4.3 A novel VAE-GAN architecture is proposed for this task. Unlike previous approaches that enforce a one-to-one mapping between lip and speech sequences, this model addresses the task's ambiguities by mapping the speech content (ASR representations) and lip sequence to similar distributions. A decoder then generates realistic speech outputs from this latent space. Additional discriminators are used to enable high-fidelity generation in unconstrained settings. 60

4.4 In addition to using a global KL-divergence loss to tie the lip and speech content distributions, these distributions are also enforced to be temporally aligned by minimizing a local KL-divergence loss on random smaller time segments. The intuition is that lips and speech are locally aligned in time in the form of visemes and phonemes. 65

4.5 A video is presented in this link: `https://youtu.be/iYehW3sd33k`. This video contains qualitative results and comparisons from the proposed VAE-GAN architecture. This image is presented as the thumbnail for this video. 72

4.6 Fine-tuning the proposed pre-trained multi-speaker model consistently outperforms the current best single-speaker model (FDSD lower is better) in the low data regime. . . . 74

4.7 The average number of unique speech outputs generated by the model for each input lip video is plotted. This "generative strength" [152] is shown at different stages of training. It can be observed that as the training progresses, the model captures more variations in the latent space, indicating an increase in the diversity of generated outputs over time. . 78

4.8 Activation maps of the visual encoder shows that the model strongly attends to the lip region while generating speech, despite variations in head pose and lip location. 79

4.9 TSNE plot of multiple instances of different words from both content and lip distributions, including homophones. The plot shows that content and lip distributions for particular words are close to each other. Additionally, homophones like "Million" and "Billion" lie close to each other, demonstrating the approach's effectiveness. 81

5.1 Overview of the proposed multi-speaker lip-to-speech system. Instead of learning a language model directly from raw speech, which provides only weak supervision due to acoustic variability (voice, accents, prosody), we leverage recent lip-to-text models to obtain noisy text transcriptions and condition a visual TTS network on both the text and the lip video. 83

5.2 An overview of the proposed approach is presented. Visual features and text predictions are first extracted from a pre-trained lip-to-text network. Speech outputs that synchronize with the silent video input are then generated using a visual text-to-speech (TTS) model. The visual and textual (in the form of phonemes) inputs are encoded and aligned in time by the visual TTS using scaled dot-product attention. For each query video time-step, the phoneme to be uttered at that time is retrieved using this attention mechanism. After the addition of the speaker identity embedding, these are upsampled and decoded into melspectrograms. Finally, the melspectrograms are converted into natural waveforms using a pre-trained vocoder. 86
5.3 The video-text alignment from the scaled dot product attention step of the model is visualized. It is observed that the model learns a strong monotonic near-diagonal attention, as expected. 94
5.4 A video is presented in this link: `https://youtu.be/6WNSazF9vyQ`. This video contains qualitative results and comparisons from the model proposed in this chapter. This image is presented as the thumbnail for this video. 95
5.5 The model is demonstrated on an ALS patient who cannot voice words but can mouth them. The speech corresponding to the silent lip movements can be generated. Lip-to-Speech can thus be a cheap and non-invasive method to assist someone who has lost their voice. 97
5.6 The top-5 wrongly predicted phonemes by the proposed lip-to-speech network are plotted in this Figure. 100
5.7 The wrongly predicted phonemes are produced without moving the lips, and thus, both the lip-to-text and lip-to-speech models struggle to predict these successfully. 101
5.8 Multiple emotions are selected from the MEAD dataset for a particular speaker to evaluate lip-to-speech synthesis using the proposed model. 103
5.9 Multiple camera positions are selected from the MEAD dataset for a particular speaker to evaluate lip-to-speech synthesis using the proposed model. The camera positions mimic different head poses in real world situations. 104

6.1 The proposed audio-visual network for speech super-resolution at large scale factors ($8\times$ and $16\times$) is illustrated. Three major components are comprised in the SR model: (i) visual encoder, (ii) speech encoder, and (iii) speech decoder. A sequence of frames is ingested by the visual encoder and processed, and visual embeddings are generated. The speech encoder takes the spectrogram representation from the linearly upsampled speech signal to create speech embeddings. These learned visual and speech embeddings are then fused and subsequently processed by the speech decoder. A residual mask is output by the network, which is added to the input spectrogram to generate realistic, high-quality (16kHz) speech signals. 109
6.2 The applicability of the proposed SR network is demonstrated by synthesizing the lip movements in cases where the visual stream is absent. A student-teacher network is set up to generate the visual stream from the LR speech input synthetically. The student model is trained to imitate the outputs from the pre-trained teacher model (Wav2Lip [27]), which ingests the HR speech and a static identity to produce accurate lip movements. . 113
6.3 A video is presented in this link: `https://youtu.be/bc0ZsTmhLM0`. This video contains qualitative results and comparisons from the Audio-Visual Speech Super-resolution model proposed in this chapter. This image is presented as the thumbnail for this video. 115

6.4 (a) Spectrograms of the ground-truth (GT), linearly upsampled speech, and the proposed predicted speech. It can be observed that the proposed network can reconstruct the LR speech, which is close to the GT speech, even at large-scale factors. (b) Performance comparison (metric: PESQ) at different scale factors. At higher scale factors, the gap in the performance of "audio-only" and "audio-visual" methods emphasizes the importance of the visual stream at larger scales. 117
6.5 Activation maps of the visual encoder for different identities. Although the proposed model is highly attentive to the lip region, the contributions from other facial areas, such as eyes and cheeks, are also noteworthy. 118
6.6 Schematic representation of lip-to-speech Synthesis with integration of noisy speech Input, depicting the process of converting visual lip movement data into synthesized speech while incorporating an additional noisy speech signal for enhanced realism and robustness. 121
6.7 The enhancement model ingests the noisy spectrogram along with the lip movements and outputs a mask for clean speech. 123

List of Tables

Table Page

2.1 Comparison of publicly available audio-visual datasets. The table presents key characteristics of each dataset, including the total duration of data in hours, the number of unique speakers, vocabulary size, presence of in-the-wild videos, availability of text transcripts, and average video dimensions. This comparison provides an overview of the scope and content of various datasets used in audio-visual research. 16

3.1 The Lip2Wav dataset is the first large-scale dataset tailored towards acting as a reliable benchmark for single-speaker lip-to-speech synthesis. 33

3.2 Structure of the Spatio-Temporal Encoder of the Lip2Wav Model 36

3.3 Decoder Components and Their Specifications 39

3.4 Objective speech quality, intelligibility and WER scores for the GRID dataset unseen test split. 44

3.5 Objective speech quality, intelligibility and WER scores for the TCD-TIMIT dataset unseen test split. 44

3.6 In unconstrained single-speaker settings, Lip2Wav model achieves almost $4\times$ more intelligible speech than the previous methods. 45

3.7 Objective Human evaluation results. The participants manually identified the percentage of (A) Mispronunciations, (B) Word skips and (C) Homophene-based errors in the test samples. 46

3.8 Mean human evaluation scores based on speech quality and intelligibility for various approaches for lip to speech. MTT denotes "manually-transcribed text". The penultimate row simulates the best possible case of an automatic lip-to-text followed by a state-of-the-art text-to-speech system. In this case, the drop in naturalness score illustrates the loss in speech style and prosody. 47

3.9 Larger context information consistently results in more accurate speech generation. The window size is limited to 3 seconds due to memory constraints. 49

3.10 Blurring the mouth region drastically affects the generated speech compared to blurring the top half of the face. 49

3.11 Gradually decaying the teacher forcing enables the model to generalize to unseen vocabulary by forcing it to look at the visual input and not just predict from the previously uttered speech. 50

3.12 Lip2Wav employs a 3D-CNN encoder to capture the spatio-temporal visual information and is the superior choice over the other alternatives. 50

3.13 Analysis of different input resolutions. It is to be noted that, in fact, 128×128 shows marginally better performance. However, there is a drop in performance as the resolution is increased to 256×256 52
3.14 Objective speech quality and intelligibility scores on the LRW dataset. WER is also calculated after using an ASR on the generated speech. Our model outperforms the baseline method proposed in [30] without any text-level supervision. The speech metrics are not applicable for [30, 136] as they are lip-to-text works. 54

4.1 Major differences between the proposed approach and the existing approaches. As shown in this table, this work deals with the most challenging task in this space. 57
4.2 Architectural details for the visual encoder. Each row represents a layer in the network, showing how the input size is transformed into the output size. Notably, the time upsampler layer upsamples the time dimension from N to $4N$, denoted as T in the presented setup 63
4.3 The architecture for the speech content encoder 63
4.4 Architecture of the speech decoder 66
4.5 Quantitative results on the constrained GRID [60] and TCD-TIMIT [61] datasets. . . . 69
4.6 All models are pre-trained on the LRW dataset and then trained on LRS2. All the comparative methods are outperformed, especially on the challenging LRS2 data, which contains unseen speakers, words, poses and a large vocabulary. 69
4.7 (A) Intelligibility (is the speech meaningful?), (B) Perceptual Quality, (C) Sync Accuracy, (D) Voice Match. The proposed approach outputs meaningful, intelligible speech that matches lip movements and voice of the target person. 73
4.8 The discriminators enforce the proposed model to produce meaningful and realistic speech outputs. 75
4.9 Optimizing both the global and local KL-divergence loss improves the overall quality of the results. 75
4.10 Quantitative comparison on the LRS3 dataset [29]. It can be seen that all competitive methods are outperformed, even in the very different setting of LRS3, which contains unseen speakers, words, and a large number of profile views. Note that the model is not fine-tuned on the LRS3 dataset. 75
4.11 Comparison of performance between frontal and non-frontal views. Similar to other lip-reading models [86], a drop in performance is observed in non-frontal views, indicating room for improvement in handling such cases. 76
4.12 Feeding the full face crop produces the best results. 76
4.13 Sampling solely from the speech distribution during training enables the decoder to learn to generate realistic, accurate outputs. 76
4.14 Using a VAE enables the model to generate meaningful, high-fidelity speech outputs. . 77
4.15 There is no distinctive variation of performance across the genders of the speakers. . . 77

5.1 This table outlines the architecture of the Conv3D feature extractor, detailing each layer's configuration, including the type, kernel size, stride, number of filters, presence of residual connections, and the input and output sizes. 88

5.2 The state-of-the-art methods are compared on several standard multi-speaker benchmarks using standard metrics. The generated outputs from the model are found to be the most natural (PESQ), the most accurate (STOI, ESTOI), and in perfect sync with the video input (LSE-C, LSE-D) in the in-the-wild videos of LRW [30], LRS2 [28], and LRS3 [29]. . . . 92
5.3 (A) Intelligibility, (B) Content clarity, (C) Sync Accuracy, (D) Overall perceptual quality. The model produces natural and realistic speech outputs that is largely preferred by the users in comparison to other approaches. . . . 96
5.4 A comparison of using generated text from different lip-to-text networks in the pipeline is presented. The WER of the lip reading model (L2T-WER) on the LRS2 test set is also reported as a reference. . . . 98
5.5 The effect of using different visual representations for training the Visual TTS module is presented in this table. . . . 98
5.6 Identity network ablation on LRS2 test set. . . . 98
5.7 Using a vocoder network during inference to generate speech produces better quality outputs. . . . 99
5.8 Comparison of WER (After ASR) for the proposed lip-to-text + Visual TTS model and LipVoicer models using the LRS3 test set. . . . 99
5.9 Comparison of PER for the proposed lip-to-text + Visual TTS model and LipVoicer models using the LRS3 test set. . . . 100
5.10 Gender bias test results on the LRS3 test set . . . 101
5.11 Age bias test results on the LRS3 test set . . . 102
5.12 Race bias test results on the LRS3 test set. . . . 102
5.13 Emotion comparison results using the MEAD dataset . . . 103
5.14 Head pose comparison results using the MEAD dataset . . . 104

6.1 Details of the speech encoder. . . . 110
6.2 Details of the visual encoder. . . . 110
6.3 Details of the speech decoder. . . . 111
6.4 Quantitative comparison of different approaches at scale-factors of $4\times$, $8\times$ and $16\times$. The proposed method outperforms the existing audio-only approaches by a large margin, illustrating the benefits from the visual stream. . . . 116
6.5 Quantitative comparison of different approaches at scale factors of $8\times$ and $16\times$ on LRS2 [28] dataset. . . . 116
6.6 Comparison of the model size (in million parameters) and the inference time (in seconds). The proposed "audio-visual" model has parameters similar to most of the "audio-only" approaches, with a very low inference time. . . . 117
6.7 Mean opinion scores of different methods based on: (i) Quality and (ii) Intelligibility. The proposed method generates plausible speech outputs with higher perceptual satisfaction. . . . 118
6.8 Feeding full face to the visual encoder achieves better performance. . . . 118
6.9 The propsoed model is robust to noisy inputs and generates plausible speech outputs. . . . 119
6.10 Addition mask achieves better performance compared to multiplication masks. . . . 119
6.11 The student network yields the best performance compared to other alternatives. . . . 120
6.12 Effect of the identity attributes such as gender and age on model's performance. . . . 120
6.13 The proposed pseudo-visual model is invariant to pseudo-lip identities. . . . 120

6.14 Quantitative comparison of different approaches. The first section contains clean speech from LRS3 [29] test set mixed with VGGSound [184] noises at different SNR levels. In the second section, the performance on "unseen noises" is specifically evaluated by mixing the LRS3 [29] test set audios with the QUT [184] city-street noises at different noise levels. Finally, in the third section, evaluation is specifically conducted on "unseen speakers" by mixing the speeches of the unseen LRS2 [28] test set speakers with VGGSound [184] noises. The proposed method outperforms the audio-only approaches in all three sections and is comparable ($< 3\%$ difference) to the real visual-stream method. 126

List of Related Publications

Thesis publications

[P1] K. R. Prajwal*, **Rudrabha Mukhopadhyay***, Vinay Namboodiri, and C. V. Jawahar, "Learning individual speaking styles for accurate lip to speech synthesis," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13796-13805.

[P2] Sindhu Hegde*, **Rudrabha Mukhopadhyay***, C. V. Jawahar, and Vinay Namboodiri, "Towards Accurate Lip-to-Speech Synthesis in-the-Wild," in Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 5523-5531.

[P3] Sindhu Hegde*, K. R. Prajwal*, **Rudrabha Mukhopadhyay***, Vinay Namboodiri, and C. V. Jawahar, "Lip-to-speech synthesis for arbitrary speakers in the wild," in Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 6250-6258.

[P4] **Rudrabha Mukhopadhyay***, Sindhu Hegde*, Vinay Namboodiri, and C. V. Jawahar, "Audio-Visual Speech Super-Resolution," BMVC 2021. [Online].
Available: `https://www.bmvc2021-virtualconference.com/conference/papers/paper_0930.html`

[P5] **Rudrabha Mukhopadhyay***, Vinay Namboodiri, and C. V. Jawahar, "A Survey on the Interplay of Speech and Lip Movements", ACM Computing Reviews, 2024 (Under Review)

Non-thesis publications

[P6] K. R. Prajwal*, **Rudrabha Mukhopadhyay***, Jerin Philip, Abhishek Jha, Vinay Namboodiri, C. V. Jawahar, "Towards Automatic Face-to-Face Translation," in Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 1428-1436.

[P7] K. R. Prajwal*, **Rudrabha Mukhopadhyay***, Vinay Namboodiri, C. V. Jawahar, "A Lip Sync Expert is All You Need for Speech to Lip Generation in the Wild," in Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 484-492.

[P8] Nimisha Srivastava, **Rudrabha Mukhopadhyay***, K. R. Prajwal*, C. V. Jawahar, "Indicspeech: Text-to-Speech Corpus for Indian Languages," in Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 6417-6422.

[P9] Sindhu B Hegde*, K. R. Prajwal*, **Rudrabha Mukhopadhyay***, Vinay Namboodiri, C. V. Jawahar, "Visual Speech Enhancement Without a Real Visual Stream," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1926-1935.

[P10] Parul Kapoor, **Rudrabha Mukhopadhyay**, Sindhu B. Hegde, Vinay Namboodiri, C. V. Jawahar, "Towards Automatic Speech to Sign Language Generation," Interspeech, 2021.

[P11] Anchit Gupta, Faizan Farooq Khan, **Rudrabha Mukhopadhyay**, Vinay Namboodiri, C. V. Jawahar, "Intelligent Video Editing: Incorporating Modern Talking Face Generation Algorithms in a Video Editor," in Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing, 2021.

[P12] Seshadri Mazumder, **Rudrabha Mukhopadhyay**, Vinay Namboodiri, C. V. Jawahar, "Translating Sign Language Videos to Talking Faces," in Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing, 2021, pp. 1-10.

[P13] Madhav Agarwal, Anchit Gupta, **Rudrabha Mukhopadhyay***, Vinay Namboodiri, C. V. Jawahar, "Compressing Video Calls Using Synthetic Talking Heads," BMVC, 2022.

[P14] Sindhu B Hegde, **Rudrabha Mukhopadhyay**, Vinay Namboodiri, C. V. Jawahar, "Extreme-Scale Talking-Face Video Upsampling with Audio-Visual Priors," in Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 6511-6520.

[P15] Aditya Agarwal, Bipasha Sen, **Rudrabha Mukhopadhyay**, Vinay Namboodiri, C. V. Jawahar, "FaceOff: A Video-to-Video Face Swapping System," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 3495-3504.

[P16] Aditya Agarwal, Bipasha Sen, **Rudrabha Mukhopadhyay**, Vinay Namboodiri, C. V. Jawahar, "Towards MOOCs for Lipreading: Using Synthetic Talking Heads to Train Humans in Lipreading at Scale," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 2217-2226.

[P17] Madhav Agarwal, **Rudrabha Mukhopadhyay**, Vinay Namboodiri, C. V. Jawahar, "Audio-Visual Face Reenactment," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 5178-5187.

[P18] Anchit Gupta, **Rudrabha Mukhopadhyay**, Sindhu Balachandra, Faizan Farooq Khan, Vinay Namboodiri, C. V. Jawahar, "Towards Generating Ultra-High Resolution Talking-Face Videos with Lip Synchronization," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023.

[P19] Jayasree Saha, **Rudrabha Mukhopadhyay**, Aparna Agarwal, Surbahi Jain, and C.V. Jawahar, "An Approach for Speech Enhancement in Low SNR Environments using Granular Speaker Embedding", in proceedings of 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD), Bangalore, India, January 2024

[P20] S. Niranjan Ramachandran*, **Rudrabha Mukhopadhyay***, Madhav Agarwal*, C. V. Jawahar, and Vinay Namboodiri, "Understanding the Generalization of Pretrained Diffusion Models on Out-of-Distribution Data," Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024.

[P21] S. Niranjan Ramachandran, **Rudrabha Mukhopadhyay**, Bipasha Sen, Aditya Agarwal, C. V. Jawahar, and Vinay Namboodiri, "Cluster Dynamics: Unraveling Semantic Associations in Diffusion Spaces," International Joint Conference on Artificial Intelligence, IJCAI 2024 (UNDER REVIEW).

Patents

[Pat1] System and Method for Lip Syncing to a Target Speech Using a Machine Learning Model

- Provisional Patent Application Filed with the USA Patent Office, January 2021
- Applicant: International Institute of Information Technology, Hyderabad
- Contributors: Prajwal K R, **Rudrabha Mukhopadhyay**, Vinay Namboodiri, and Jawahar C V

[Pat2] System and Method for Automatically Generating Synthetic Head Videos Using a Machine Learning Model

- Provisional Patent Application Filed with the Indian Patent Office, January 2023
- Applicant: International Institute of Information Technology, Hyderabad
- Contributors: Jawahar C V, Aditya Agarwal, Bipasha Sen, **Rudrabha Mukhopadhyay**, Vinay Namboodiri

[Pat3] System and Method for Automatically Generating a Sign Language Video with an Input Speech Using a Machine Learning Model

- Provisional Patent Application Filed with the Indian Patent Office, January 2023
- Applicant: International Institute of Information Technology, Hyderabad
- Contributors: CV Jawahar, Parul Kapoor, Sindhu B Hegde, **Rudrabha Mukhopadhyay**, Vinay Namboodiri

Other presentations

[O1] **Rudrabha Mukhopadhyay***, Sindhu Hegde*, K. R. Prajwal*, Vinay Namboodiri, and C. V. Jawahar, "The Interplay of Speech and Lip Movements," public demonstration at ICPR, 2021, Milan, Italy (ONLINE).

[O2] K. R. Prajwal*, **Rudrabha Mukhopadhyay***, Vinay Namboodiri, and C. V. Jawahar, "Learning individual speaking styles for accurate lip to speech synthesis," in Vision India track, ICVGIP, 2020 (ONLINE)

[O3] K. R. Prajwal*, **Rudrabha Mukhopadhyay***, Vinay Namboodiri, and C. V. Jawahar, "The Interplay of Speech and Lip Movements," public demonstration at ECCV, 2020 (ONLINE).

[O4] K. R. Prajwal*, **Rudrabha Mukhopadhyay***, Jerin Philip, Abhishek Jha, Vinay Namboodiri, C. V. Jawahar, "Towards Automatic Face-to-Face Translation," in Vision India track, NCVPRIPG, 2019, Hubli, India.

Chapter 1

Introduction

For centuries, speech has been a fundamental aspect of human communication, social interaction, and the transmission of knowledge and culture, captivating researchers from various disciplines. Researchers have explored the complexities of speech, including its production, processing, acquisition, development, and social implications. Speech production involves multiple components of the human body, including the larynx (voice box), tongue, teeth, and nasal cavity. However, lip movements are the most direct and observable link between speech and vision, making them particularly interesting to the computer vision community, as they provide the only easily identifiable visual cue related to speech production that can be captured and analyzed by vision systems. Lip movements and speech are closely connected and work together to help us communicate better. This interesting link has caught the attention of researchers in artificial intelligence, leading to the creation of various technologies that involve both lip movements and speech. Historically, the first significant contribution is Alexander Melville Bell's "The Visible Speech Manual," [1] created in 1867 to teach lip-reading. His son, Alexander Graham Bell, and colleague David Murray later expanded upon this work in the early 20th century with "The Visible Speech Movement" [2]. These pioneering studies laid the groundwork for understanding the intricate relationship between lip movements and speech. Building on the groundwork laid by the Bells, various researchers have delved deeper into the relationship between lip movements and speech. In the early 20th century, neurosurgeon Wilder Penfield used electrical brain stimulation to identify regions responsible for speech, including lip articulation [3]. Concurrently, linguist Benjamin Lee Whorf explored how lip movements and the sounds of different languages could influence thought [4]. One of the earliest specific studies [5] on this topic came in 1929 from Utzinger et al., who highlighted the importance of lip movements in articulation and speech production.

1.1 Audio-Visual Perception in Humans

According to a study published in the Journal of Experimental Psychology - Frontiers in Developmental and Related Phenomena (JPD) [6], speech perception in infants as young as 4.5 months is influenced by sensorimotor information related to lip movements, such as those associated with chewing or sucking. Another study [7] suggests that even very young infants can detect when sounds and

lip movements do not match, even if they do not yet understand the meaning of spoken words. This study found that the amount of inattention measured was not related to age or birth weight and that there were no differences between trials, which may suggest that the awareness of the relationship between lip movements and speech sounds is innate. According to [8], the function and structure of the dorsal auditory stream play a crucial role in the visual enhancement of speech perception in noise. Lip movements can improve the specificity of phoneme representations and enhance the network connectivity of the dorsal stream, leading to improved speech perception. The influential paper "Hearing lips and Seeing Voices" [9] by Harry McGurk and John MacDonald, published in 1976, demonstrated the impact of lip movements on our hearing abilities. The McGurk effect illustrates that *humans perceive the same sound differently depending on the accompanying lip movements.* Masapollo et al. conducted a study [10] using magnetic resonance imaging (MRI) to scan speakers' mouths while they pronounced various vowel sounds. The scans showed that the lips and tongue make distinct movements for each vowel sound and that the movements of these muscles contribute significantly to the production of vowel sounds.

1.2 Multimodal Learning: Mirroring Human Perception

Multimodal deep learning has emerged as a subfield of artificial intelligence that focuses on integrating multiple modalities or types of data to improve the performance of machine learning models. One of the main goals of multimodal deep learning is to develop machine learning models that can extract and use relevant information from multiple modalities to make better decisions or predictions. This can be particularly useful in tasks where different modalities provide complementary information, such as natural language processing, image and video analysis, or speech recognition. There has been significant progress in multimodal research on several tasks that need to process and understand visual and textual information, such as visual question answering (VQA) [11, 12] and image caption generation [13]. VQA involves answering natural language questions about visual content, such as images or videos. Image caption generation creates textual descriptions of images, combining visual and language information. While this task links static images with text, another interesting area explores the dynamic relationship between visual and auditory information in human speech. The natural connection between lip movements and speech has long fascinated researchers. This correlation makes it an exciting field for artificial intelligence research. One key application is multimodal speech recognition [14], which uses both audio and lip movement information to transcribe speech. This approach is particularly useful in noisy environments where audio alone is insufficient. The strong link between lip movements and speech not only improves such practical applications but also helps us better understand and replicate human communication through artificial intelligence.

1.2.1 Why using lip movements and speech for multimodal deep learning is beneficial?

Researchers have formally characterized lip movements as a sequence of visemes, each representing the visual aspects of speech production, such as the positioning and movement of lips and mouth. In

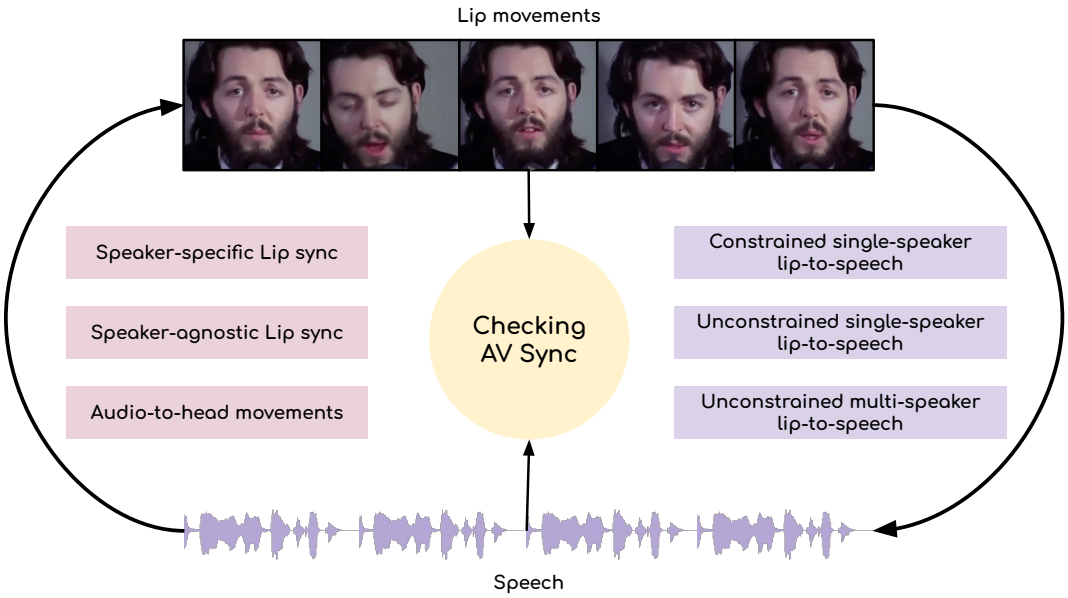

Figure 1.1 The multimodal ecosystem of audiovisual speech processing. This diagram illustrates the bidirectional relationship between visual information (lip movements) and auditory signals (speech waveforms). At its center, audiovisual synchronization serves as the foundational mechanism connecting these modalities. The diagram showcases how information from one modality can generate or augment another: lip movements can be synthesized into speech with varying degrees of complexity (progressing from constrained single-speaker to unconstrained multi-speaker systems), while speech audio can drive visual representations through both speaker-specific and speaker-agnostic lip synchronization, as well as broader head movement generation. This circular architecture highlights how advancements in cross-modal synthesis enable increasingly natural human-computer interaction systems, with each approach building upon shared principles of temporal alignment and modality translation.

contrast, the auditory aspect of speech is captured by a sequence of phonemes, which are the distinct units of sound that differentiate meaning in language.

Phonemes: A phoneme is the smallest unit of sound in a language that can distinguish meaning. Phonemes are abstract representations of sounds and do not necessarily correspond to a single, specific sound; rather, they encompass a set of sounds that are perceived as equivalent within a given language. For example, in English, the sounds represented by "p" in "pat" and "spat" are considered the same phoneme, even though they are slightly different acoustically. Phonemes are crucial in understanding the structure of languages and in the study of phonology, which is the study of how sounds are organized and used in natural languages.

Visemes: Visemes, on the other hand, are the visual equivalent of phonemes. They refer to the set of facial movements and positions (like lip shapes and tongue positions) that are used to produce speech sounds. In lip reading (or speech reading), visemes are the units that are recognized and interpreted. Since multiple phonemes can produce similar or identical facial movements, a single viseme may cor-

respond to multiple phonemes. For instance, the phonemes /p/, /b/, and /m/ might all be associated with the same viseme involving closed lips.

It is well known that the two modalities can provide complementary information streams. For example, the visemes for the phonemes “ma,” “pa,” and “ba” are identical while sounding different, whereas the phonemes “ma” and “na” sound similar but have distinct lip movements.

The research community has explored how lip movements and speech can be used to solve a variety of problems in both the fields of computer vision and speech processing. The first fundamental problem was, of course, to determine whether there is **synchronization between lip movements and speech**. This is a critical issue because if the synchronization problem cannot be solved, it would be challenging to tackle other related problems effectively. Ensuring that lip movements and speech are in sync is essential for applications such as dubbed movies, video conferencing, and virtual avatars, where any mismatch between the two can be highly noticeable and distracting.

The second most important problem that has been addressed using lip movements and speech is the **generation of lip movements from speech**, also known as speech-to-lip synthesis. This task involves predicting the corresponding lip positions for a given speech input. Lip movements and speech follow a one-to-many relationship. In other words, multiple speech sounds can correspond to the same lip position, making speech-to-lip synthesis a relatively straightforward problem to solve.

On the other hand, the reverse task of **generating speech from lip movements**, or lip-to-speech synthesis, is considered to be far more challenging. Lip-to-speech synthesis faces the challenge of one-to-many mapping between lip positions and speech sounds. This means that a single lip position can correspond to multiple speech sounds, making it more difficult to predict the corresponding speech accurately.

1.2.1.0.1 Applications of lip-to-speech Lip-to-speech synthesis is particularly motivating because it can revolutionize how we interact with technology and each other. Imagine a world where silent lip movements can be transformed into audible speech, enabling individuals who have lost their ability to speak to communicate more effectively. Lip-to-speech synthesis could also enhance the naturalness of voice assistants and create more engaging virtual reality experiences. Moreover, this technology could be used to generate speech in different languages or accents, opening up new possibilities for cross-cultural communication and language learning.

1.2.2 Overview of the thesis

This thesis addresses the challenging problem of lip-to-speech synthesis in detail. However, before delving into the specifics of the approach and contributions, it is essential to provide a comprehensive overview of the entire field that explores the interplay between lip movements and speech. This introductory chapter summarizes the major problems and research areas within this domain, setting the stage for a better understanding of the overall landscape. By presenting a broad perspective on the various aspects of lip movements and speech, the author aims to contextualize the significance of the contributions

in this thesis. Furthermore, this chapter also highlights some of the other notable contributions made by the research group, which, although not directly part of this thesis, have played a crucial role in shaping the understanding and advancing the state-of-the-art in this field. Through this introduction, the author strives to provide readers with a solid foundation and a clear picture of the challenges, opportunities, and recent developments in the area of lip movements and speech.

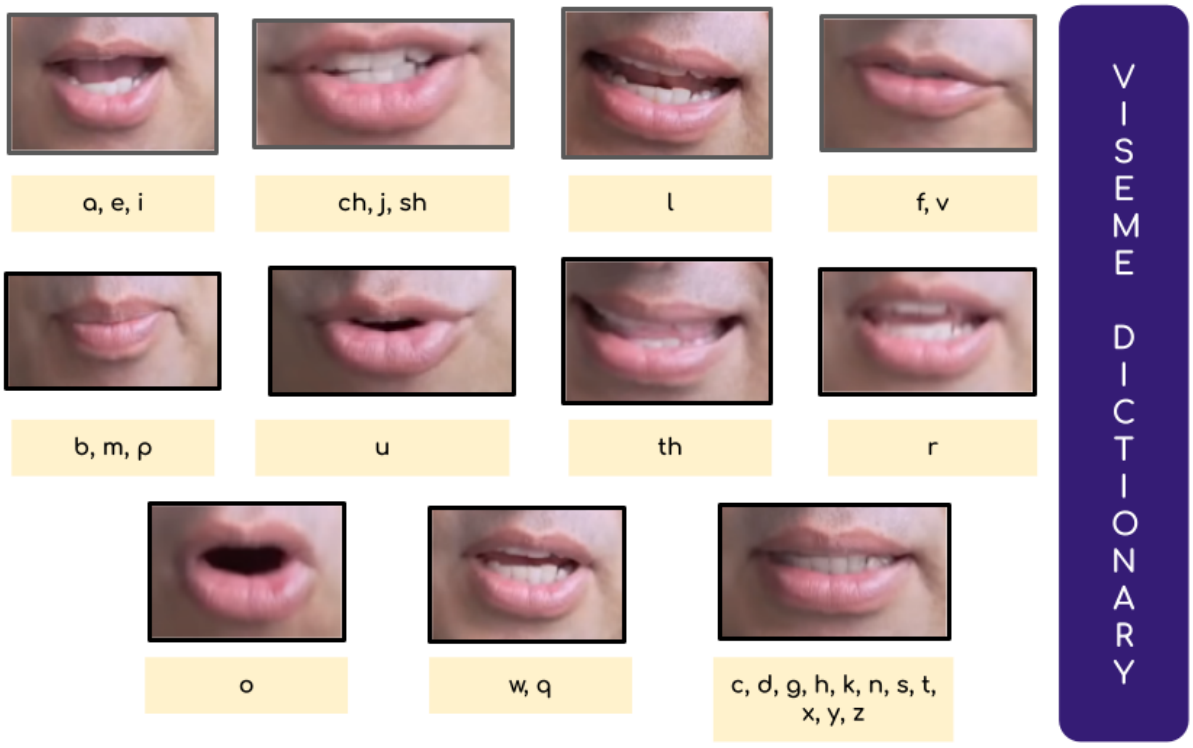

Figure 1.2 In this Figure, we show the different lip shapes (visemes) that are associated with particular phonemes. Researchers have long envisioned using this association to solve several challenging tasks.

1.3 Audio-Visual Sync

1.3.1 Motivation:

While speech and lip movements are naturally correlated, it was paramount for a neural network to learn the concept of "sync" between lip shapes (visemes) and phonemes in speech. Thus, the first works in this space explored this relationship with great interest and attempted to learn it. The general question was "can a deep learning network be trained to understand what is in sync and what is not?". In the realm of audio-visual speech technologies, understanding this synchronization between lip shapes (visemes) and speech sounds (phonemes) is a foundational aspect. This synchronization forms the core of accurately modeling tasks like lip-to-speech synthesis or speech-to-lip generation. The ability of a network to discern and maintain this sync is beneficial and essential for the success of these tasks.

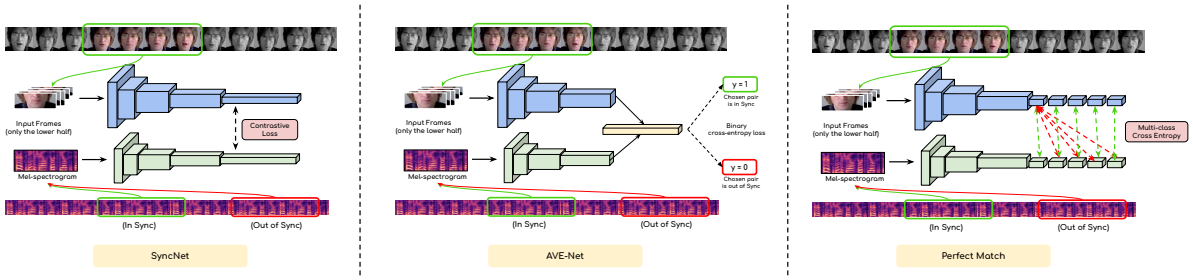

Figure 1.3 SyncNet-like models are trained by creating positive and negative audio-visual pairs and are trained with a contrastive learning strategy. The three most popular models improved the loss function incrementally, improving the detection accuracy of sync between audio and video pairs.

Contrastive learning between lip movements and speech The first work to solve this problem was SyncNet [15] in 2016, which used a standard contrastive learning approach.

SyncNet [15] helps to determine whether a given set of lip movements and an audio segment match up. It uses a technique called contrastive learning [16] to achieve this. The system has two parts: a visual encoder and an audio encoder. The visual encoder takes five consecutive frames of grayscale images and processes them to produce a 256-dimensional vector. The audio encoder takes in a short chunk of audio, represented as a set of numbers called Mel-frequency cepstral coefficients, and produces a 256-dimensional vector. The system is trained by showing pairs of audio and video segments, with some pairs being correctly matched (called positive samples) and others being mismatched (called negative samples). SyncNet uses these pairs to learn how to differentiate between a matching and mismatched pair.

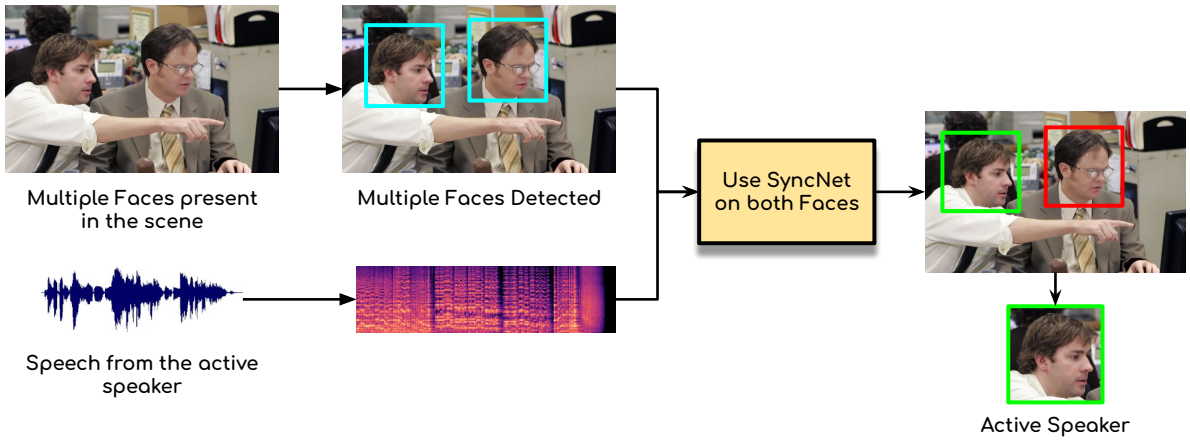

Figure 1.4 Detecting active speakers in a video is a major application of SyncNet-like networks.

During inference, any audio-visual pair is passed through the network, and the distance between the embeddings determines whether the speech and lip movements are in sync or not.

Using a better loss function A version of SyncNet similar to the original was published in [17]. The network uses a SyncNet-like architecture with two key changes: instead of using a contrastive loss

between the audio and video embeddings, the authors propose to use a classification head with a binary cross-entropy loss function, and they also use the L2-norm on the audio and video embeddings before passing them to the classifier. The network uses the same training strategy as SyncNet, sampling positive and negative pairs to classify.

Using multiple negatives The study published in [18] stands as one of the most effective methodologies for classifying audio-visual synchronization. Utilizing a dual-encoder architecture for processing audio and video streams, the network is trained on the LRS2 dataset using multi-class cross-entropy loss. A key innovation lies in the incorporation of multiple negative pairs during training. For each video segment, 'N' corresponding audio segments are sampled, only one of which is positively correlated with the video. The rest are randomly selected to serve as negative pairs. The model computes the L2 distance between the encoded audio and video features, which is then passed through a softmax classifier. This classifier produces a one-hot vector, indicating the index of the positive pair, thereby significantly enhancing the network's performance. A brief graphical description of all three networks is presented in Figure 1.3.

Practical Usage of SyncNet [15]: The high accuracy of these models helped in the collection of large datasets like VoxCeleb, VoxCeleb2 use SyncNet [15] for detecting active speakers as shown in Figure 1.4. Researchers ensured the large-scale datasets did not contain any out-of-sync segments by automatically cross-checking the lip-sync accuracy using networks like SyncNet [15], saving hundreds of human hours. Our research takes this concept further by employing networks like SyncNet as a novel approach to evaluate synchronization. In the context of our thesis, which focuses on lip-to-speech synthesis, SyncNet serves another purpose. It acts as a metric to assess the quality of synchronization. This method lets us quantitatively measure how well the generated speech aligns with lip movements. In the realm of speech-to-lip generation, SyncNet also plays a pivotal role as a discriminator, serving both as a loss function during the training process and as a metric for evaluating the final output. This dual functionality is crucial in fine-tuning speech-to-lip models to achieve high accuracy in generating lip movements corresponding to a given speech.

1.4 Speech-to-Lip Synthesis

Speech-to-lip synthesis, the generation of talking head videos from speech, is a simpler task compared to lip-to-speech synthesis due to the many-to-one relationship between speech sounds and lip positions. This characteristic simplifies the learning process for neural networks and reduces the dependency on long-term context. Research in this field has led to various approaches and methodologies for creating realistic and synchronized lip movements from speech inputs [19–24]. Early works focused on speaker-specific models, training networks on substantial amounts of data from individual speakers to capture the unique relationship between their speech and lip movements. These models excel at generating personalized lip movements but often require extensive speaker-specific data and computational resources. To address the limitations of speaker-specific models, researchers developed

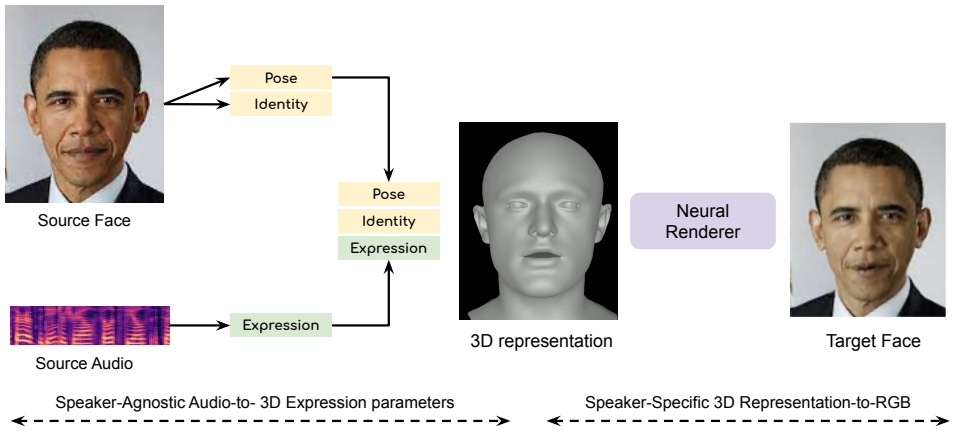

Figure 1.5 The general idea for the speaker-specific talking head generation works are presented in this figure.

speaker-agnostic approaches that can generalize to new speakers without requiring individualized data. These models, such as You-said-that? [25, 26], LipGAN, and Wav2Lip [27], are trained on diverse datasets containing talking head videos from a wide range of speakers [28–32]. While they may lack the nuanced expressiveness of speaker-specific models, they offer scalability and broader applicability. Recent advancements in speech-to-lip synthesis have focused on improving the quality and resolu-

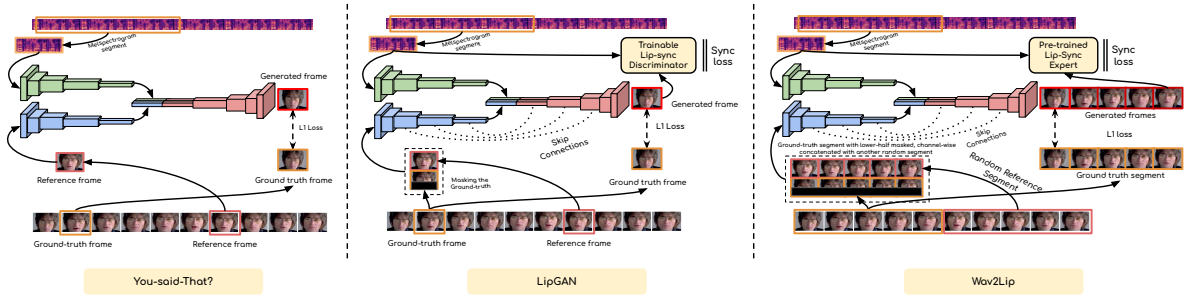

Figure 1.6 The different training strategies of seminal speaker-agnostic lip sync generation works are presented in this Figure. Each method made incremental improvements, leading to improved generation quality.

tion of generated videos, as well as incorporating head movements and facial expressions. Works like Wav2Lip-VQ [33] and VideoReTalking [34] have achieved higher fidelity lip movements and enhanced visual quality by leveraging intermediate representations and face enhancement techniques. In addition to lip movements, researchers have explored generating realistic head motions, expressions, and emotions from speech. Although more challenging due to the indirect relationship between speech and head motion, works like MakeItTalk [35], Audio2Head [36], and PiRenderer [37] have made progress in this area by predicting facial landmarks, trainable key-points, or 3D morphable model parameters from speech [38–40]. Speech-to-lip synthesis has witnessed significant advancements, with models capable

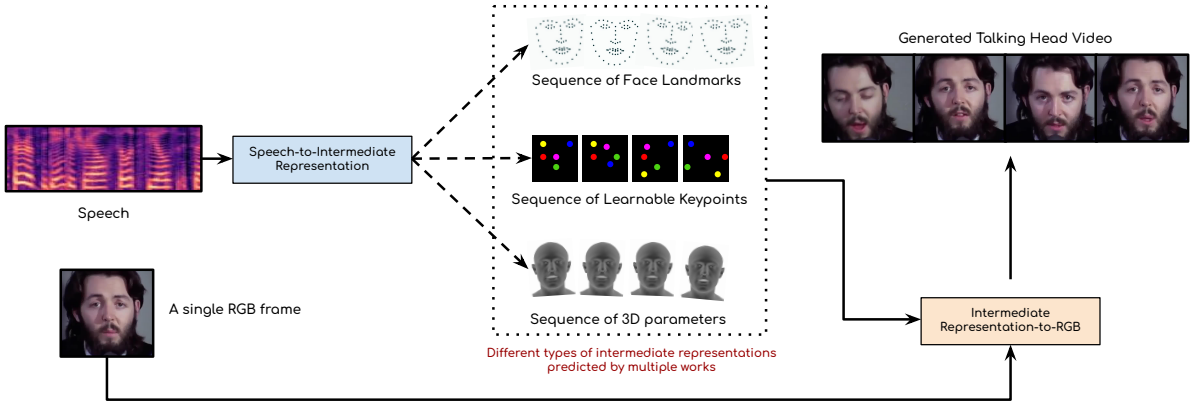

Figure 1.7 Networks designed to predict head motion from audio commonly employ intermediate representations to capture facial nuances. These models typically use a sequential architecture to forecast various forms of face representations from audio, such as face landmarks, learnable key points, and 3D parameters. Subsequently, a separate rendering network takes a single RGB frame and morphs it in accordance with the predicted intermediate representations, enabling more accurate and detailed simulations of head movements.

of generating increasingly realistic and personalized talking head videos from speech inputs. However, challenges remain in capturing the nuanced expressiveness of individual speakers and generating natural head movements that align with the speech content [41, 42].

In summary, the field of speech-to-lip synthesis has progressed from simpler to more complex challenges over time. Early works focused on speaker-specific models, often using sequence-to-sequence architectures. Later, researchers developed multi-speaker approaches to generate lip movements for a wider range of speakers. More recently, the field has tackled the even harder problem of audio-to-head movement generation, which involves synthesizing lip movements, head gestures, and facial expressions from audio. This is considered an ill-posed problem due to the many possible head movement sequences for a given audio input.

In this thesis, the focus is on the reverse problem: lip-to-speech synthesis, which generates speech from lip movements. While this task has its own challenges, inspiration is taken from the progress made in speech-to-lip synthesis, adapting those techniques and ideas to effectively address lip-to-speech synthesis. Progress in this field has guided the development and refinement of methods used in this research.

1.5 Lip-to-Speech Synthesis: Contribution of this thesis

This thesis focuses on the challenging problem of lip-to-speech synthesis, which has received less attention than the generation of talking face videos. The complexity of this task arises from the ambiguous nature of lip movements, as a single viseme can correspond to multiple phonemes. Additionally, speech generation itself presents its own set of challenges related to voice generation, maintaining prosody,

etc. Lip-to-speech synthesis is a cross-modal process where visual information from lip movements is transformed into corresponding speech signals. Unlike text-to-speech systems that work with explicit linguistic information, lip-to-speech must infer acoustic characteristics from visual cues alone, navigating the inherent ambiguity where multiple sounds share identical lip shapes. This technology has significant potential for applications ranging from assistive devices for speech-impaired individuals to enhanced communication in noisy environments, making it an important yet underexplored area within audiovisual speech processing.

1.5.1 Motivation for lip-to-speech synthesis

The human ability to interpret speech from lip movements is a natural phenomenon deeply ingrained in our communication process. This innate skill becomes particularly prominent in situations where auditory speech is absent or obscured. Lip-to-speech synthesis has numerous potential applications across various domains. In assistive technology, it can aid individuals with speech impairments or disorders to communicate more effectively. Speech therapists can use lip-to-speech synthesis to provide visual and auditory feedback, helping patients improve pronunciation and speaking skills. In telecommunications, it can fill in missing audio content during video calls or conferences with poor audio quality. The entertainment industry can benefit from lip-to-speech synthesis by enhancing dubbing in foreign-language films or creating realistic voice-overs for animated characters. In education, this technology can create accessible learning materials for students who are deaf or hard of hearing. Furthermore, lip-to-speech synthesis can contribute to the development of more natural and intuitive voice interfaces for virtual assistants or smart home devices in the field of human-computer interaction. As research in this area continues to advance, it is likely that even more innovative and impactful applications will emerge, demonstrating the value and importance of lip-to-speech synthesis in improving communication and accessibility across various sectors of society.

Early works on lip-to-speech synthesis, such as Vid2Speech [43], Improved Vid2Speech [44], Lipper [45], Lip2AudSpec [46], and others [47–49], were groundbreaking in their approach to generating speech from lip movements. However, these studies were typically conducted in controlled environments, which limited their applicability to real-world scenarios. The videos used in these works were often recorded in laboratory settings, featuring talking heads with minimal head movement, restricted vocabulary, and limited expression changes. While these controlled conditions allowed researchers to focus on the fundamental challenges of lip-to-speech synthesis, they also introduced certain limitations. The lack of natural head movement and expressions in the training data may have hindered the models' generalization of more dynamic and realistic speaking scenarios. Additionally, the restricted vocabulary used in these studies may have limited the models' capacity to handle diverse speech content encountered in real-life situations. Furthermore, the laboratory setting in which these videos were recorded may not have adequately captured the variations in lighting, camera angles, and background noise that are common in real-world environments. As a result, the models trained on such data may struggle to perform well when applied to videos captured in uncontrolled settings. Despite these limitations, early

works on lip-to-speech synthesis laid a crucial foundation for future research in this field. They demonstrated the feasibility of generating speech from lip movements and provided valuable insights into this task's challenges and potential solutions. These pioneering studies paved the way for subsequent research efforts mentioned in this thesis to address the limitations of controlled environments and develop more robust and generalized lip-to-speech synthesis models.

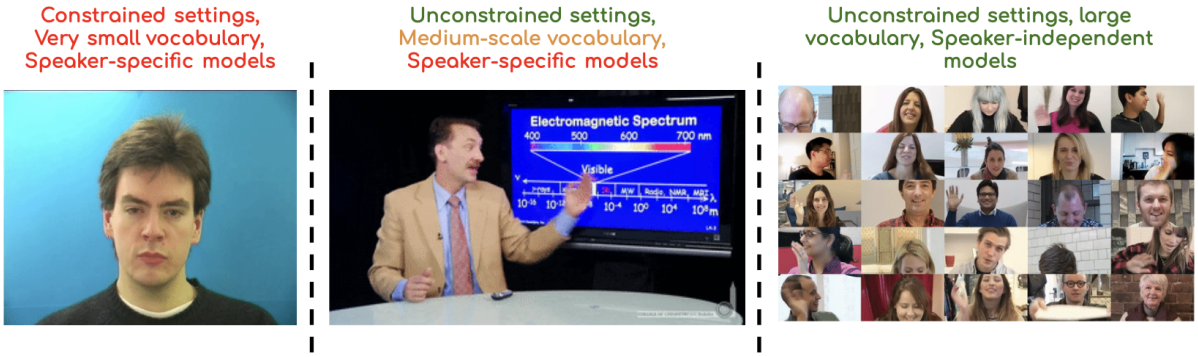

Figure 1.8 In recent years, the field of lip-to-speech synthesis has undergone remarkable advancements, evolving from its initial focus on constrained single-speaker scenarios to more complex and realistic settings. The journey commenced with studies centered on a single speaker under controlled conditions, laying the foundational framework for the domain. Subsequently, research efforts shifted to unconstrained single-speaker models, broadening the applicability and robustness of these systems by accounting for variations in speech and facial expressions. Most recently, the frontier has extended to unconstrained multi-speaker lip-to-speech synthesis, representing a quantum leap in complexity and real-world relevance. This evolution underscores the significant strides the field has made, moving closer to robust and versatile applications that can handle various scenarios.

The first significant contribution of this thesis is the development of Lip2Wav [50], a novel approach that addresses the problem of lip-to-speech synthesis in unconstrained single-speaker settings. Lip2Wav focuses on learning the speech patterns of a specific speaker by training on large amounts of in-the-wild videos. It introduces sequence-to-sequence learning for this problem and uses a modified Tacotron-2 network [51] to generate speech from lip movements. A follow-up work [52] builds upon Lip2Wav by incorporating self-supervised speech representations and acoustic variance information to improve speech synthesis quality.

The second contribution of this thesis is the development of a variational approach [53] for unconstrained multi-speaker lip-to-speech synthesis. This approach learns a mapping between lip movements and speech distributions, using an ASR like DeepSpeech2 [54] to generate content embeddings and a speaker-embedding from [55] to condition the decoder. Other notable works in this domain include VCA-GAN [56], which employs a visual context attention module, and [57], which uses a conformer-based architecture.

The third contribution of this thesis is the use of pre-trained lip-to-text models to assist in lip-to-speech synthesis [58]. This approach generates text transcriptions from silent video input using

subword-level lip reading techniques and employs Visual Transformer Pooling (VTP) embeddings to condition the subsequent text-to-speech (TTS) synthesis process. The TTS network, adapted from FastSpeech2 [59], is modified to incorporate these VTP embeddings, allowing it to be conditioned on both the text and the lip movements. This dual conditioning enables the network to generate accurate, natural, and high-quality speech output. For the first time, speech generated solely from lip movements from a neural network is shown to be practically usable in this work.

The fourth and final major contribution of this thesis is the exploration of noisy speech-assisted lip-to-speech synthesis, which extends the scope and applicability of lip-to-speech (L2S) technology. We demonstrate the effectiveness of our approach in two types of speech enhancement: noise reduction and super-resolution of speech signals. By leveraging the complementary information provided by noisy speech and lip movements, our proposed method improves the robustness and practicality of L2S technology in real-world scenarios.

In summary, this thesis makes four main contributions to the field of lip-to-speech synthesis. First, it tackles challenges in single-speaker settings with approaches like Lip2Wav. Second, it addresses multi-speaker settings using variational methods. Third, it explores the use of pre-trained lip-to-text models to improve synthesis quality. Finally, it investigates the integration of noisy speech inputs to enhance the practicality of lip-to-speech systems in real-world scenarios, demonstrating the effectiveness of this approach in both noise reduction and speech signal super-resolution. While there is still room for improvement, these contributions collectively advance the state-of-the-art in lip-to-speech synthesis and bring us closer to developing practical applications that can work well in various real-world situations.

1.6 Structure of the thesis

The thesis is systematically structured into seven chapters, each focusing on a distinct aspect of lip-to-speech synthesis. The chapters are organized as follows.

1. **Chapter 1: Introduction**
   This introductory chapter provides an overview of the research problem, its significance, and the main contributions of the thesis.

2. **Chapter 2: Background**
   This chapter provides a comprehensive background on the fields of lip-to-text, text-to-speech, and lip-to-speech synthesis. It will summarize the key contributions of this thesis in the context of lip-to-speech synthesis and highlight other contemporary work in these areas.

3. **Chapter 3: Unconstrained Single-speaker Lip-to-Speech Synthesis**
   Focusing on single-speaker scenarios, this chapter details the development and nuances of the Lip2Wav model. It delves into the technical aspects of the model, the sequence-to-sequence approach, and the evaluation of its performance.

4. **Chapter 4: Towards Lip-to-Speech Synthesis for Arbitrary Identities in the Wild**
   This chapter discusses the novel VAE-GAN-based approach for multi-speaker lip-to-speech synthesis. It examines the model architecture, its ability to handle the inherent ambiguity in lip movements, and its performance across diverse speaker profiles.

5. **Chapter 5: Accurate Lip-to-Speech Synthesis for Arbitrary Identities in the Wild**
   The chapter explores an advanced approach for multi-speaker lip-to-speech, utilizing a lip-to-text network to distill content and lip-shape information. This chapter describes the process of training a visual text-to-speech system using this distilled information.

6. **Chapter 6: Adding Degraded Speech to Lip-to-Speech Synthesis setup**
   This chapter discusses how adding degraded speech inputs in lip-to-speech can enhance speech quality, thereby extending the scope of lip-to-speech networks.

7. **Chapter 7: Conclusion**
   The final chapter summarizes the findings, discusses the research implications, and suggests directions for future work in the field of lip-to-speech synthesis.

Chapter 2

Background

This background chapter provides a comprehensive overview of three interconnected problems. The chapter begins with a comprehensive overview of the primary datasets used in the field of speech and lip movement analysis. It discusses the characteristics, strengths, and limitations of these datasets, providing context for their use in various research efforts. Following the discussion of datasets, the face and speech representations used in several contributions of this thesis are examined in detail. These representations play a crucial role in the lip-to-speech synthesis process and are fundamental to the approaches developed in this research. Then, this chapter explores the task of lip-to-text synthesis, a sister task to lip-to-speech. Next, the chapter delves into the advancements in text-to-speech technologies, tracing their evolution and current state-of-the-art approaches. This discussion provides crucial context for understanding the challenges and possibilities in speech synthesis. Finally, it thoroughly examines the main works in lip-to-speech synthesis, the central focus of this thesis. In this subsection, the progression of research in this field is chronicled, including the author's published works. These contributions are positioned within the broader context of the field, highlighting how they have advanced the state-of-the-art and addressing the challenges they've overcome.

2.1 Audio-Visual Datasets

2.1.1 Self-supervision between Lip Movements and Speech

Self-supervised learning in machine learning is leveraged to utilize the inherent structure of unlabeled data for training. In the context of lip movements and speech, one is used as the supervisory signal for the other, eliminating the need for manual labeling. A rich source of self-supervised data is provided by the abundance of talking face videos online, from lectures to interviews. These datasets, featuring synchronized lip movements and speech from thousands of individuals, are considered valuable for various tasks in this domain. In this survey, several public datasets are evaluated on key metrics like data volume, speaker diversity, and vocabulary. The datasets are categorized into 'constrained,' collected in a lab setting, and 'unconstrained,' featuring in-the-wild videos. This information and other attributes like text transcripts and video dimensions are summarized in Table 2.1 for easy comparison.

2.1.2 Constrained Datasets

Early audio-visual datasets, such as GRID [60] and TCD-TIMIT [61], were recorded in a laboratory setting with a limited vocabulary. The GRID [60] Corpus is a collection of audio and video recordings of 34 different people speaking 1000 sentences each. The recordings are of high quality and are intended to be used for research on speech perception. The corpus includes both audio and video recordings of the talkers' faces and consists of 18 male and 16 female talkers. The TCD-TIMIT [61] database is also a collection of audio-visual recordings of continuous speech. While it includes 13826 video clips in MP4 format, featuring 62 speakers reading 6913 sentences, videos from 3 professional lip speakers are considered for most of the tasks. Both of these datasets consist of frontal talking face videos of a small number of individuals without any head movement. They also include text transcripts alongside the audio-visual data to aid in lip reading. The Multi-view Emotional Audio-visual Dataset (MEAD) [62] is a comprehensive talking-face video corpus featuring 60 actors speaking with eight different emotions at three different intensity levels, with the exception of neutral. The videos were recorded simultaneously from seven different perspectives in a controlled environment to capture high-quality details of facial expressions. The corpus comprises approximately 40 hours of audio-visual clips per person and view. The constrained datasets also have a consistent illumination along with negligible pose variation.

2.1.3 Unconstrained Datasets

Unconstrained datasets are typically collected by scraping the internet for talking face videos recorded in the real world. Several datasets, such as LRW [30], LRS2 [28], VoxCeleb [32], and VoxCeleb2 [31], were collected from BBC TV programs, news reports, and interviews. Another dataset, LRS3 [29], was similarly collected from TED talks. LRW [30] includes word-level annotations, with each video tagged with a word label. LRS2 and LRS3 include sentence-level annotations with each video file. VoxCeleb and VoxCeleb2 are extensive datasets containing over $2,000$ hours of talking face videos. These datasets were collected using similar preprocessing involving detecting the active speaker and then tracking the face. The face tracks are then used to create a loose bounding box around the face, which is cropped and released, preserving pose changes, lighting, background changes, and other factors. All the videos in the datasets are resized to the same dimensions (which vary between the datasets) and resampled to 25 fps. However, a drawback of these datasets is the low resolution of the videos, which varies from 160×160 to 256×256. During the same time period, Google released the AVSpeech dataset [63], which consists of $4,700$ hours of talking face videos uploaded to YouTube. The videos vary in dimensions and frame rate and do not include text transcripts. The dataset includes HD videos at 1080p, which is useful for high-resolution video generation tasks. Similarly, a recent work introduced another HD talking face video dataset called HDTF [38], significantly smaller than AVSpeech but containing only 720p or 1080p videos. All the unconstrained datasets mentioned so far have been collected from thousands of individuals and include only small amounts of speaker-specific data. A dataset with large amounts of

speaker-specific data was released as Lip2Wav dataset [50]. This dataset contains over 10 hours of data for each speaker and useful for learning personalized traits. The datasets are summarized in Table 2.1.

Table 2.1 Comparison of publicly available audio-visual datasets. The table presents key characteristics of each dataset, including the total duration of data in hours, the number of unique speakers, vocabulary size, presence of in-the-wild videos, availability of text transcripts, and average video dimensions. This comparison provides an overview of the scope and content of various datasets used in audio-visual research.

Name	#hours	#identities	#vocab	in-the-wild?	High-Resolution?
GRID [60]	28	34	56	×	✓
TCD-TIMIT [61]	1.5	3	82	×	✓
Lip2Wav [50]	100	5	≈ 5000	✓	✓
LRW [30]	160	100+	500	✓	×
LRS2 [28]	200	500+	50000	✓	×
LRS3 [29]	450	5000+	51000	✓	×
VoxCeleb [32]	1000+	1211	-	✓	×
VoxCeleb2 [31]	2000+	5994	-	✓	×
HDTF [38]	15.8	300+	-	✓	✓
AVSpeech [63]	4000+	150000	-	✓	✓

2.2 How are faces and speech represented in our works?

In this work, face and speech signals are represented using specific approaches tailored to neural network processing. The researchers have adopted particular methods to effectively represent these complex modalities -

For facial data, the focus is primarily on the lip region, which contains the most relevant information for speech-related tasks. Lip movements are extracted from video frames and typically represented as sequences of cropped images or as feature vectors derived from these crops. This approach allows for the capture of temporal dynamics of lip movements while minimizing irrelevant facial information. In terms of speech representation, the researchers utilize melspectrograms, which offer a balance between capturing the frequency content of speech and mimicking human auditory perception. melspectrograms provide a time-frequency representation of speech that is both computationally efficient and perceptually relevant. Raw audio signals are processed to generate these spectrograms, which serve as the primary input for the speech-related neural network models. These representations form the foundation of the multimodal approaches employed in this work, enabling the models to effectively process and synthesize speech from visual lip movement data.

2.2.1 Representing faces

The video of the speaker is often presented with extensive background, additional non-speaking faces, and other types of distractions in the training datasets. Therefore, talking heads are generally

extracted from the full videos using accurate face detectors like S3FD [64], MediaPipe [65], RetinaFace [66]. For each video file, the algorithm begins by opening the video stream and setting a frame counter to zero. It then proceeds to read each frame of the video sequentially. A face detection function is applied in every frame, utilizing a face detection model to identify and locate faces within the frame. The function returns the coordinates of all detected faces. If multiple faces are detected in a frame, the algorithm employs a process to determine the largest face by calculating the area of each detected face. It selects the face with the largest area for cropping, ensuring focus on the most prominent face in the frame. This step is crucial in scenarios where multiple faces are present, as it allows the algorithm to prioritize the primary subject of the video. Once the largest face is identified, the algorithm crops the frame to the region containing this face based on the provided coordinates. The cropped frame is then saved, with the file name incorporating the frame number, allowing for easy identification and reference of specific frames. In the event that no face is detected in a frame, the algorithm skips that frame and continues to the next one. This process is repeated for all frames in the video, ensuring that only frames containing the largest, most prominent faces are extracted and stored. The cropped frames, particularly the lower half of the detected faces, are often used to extract the mouth region of the speaker, which includes the lips and jaw. This method is instrumental in tasks that require focus on the mouth for lip reading or speech analysis. Furthermore, researchers often need their networks to focus on more specific areas of the face (like eyes, nose, etc.), which can be extracted through Face-Alignment [67], generating face landmarks corresponding to different face regions. Face landmarks also serve as an excellent representation of facial structure and are utilized in several tasks where only gross facial features like head pose, gross expressions, etc., are required. The head pose is also often extracted using works like [68]. Deep Facial Features are also obtained using pretrained networks like FaceNet [69], VGGFace [70], etc., that can be used instead of RGB frames or face landmarks as input to various audio-visual neural networks. Recently, a lip-reading network [71] proposed a feature extraction network that specifically attended to the lip region in a face image. This representation is used in Chapter 5 extensively to train a state-of-the-art lip-to-speech network.

2.2.2 Representing speech

Audio signals are often directly represented using MFCC features [72], melspectrograms [73], and linear spectrograms [74]. The algorithms to calculate a magnitude-phase linear spectrogram in Algorithm 1 and a melspectrogram in Algorithm 2 are provided below. Both of these representations are used extensively in this thesis, and thus, the algorithms for calculating both the representations are presented in detail.

2.2.2.1 Capturing the content information in speech

Advanced speech recognition networks are employed to extract and process the information from speech, primarily including what is being said. Models like DeepSpeech2 [54] are specifically designed to convert raw speech signals into textual representations, effectively transcribing spoken language. In

Algorithm 1 Magnitude-Phase Representation of an Audio Signal

Require: Audio signal $y[n]$, sampling rate sr, window size N, hop length H, window function $w[n]$
Ensure: Magnitude and Phase matrices Magnitude$[k, m]$ and Phase$[k, m]$

function STFT(y, N, H, w)
  for m from 0 to $M - 1$ **do**
    for k from 0 to $K - 1$ **do**
      STFT$(y)[k, m] \leftarrow \sum_{n=0}^{N-1} y[n] \cdot w[n - mH] \cdot e^{-j\frac{2\pi kn}{N}}$
    end for
  end for
  return STFT(y)
end function
function MAGNITUDEPHASE(STFT)
  for each element $[k, m]$ in STFT **do**
    Magnitude$[k, m] \leftarrow |\text{STFT}(y)[k, m]|$
    Phase$[k, m] \leftarrow \arg(\text{STFT}(y)[k, m])$
  end for
  return Magnitude, Phase
end function
STFT$(y) \leftarrow$ STFT(y, N, H, w)
Magnitude, Phase $\leftarrow$ MAGNITUDEPHASE(STFT(y))
return Magnitude, Phase

this thesis, the authors use the DeepSpeech2 embedding for distilling content information from speech in Chapter 4. They note that other models like Wav2Vec [75] and Wav2Vec2 [76] have further advanced this field by utilizing deep neural networks that learn complex temporal hierarchies of speech signals.

2.2.2.2 Capturing the style information in speech

For distilling the style of the speaker and capturing the accent, pitch, tone, and voice, speaker embeddings are utilized heavily. Speaker embeddings condense a speaker's unique vocal attributes into a concise vector representation. Speaker embeddings serve as instrumental tools in various applications, including speaker verification [77, 78], where they facilitate precise authentication based on individual vocal traits. Furthermore, in the context of multi-speaker text-to-speech (TTS) synthesis [55, 79, 80], speaker embeddings assume a critical role in ensuring that the synthesized speech accurately conveys both the intended linguistic content and the nuanced style and expressive characteristics specific to the chosen speaker. Furthermore, speaker embeddings also find relevance in voice conversion [81], where they contribute to transforming one speaker's voice into another's while preserving original content. Consequently, the integration of speaker embeddings serves as a fundamental bridge between content and style in spoken language, contributing significantly to advancements in speaker-centric speech technologies. In this thesis, the author has utilized SV2TTS speaker embeddings [55] to provide voice and style information about speakers in multiple contributory works. These embeddings are employed in the studies discussed in Chapters 3, 4, and 5.

Algorithm 2 Compute Melspectrogram

Require: Audio signal $y[n]$, sampling rate sr, window size N, hop length H, window function $w[n]$, power p, Mel filter parameters M_{params}
Ensure: Melspectrogram Melspec$[k, m]$

function COMPUTEMELSPECTROGRAM($y, sr, N, H, w, p, M_{params}$)
    STFT$(y) \leftarrow$ STFT(y, N, H, w)
    Magnitude,$\leftarrow$ MAGNITUDEPHASE(STFT(y))
    $S[k, m] \leftarrow$ Magnitude$[k, m]^p$
    $M \leftarrow$ CONSTRUCTMELFILTERBANK(sr, M_{params})
    for each m **do**
        Melspec$[:, m] \leftarrow M \cdot S[:, m]$
    end for
    return Melspec
end function
Melspec $\leftarrow$ COMPUTEMELSPECTROGRAM$(y, sr, N, H, w, p, M_{params})$
return Melspec

2.2.2.2.1 The GE2E loss used in SV2TTS The Generalized End-to-End (GE2E) loss is utilized to train the SV2TTS framework for generating speaker embeddings. The author provides a background for this loss in the thesis, as he uses SV2TTS speaker embeddings to provide speaker information for multi-speaker lip-to-speech networks. This loss function generates and optimizes speaker embeddings - vector representations of spoken phrases or sentences. In contrast to traditional methods like triplet or contrastive loss, which consider pairs or triplets of samples, GE2E loss efficiently handles batches of utterances from multiple speakers simultaneously. It calculates loss based on the similarity between each embedding and the centroids of all speakers in the batch, with centroids being the mean of embeddings for each speaker. The essence of GE2E loss lies in its optimization objective: to minimize the distance between an embedding and its corresponding speaker's centroid while maximizing the distance from other speakers' centroids. This approach ensures that the model learns discriminative features crucial for differentiating between speakers, thereby enhancing the effectiveness of speaker verification systems and making GE2E loss a preferred choice in modern speaker recognition solutions.

2.3 Recognizing content solely from lip movements: Lip-to-Text

Lip-to-text, often considered the sister task of lip-to-speech synthesis, has seen considerable advancements in recent years. While both tasks revolve around interpreting and translating visual speech cues, lip-to-text is relatively more straightforward, primarily because it involves mapping visual information to a textual format rather than the more complex audio output. This relative simplicity has allowed for more rapid progress in the field of lip-to-text, with deep learning technologies playing a pivotal role in these advancements.

2.3.1 The initial approaches for lip-to-text

The trajectory of lip-to-text technology has been marked by a series of evolutionary steps, each building upon the advancements of its predecessors. The journey began with the implementation of relatively simple GRU/LSTM networks, exemplified by systems like LipNet [82], which made strides on constrained datasets such as GRID [60] and TCD-TIMIT [61].

2.3.2 Word-level lip-to-text approaches

As the field progressed, researchers expanded their focus to more complex and unconstrained data, like the LRW dataset [30], which led to the development of word-level models using 3D CNNs [30]. The introduction of these models marked a significant improvement, as they could classify a broader range of visual speech data with greater accuracy. The subsequent integration of 3D CNNs with LSTM/GRU models further enhanced the performance, enabling more nuanced and detailed lip-reading capabilities.

2.3.3 Sentence-level lip-to-text

The integration of Long Short-Term Memory (LSTM) [83] networks and Bahdanau attention mechanisms [84] marked a significant development in the field. LSTM networks, known for their efficiency in handling sequential data, brought a new level of depth to lip-reading models [28, 85]. These networks excel in capturing temporal dependencies, a critical factor in accurately interpreting the sequential nature of lip movements during speech. The LSTM's ability to retain information over long sequences suited it, particularly for the complexities inherent in lip reading, where understanding the context and progression of lip movements is vital. In recent years, the advent of transformer-based models has marked a significant leap in the field of lip reading and audio-visual speech recognition. These models [86, 87], leveraging the powerful architecture of transformers, have shown remarkable proficiency in deciphering speech from visual information.

2.3.4 Recent advances in lip-to-text

One of the most recent and significant developments in this field has been the introduction of AVHubert [88] from Meta. This model represents a novel approach to lip reading, employing advanced audio and visual representation learning. AVHubert first pretrains on a large dataset using video and audio tokens, effectively creating a language model. This pretrained model is then fine-tuned specifically for lip reading, resulting in unprecedented levels of accuracy and efficiency. Complementing this, research focused on subword level lip reading [71] has made another significant advancement. This approach, which trains on large volumes of data, utilizes a specialized attention-based lip encoder known as Visual Transformer Pooling (VTP). The key innovation in this method lies in its strategy to predict subwords rather than entire words. Subwords, which occupy a linguistic space between phonemes and full words, offer a more granular level of analysis and recognition. By targeting these subword units, the model achieves a finer balance between the specificity of phonemes and the broader context provided

by words. This shift towards subword prediction has resulted in a marked performance improvement, further enhancing the model's accuracy and efficiency.

In their latest work [89], Ma et al. demonstrate the effectiveness of using automatically generated transcriptions from unlabelled datasets to augment training data for audio-visual speech recognition (AV-ASR). By leveraging pre-trained ASR models to transcribe large unlabelled datasets and combining them with manually labeled data, the authors achieve state-of-the-art performance on AV-ASR tasks for the LRS2 and LRS3 datasets. Their approach notably achieves a 0.9% Word Error Rate (WER) on LRS3, marking a 30% relative improvement over previous methods while using significantly less training data than competing approaches. A more comprehensive analysis of lip reading models is presented in [90].

2.3.5 Why is lip-to-speech important, when lip-to-text is accurate?

A comparison between lip-to-text and lip-to-speech synthesis reveals that lip-to-text is generally an easier task to tackle. This is primarily due to the discrete nature of text outputs, which contrasts with the continuous, time-varying nature of speech signals. Lip-to-text systems can more easily integrate language models, providing powerful constraints and context to improve accuracy. Additionally, the discrete text outputs can benefit significantly from post-processing techniques, especially with the advent of Large Language Models (LLMs), which can refine and correct initial predictions. Text outputs are also more forgiving of minor errors, computationally more efficient to generate, and easier to evaluate and iterate upon. In contrast, lip-to-speech synthesis must capture nuances in pronunciation, intonation, and timing, making it inherently more complex. However, it's crucial to recognize that text alone does not convey all the information present in speech. Speech carries emotional nuances, tonal variations, and subtle inflections that text cannot fully capture. This is one reason why movies often resonate more deeply with audiences than books - the combination of speech and visual cues allows for a more immersive and emotionally engaging experience. The richness of information conveyed through speech underscores the importance of advancing lip-to-speech synthesis despite its challenges. By striving to generate accurate and natural speech from lip movements, we aim to create more comprehensive and emotionally resonant communication systems that can better serve human needs and experiences.

2.4 Speech generation using text-to-speech algorithms

A survey of text-to-speech (TTS) technologies is crucial for advancing lip-to-speech synthesis. While lip-to-text systems capture spoken words, they lack the nuances of human speech, such as voice quality, articulation, tone, pitch, and emotion. To address these limitations, insights from TTS can be leveraged. TTS is regarded as the standard speech generation task, offering valuable methodologies and techniques that can be adopted or adapted for lip-to-speech synthesis. Advanced TTS models, particularly those based on neural networks, have revolutionized speech synthesis, achieving unprecedented levels of nat-

uralness and expressiveness. These models encompass sophisticated techniques for handling various speech aspects, from basic articulation to complex emotional intonations.

In lip-to-speech synthesis, TTS advancements can be incorporated in several ways. At a minimum, decoders from TTS systems, which are adept at generating realistic speech, can be utilized. Additionally, techniques such as prosody modeling and emotion injection can be adapted to enhance the quality of speech generated from lip movements. The parallels between TTS and lip-to-speech, including the management of linguistic content, intonation, and expressiveness, underscore the importance of this survey. By examining TTS's evolution, a deeper understanding is gained of how to translate lip movements into coherent and natural speech, thereby enhancing the capabilities and effectiveness of lip-to-speech systems.

2.4.1 Pre-deep learning TTS systems

The initial Text-to-Speech (TTS) systems employed classical approaches to speech synthesis, laying the foundation for developing modern TTS technologies. These early systems [91, 92] aimed to generate speech from text input, focusing on achieving intelligibility and, to some extent, naturalness in the synthesized speech. The earlier approaches to speech synthesis involved using a database of sound units, where multiple variations of all possible sounds that could be uttered in a specific language were recorded. The raw speech waveforms were generated by concatenating these small speech units in appropriate order. However, the resulting speech was intelligible but not natural sounding. One of the classical approaches [93, 94] to speech synthesis was diphone-based synthesis, which involved connecting two phones (simplest speech sounds) to form a diphone. Various signal processing techniques such as Pitch Synchronous Overlap-Add (PSOLA) [95] were used in diphone-based approaches. These techniques decomposed speech into smaller segments (at the level of diphones) and then combined them to produce the expected output. The breakdown of text at the character level (graphemes or converted to phonemes) served as the information for selecting the smaller speech unit. Even the most modern TTS systems still uses similar representations for text.

2.4.2 Neural Text-to-Speech Systems

Contemporary neural TTS models typically employ a three-stage process. Initially, they transliterate input graphemes to phonemes using a phonemizer. Subsequently, these phoneme sequences are transformed into time-frequency representations, known as melspectrograms. The final stage involves generating raw speech waveforms from these melspectrograms. These models [51, 96–98] harness the power of deep learning, utilizing convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention mechanisms to learn the complex mapping between text and speech. Recent literature has been abundant with innovative architectures and techniques aimed at enhancing neural TTS models. Notable examples include Tacotron 2 [51], which introduced an attention-based sequence-to-sequence model for generating melspectrograms from text. FastSpeech [59] proposed a non-autoregressive ap-

proach using transformers for parallel melspectrogram generation, substantially accelerating inference time.

2.4.3 Vocoders in TTS

In the evolution of Text-to-Speech (TTS) systems, vocoders play a crucial role in converting melspectrograms into speech. These tools have undergone significant improvements over time. WaveNet [99], developed by DeepMind, marked a major breakthrough with its complex system for generating highly natural speech. Google's WaveRNN [100] followed, offering a simpler and faster approach while maintaining good quality. MelGAN [101] introduced a new direction using Generative Adversarial Networks (GANs), sacrificing some quality for increased speed, which is particularly useful for real-time applications. HiFi-GAN [102] built upon MelGAN's foundation, focusing on high-quality speech production while balancing speed, and excelling in capturing both fine details and overall speech characteristics. The most recent advancement, BigVGAN [103], offers versatility in handling diverse voices and speech styles, with further improvements in naturalness and clarity. Each of these vocoders has contributed uniquely to the field of TTS, with some prioritizing speech realism, others computational efficiency, and some focusing on adaptability to various speaking styles.

2.5 Generating Speech Solely from Lip Movements: Lip-to-Speech Synthesis

Lip-to-speech synthesis is explored in this thesis, a relatively new field where speech is generated from lip movements. The first papers on this topic emerged only in 2017, marking it as a recent area of research. Significant progress has been observed since then, with advancements moving from simple lab-based models to complex systems capable of handling multiple speakers in real-world settings. A comprehensive survey of these developments is presented, placing different works into perspective and highlighting the rapid evolution of the field. This overview includes contributions made as part of this thesis, alongside other significant works in the area. The challenges addressed, such as working with different speakers and varied real-world conditions, are examined, illuminating the current state and future potential of lip-to-speech synthesis.

2.5.1 The initial works: Constrained Single Speaker Lip-to-Speech Synthesis

CNN-based Encoder-Decoders In constrained single-speaker lip-to-speech, (1) The networks were trained on highly constrained datasets. (2) The networks consisted of standard CNN-based encoder-decoder architectures, which are unsuitable for speech generation in the wild. The first work [43] for this problem was proposed in 2017. The network consisted of a simple 2D CNN-based encoder-decoder architecture that required video frames (containing lip movements) as input and generated low-level LPC features, which were then used to create raw waveforms. The networks were trained on

the GRID dataset [60] that consisted of videos captured in a laboratory environment. Vid2Speech [43] was improved in a follow-up work [44]. In this work, the authors replaced the LPC features with high dimensional melspectrograms, which improved the speech generation quality. The authors also proposed using optical flow as an additional input to the encoder. Ephrat et al. [44] also train on a second speaker-specific dataset, TCD-TIMIT [61], which proves to be slightly more challenging than the GRID dataset.

Adding multiple views of the speaker A very similar approach was also taken by a different work called Lipper [45]. The authors of Lipper used a multiview dataset to train, which was also collected in a laboratory setting. The multiview dataset contained multiple views of 4 speakers speaking different phrases. In this work, all the views were fed in unison to the model, which uses a multi-class classifier to decide the best view required for a particular video. The features from the best view are then passed on to a bidirectional GRU-based model, which decodes both the audio and text from the given input. This work also shows preliminary results on unseen speaker lip-to-speech synthesis and out-of-vocabulary word generation. Akbari et al. [46] proposes a two-stage technique that first pretrains an audio auto-encoder to encode spectrogram into a latent space. A separate GRU-based network is then used to translate lip movements into this pretrained audio-latent space. However, collecting datasets with multiple views remains a challenging aspect and thus was not explored much by followup works.

Adding a discriminator in a GAN-setup A Wasserstein GAN [104] based model was proposed by [47], which adds a WGAN discriminator while training a standard 3D CNN-based network to convert lip movements into raw waveform. Both [46] and [47] are trained on GRID and TCD-TIMIT datasets in a speaker-specific fashion.

Mapping lip movements and speech into a joint distribution While most of these approaches directly try to translate a sequence of lip movements into plausible speech, Yadav et al. [48] proposed a newer approach involving a Variational Autoencoder setup [105]. In this approach, lip movement and speech distributions are attempted to be jointly learned in a variational setup. During training, the lip movements and corresponding speech are used to create two Gaussian distributions, both of which are tangled with each other using a KL-divergence loss. Points are sampled from either of these distributions and passed through a decoder (also jointly trained) to generate the final output. During inference, only the lip movement's distribution is used to generate speech. A newer version of this work was proposed in [49] where transformer [106] layers were used to replace the convolution layers in the previous work.

2.5.2 Unconstrained Single Speaker Lip-to-Speech Synthesis

While the authors of [44] were able to improve the performance of the original Vid2Speech network, the networks were still trained only on GRID [60] and the TCD-TIMIT [61] corpus which was both recorded in laboratory settings. The absence of realistic head movements, the extremely constrained vocabularies, and the overall lack of variations in these datasets made the task of lip-to-speech synthesis relatively simpler, but less applicable to real-world scenarios. Recognizing this gap, our work with Lip2Wav [50] was the first work that was primarily motivated to address and overcome these limita-

tions. We were the first to develop a model that could handle more natural, varied, and realistic speech scenarios, moving beyond the confines of laboratory settings.

Sequence-to-Sequence modelling of the problem One of the key contributions of this thesis, detailed in Chapter 3, is the development of Lip2Wav [50], a novel approach to lip-to-speech synthesis. This method is inspired by the observation that deaf individuals and professional lip readers find it easier to lip-read familiar speakers. Instead of attempting lip-to-speech on random speakers in the wild, the focus is placed on learning the speech patterns of a specific speaker through extended observation of their speech. Large amounts of in-the-wild videos are used for training, departing from previous CNN-based encoder-decoder networks. Lip2Wav [50] introduces sequence-to-sequence learning for this problem, utilizing a modified Tacotron-2 network [51] that processes lip movements instead of text. The model's architecture consists of a 3D face encoder for processing lip movements and a decoder for generating mel timesteps. Long Short-Term Memory (LSTM) layers and Bahdanau's attention mechanism [84] are employed in the decoder to focus on specific input data parts while generating output mel timesteps. The final melspectrogram is converted into a waveform using the Griffin-Lim [107] algorithm. Through training on a large lip-to-speech dataset, Lip2Wav demonstrates the ability to generate high-quality, natural-sounding speech when trained on specific speakers.

A follow-up work [52] by Kim et al. employs a multi-faceted approach built on Lip2Wav to improve speech synthesis. Specifically, they incorporated two key elements: (1) self-supervised speech representations to disambiguate homophones and (2) acoustic variance information to capture diverse speech styles. To further refine and enhance the quality of the generated speech, the authors also utilized a flow-based post-net designed to capture intricate details and add a layer of refinement to the synthesized speech.

2.5.3 Unconstrained Multi-Speaker Lip-to-Speech Synthesis

Initial approaches The most challenging version of lip-to-speech synthesis is undoubtedly unconstrained and multispeaker settings. In this setting, a lip-to-speech model is expected to handle unseen speakers, i.e., speakers not seen during training and work for videos captured in the wild. Here a multi-speaker network indicates the network handles unseen speakers and not multiple speakers present in the same video. While some constrained single-speaker works, such as [46,47], have attempted to tackle the multispeaker Lip-to-Speech task, they only did so using the limited GRID dataset, which lacks variation in terms of head motion, background change, speaking style, and vocabulary. As a result, these works cannot be considered true multispeaker efforts. Lip movements mainly contain information regarding the spoken content and prosody. However, attributes like the speaker's accent and voice [108–112] can only partially be determined from this source. Therefore, a multi-speaker lip-to-speech network will require a sequence of lip movements as input and a style token containing a target voice and accent information.

Using a pre-trained speaker embedding The proper multispeaker setup was first attempted in Lip2Wav [50], which used the word-level LRW dataset to train a version of their model with an ad-

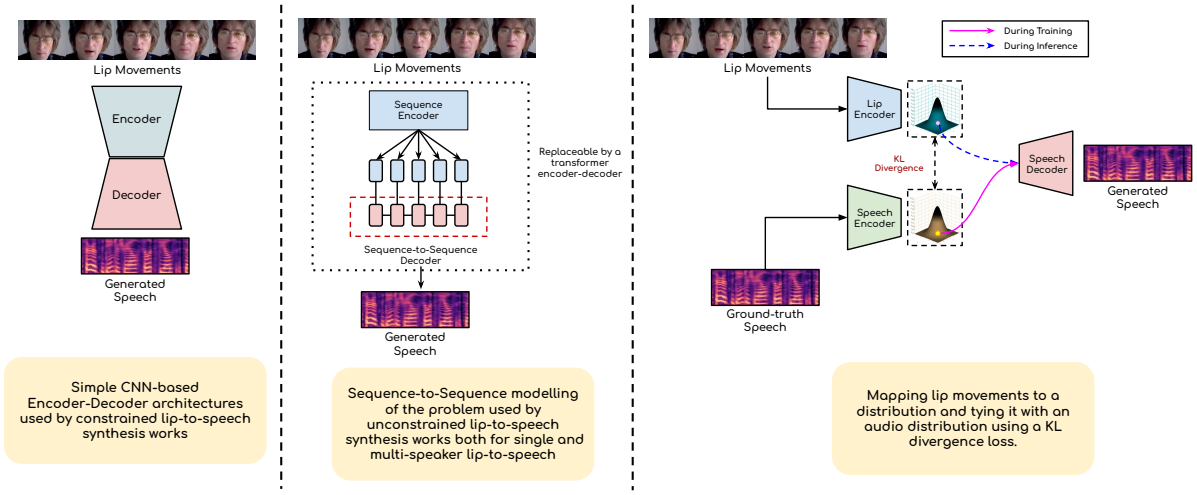

Figure 2.1 In the figure, we present a comprehensive overview of the primary training strategies employed in lip-to-speech networks over time. Early models addressing constrained lip-to-speech synthesis typically utilized a standard encoder-decoder architecture. This has since been improved by sequence-to-sequence learning models, which often leverage transformer architectures or other advanced sequence-to-sequence techniques. Additionally, some recent works have ventured into mapping lip movements and speech to interrelated distributions, aiming to achieve a more accurate and nuanced correspondence between the two modalities.

ditional speaker embedding as an input. Lip2Wav was trained with a speaker embedding from a pre-trained SV2TTS [55] network, which uses $1-5$ seconds of speech from a target speaker to generate the embedding.

Using visual context attention More recently, Kim et al. proposed VCA-GAN [56] that synthesizes speech from local lip visual features by finding a mapping function from viseme to phoneme while incorporating global visual context in the intermediate layers of the generator to disambiguate the mapping, which can be confused by homophene. To do this, VCA-GAN includes a visual context attention module that encodes global representations from local visual features and provides the desired global visual context to the generator through audio-visual attention. VCA-GAN also employs a sync loss (similar to Wav2Lip [27] to synthesize speech that is synchronized with the input lip movements. VCA-GAN [56] also attempted the problem in the same setup but did not use additional speaker embeddings. The cross-modal attention used by the network learned an inherent mapping between a face and the voice, which proved to be effective too. However, both of these works were limited to word-level generations, and Lip2Wav had been shown to be extremely limited on sentence-level datasets like LRS2 [28].

Mapping content from lip movements and speech into a joint distribution This led to the development of newer approaches that attempted to learn a mapping between lip movements and speech distributions. Hegde et al. [53] approached this problem in a variational setup like [48]. Hegde et al., too, learned an audio distribution and an entangled lip distribution during training while using only the lip distribution during inference. However, unlike Yadav et al. [48], the authors did not directly learn a distribution from the speech. Instead, they first used an ASR like the DeepSpeech2 [54] to generate

content embeddings, which were then used to learn a "content" distribution. The lip movements were used to learn a similar distribution, tied with a local and global KL-divergence loss. The decoder then sampled from either of these distributions during training (only from the lip movement's distribution during inference) and used a speaker-embedding from [55]. An improved Wasserstein GAN setup [113] discriminator was further utilized to enhance the quality of the generations. The generated melspectrograms were then converted back to the raw wave format using the Griffin-Lim [107] algorithm. The outputs from the VAE-GAN are clearer and louder thus helping the user understand the spoken content. However, these outputs were still gargled to an extent and did not have have maintain linguistic correctness throughout the generated speech. Chapter 4 of this thesis is dedicated to this work.

A separate work from [57] used a similar approach to Lip2Wav [50] but replaced the sequence-to-sequence model with a conformer [114] based architecture for multispeaker lip-to-speech. This approach from Mira et al. [57] used a pretrained ResNet18 [115] as a feature extractor for the mouth features and SV2TTS [55] for extracting target speaker's voice information. Both of these were passed through a conformer block, generating a melspectrogram. However, the authors used a Parallel WaveGAN [116, 117] to generate the final raw audio outputs.

Using multi-task learning to improve the clarity of generated speech Kim et al. introduced Multi-task Lip-to-Speech [118] that uses multi-task learning with both text and audio as supervisory signals to overcome limitations in word representation. The result is a system capable of synthesizing speech with the correct content for multiple speakers and unconstrained sentences.

Pre-trained lip-to-text assisted lip-to-speech synthesis The research conducted by our group, particularly in the study [58] which is a part of this thesis, represents a significant advancement in the domain of lip-to-speech synthesis. This approach uniquely employs pre-trained lip-to-text models, specifically utilizing the subword level lip reading technique discussed earlier. By feeding silent video input into these models, we generate text transcriptions, which, although imperfect, serve as a crucial intermediate step in speech synthesis. A key innovation in [58] is the use of Visual Transformer Pooling (VTP) embeddings, derived from the lip reading network, as visual features. These features play a vital role in conditioning the subsequent text-to-speech (TTS) synthesis process. Our TTS network, inspired by and adapted from the FastSpeech2 [59] architecture, is modified to incorporate these additional VTP embeddings. This modification allows the network to be conditioned not only on the text but also on the lip movements. The dual conditioning on both the predicted text and the lip movements equips the network with a nuanced understanding of speech generation. It knows what phonemes to utter based on the lip-derived text and precisely when to utter them, guided by the lip movements. This synergy between text prediction and lip movement timing significantly enhances the quality of the synthesized speech. During inference, this network effectively converts lip movements to speech with this intermediate text generation step. The result is a more accurate, natural, and high-quality speech output, closely mirroring the nuances of natural speech. This work is presented as a contribution in Chapter 5 of this thesis.

In a contemporary work by Yemini et al. [119], pre-trained lip-to-text models are also employed in the pipeline, but with a significantly different approach. Rather than using a pre-trained lip-to-text model to generate text that is subsequently converted into speech through a lip movement-conditioned TTS system, a different strategy is adopted. The pre-trained lip-to-text model is utilized more as a guiding mechanism in their lip-to-speech system. At the core of this method is MelGen, a diffusion-based speech generation module that directly translates lip movements into speech. To ensure accuracy and coherence of the generated speech, an additional verification step is incorporated. A pretrained Automatic Speech Recognizer (ASR) is used to transcribe the generated speech, and this transcription is then compared with the text derived from silent lip videos using the pre-trained lip-to-text network.

Notably, this work employs diffusion models and uses text for classifier guidance-based conditioning in diffusion, while also offering a classifier guidance-free version. The use of diffusion models and the innovative application of text-based guidance have contributed to this work's current status as the state-of-the-art in lip-to-speech generation, demonstrating superior performance in terms of speech quality and accuracy.

2.6 Lip-to-Speech synthesis conditioned on low-quality speech

Lip-to-speech synthesis can be enhanced by the incorporation of noisy speech inputs, a concept that can also be viewed as the integration of lip movements into speech enhancement processes. This approach, known as Audio-Visual Speech Enhancement (AVSE), is essentially an extension of plain lip-to-speech synthesis. In AVSE, the challenge of generating speech from lip movements is compounded by the need to account for and mitigate background noise in the audio signal. This methodology is particularly relevant in scenarios where audio quality is compromised but visual information remains clear. By combining visual cues from lip movements with noisy auditory inputs, AVSE models offer a more comprehensive approach to speech synthesis and enhancement, potentially improving the performance of speech systems in challenging acoustic environments.

Breakthroughs in this domain were marked by three seminal studies in 2018 [63, 120, 121]. Each of these approaches begins with a mixed speech signal where multiple speakers are talking simultaneously. Utilizing an encoder-decoder architecture, these studies employ both a speech encoder and a visual encoder focused on lip movements, followed by a speech decoder. The underlying assumption is that the lip movements are inherently synchronized with the speech of the individual speaker in question, thus offering a pathway to isolate that speaker's speech. To train these sophisticated systems, authors artificially introduce superimpose speeches from multiple speakers onto a clean signal. This noisy audio is then fed into the speech encoder while a separate, parallel visual encoder processes the lip movements of the target speaker. The outputs of both encoders are subsequently fused and passed to the speech decoder. This decoder is specifically trained to generate a clean speech signal that aligns with the lip movements. Through this innovative architecture, the network effectively learns the intricate correlation

between speech and lip movements, resulting in a model capable of separating a speech matching the lip movements from the mixed speech.

2.6.1 Adding visual input for speech super-resolution and denoising

In Chapter 6 of this thesis, significant contributions to the fields of speech super-resolution and speech denoising are presented. The integration of lip movements has been shown to enhance speech denoising and affect other speech tasks, notably speech super-resolution. In the work by Mukhopadhyay et al. [122], which forms a key contribution of this thesis, a remarkable advancement in speech super-resolution was demonstrated. A 16-fold increase in performance was achieved, significantly outperforming traditional audio-only methods. This substantial improvement was primarily attributed to the novel integration of lip movements with low-frequency speech inputs in the algorithm. As a result of this innovation, the audio-visual speech super-resolution algorithm has been established as state-of-the-art, capable of super-resolving at much higher scale factors than previously attainable. Following the success in speech super-resolution, the research was extended to the field of speech denoising. A similar hypothesis was employed, leveraging the synergy between audio and visual components. This exploration into audio-visual speech denoising was driven by the premise that visual cues, particularly lip movements, can significantly enhance the process of isolating clean speech from noisy environments. Both of these contributions, representing significant advancements in audio-visual speech processing, are thoroughly discussed in Chapter 6 of this thesis.

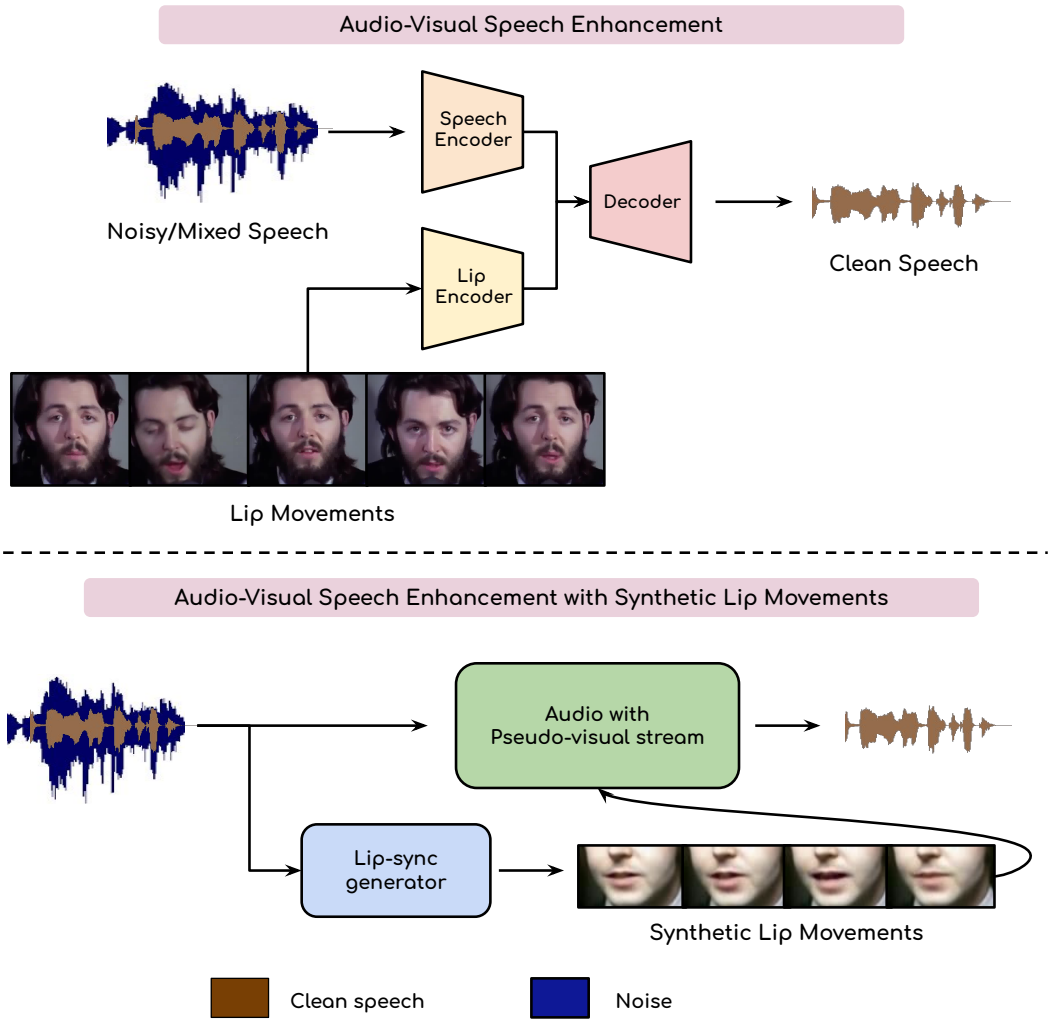

Figure 2.2 A visual representation of the general idea utilized by all the audio-visual denoising works. We also depict the use of a synthetic visual stream generated from the noisy speech to replace the real visual stream, allowing the standard audio-visual models to extend to audio-only settings where a visual stream is not naturally present.

Chapter 3

Unconstrained Single-speaker Lip-to-Speech Synthesis

3.1 Introduction

The first contribution of this thesis addresses lip-to-speech synthesis for individual speakers in unconstrained environments. Prior to this work, research in this field was primarily conducted in controlled laboratory settings, focusing on specific speakers. This contribution extends the paradigm to ”in-the-wild” scenarios while maintaining a speaker-specific approach, thereby bridging the gap between controlled experiments and real-world applications.

Lip movements play a crucial role in human communication and speech perception. Infants observe lip movements intently when learning to speak [123], while adults rely on visual cues to enhance speech comprehension in noisy environments. Individuals with hearing impairments often develop the ability to lip-read familiar speakers over time [124], facilitating more fluid conversations. These observations naturally lead to the question of whether a computational model can be developed to generate speech from lip movements by ”observing” a speaker for an extended period. Such a model would have significant advantages, requiring only videos of people talking without additional manual annotation. However, one of the fundamental challenges in this domain is the presence of homophones — visually similar lip shapes corresponding to different, auditorily distinct phonemes. This phenomenon significantly complicates the task of lip-to-speech synthesis, as it introduces a high degree of ambiguity in interpreting lip movements. Recognizing this a unique approach was formulated to focus on learning speech patterns from extended observation of individual speakers, inspired by the enhanced abilities of professional lip readers and individuals with hearing impairments to understand familiar speakers. The potential applications of this technology are diverse and impactful, ranging from enhancing video conferencing in silent environments to recovering high-quality speech from noisy backgrounds [120]. It could also be utilized for long-range surveillance, generating voices for individuals with aphonia, and even ”voice inpainting” [125] to replace corrupted speech segments. This chapter focuses on the development of such a model for single-speaker lip-to-speech synthesis in unconstrained environments, addressing the challenges and opportunities presented by this novel approach to speech generation.

Utilizing this dataset, Lip2Wav, a state-of-the-art sequence-to-sequence model, has been developed for generating natural, accurate speech that aligns with the lip movements of specific speakers. This

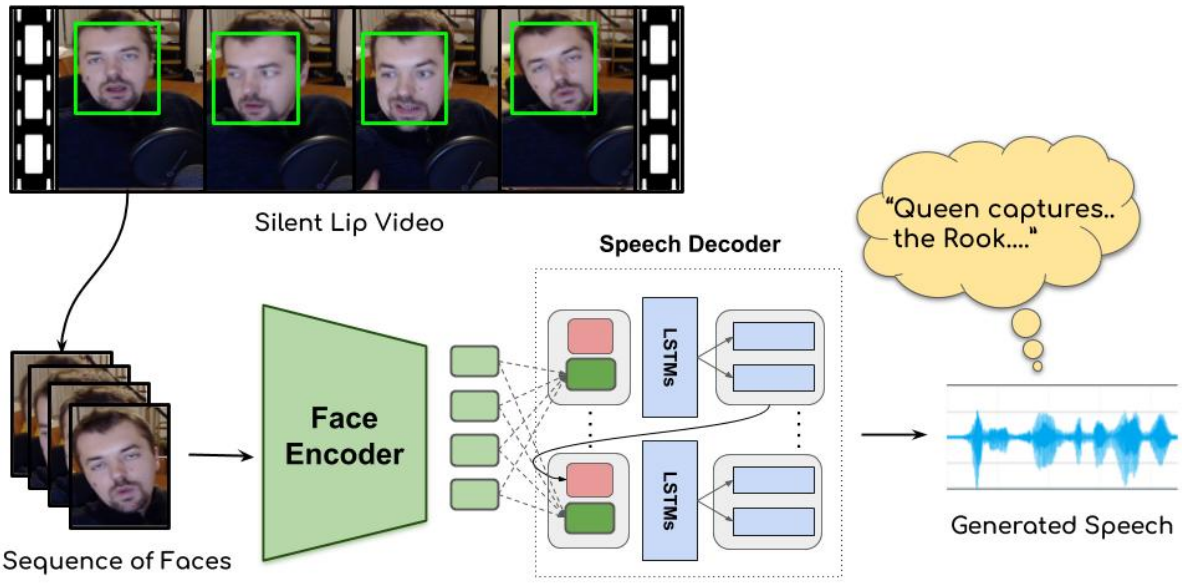

Figure 3.1 A sequence-to-sequence architecture named "Lip2Wav" is proposed for accurate speech generation from silent lip videos in unconstrained settings for the first time. The text in the bubble is manually transcribed and is shown for presentation purposes.

approach represents a departure from earlier works, offering enhanced intelligibility and naturalness in the generated speech. The model's effectiveness is supported by extensive quantitative and qualitative evaluations, including human studies that demonstrate its superior performance over previous models.

The key contributions of this chapter can be summarized as follows:

- A comprehensive investigation has been conducted on the silent lip video-to-speech generation problem in large vocabulary, unconstrained settings, marking a first in this field.
- The Lip2Wav dataset has been released, providing an unprecedented volume of data per speaker, coupled with a diverse vocabulary, to support the development of accurate, speaker-specific lip-to-speech models.
- Lip2Wav, a novel sequence-to-sequence model, has been introduced, significantly outperforming existing models in terms of intelligibility and naturalness in unconstrained environments.

This chapter aims to present the proposed novel approach and findings and set a new benchmark in the field of single-speaker lip-to-speech synthesis, paving the way for future research and applications in this exciting and rapidly evolving domain.

3.2 Speaker-specific Lip2Wav Dataset

As mentioned in Section 2.1, The datasets for lip-to-speech (or) text were at the two opposite ends of the spectrum: (i) small, constrained narrow vocabulary like GRID [60], TCD-TIMIT [61] or (ii) unconstrained, open vocabularies multi-speaker like LRS2 [28], LRW [30] and LRS3 [29]. The latter class of datasets contains only about 2 - 5 minutes of data per speaker, making it significantly harder for models to learn speaker-specific visual cues that are essential for inferring accurate speech from lip movements. Further, the results would also be directly affected by the existing challenges of multi-speaker speech synthesis [55, 97]. Conversely, the single-speaker lip-to-speech datasets [60, 61] do not emulate the natural settings as they are constrained to narrow vocabularies and artificial environments. Thus, neither of these extreme cases tests the limits of the unconstrained single-speaker lip-to-speech synthesis.

A new benchmark dataset for unconstrained lip-to-speech synthesis has been introduced, tailored towards exploring the following line of thought: *How accurately can an individual's speech style and content be inferred from his/her lip movements?* To create the Lip2Wav dataset, a total of about 120 hours of talking face videos across 5 speakers has been collected. The speakers were selected from various online lecture series and chess analysis videos. English was chosen as the sole language of the dataset. With approximately 20 hours of natural speech per speaker and vocabulary sizes over 5000 words[1] for each of them, this dataset is significantly more unconstrained and realistic than GRID [60] or TIMIT [61] datasets. It is thus considered ideal for learning and evaluating accurate person-specific models for the lip-to-speech task. The features of the Lip2Wav dataset are compared with other standard single-speaker lip-reading datasets in Table 3.1. It should be noted that a word is included in the vocabulary calculation for Table 2.1 only if its frequency in the dataset is at least five.

Dataset	Num. speakers	Total #hours	#hours per speaker	Vocab per speaker	Natural setting?
GRID [60]	34	28	0.8	56	×
TIMIT [61]	3	1.5	0.5	82	×
Lip2Wav (Ours)	5	**120**	**≈ 20**	**≈ 5K**	✓

Table 3.1 The Lip2Wav dataset is the first large-scale dataset tailored towards acting as a reliable benchmark for single-speaker lip-to-speech synthesis.

3.2.1 Steps to collect the dataset

The dataset collection process was systematically executed through a custom-designed script, which automated key tasks such as video downloading and segmentation. This methodological approach ensured consistency and efficiency in data gathering, a critical aspect of this research. The process comprised three main phases.

[1] approximate; texts obtained using Google ASR API

Directory Structure Initialization A target directory was specified as an input argument for all operations. Within this, a subdirectory was automatically created for storing downloaded videos, establishing a structured data repository.

Video Acquisition Videos were acquired using a command-line tool interfaced with an online platform. The acquisition process was guided by predefined text files containing URLs for training, validation, and testing sets. All downloaded content was systematically stored in the designated video storage subdirectory.

Video Segmentation Each acquired video file was subjected to a segmentation process. This involved dividing the videos into uniform segments of specified duration using a multimedia processing tool. The resulting segmented files were stored in a separate directory, with a naming convention designed for easy identification and retrieval.

This automated workflow facilitated efficient handling of large volumes of video data, rendering it suitable for subsequent tasks such as speech recognition and lip movement analysis. The following section presents a detailed examination of the script used, providing insight into the technical implementation of these procedures.

3.3 Lip2Wav: Speaker-specific lip-to-speech synthesis in unconstrained environments

Given a sequence of face images $I = (I_1, I_2, \ldots, I_N)$ with lip motion, the goal is to generate the corresponding speech segment $S = (S_1, S_2, \ldots, S_{T'}$. Here N is the total number of video frames, and T' is the number of speech time steps corresponding to those N frames. Numerous key design choices are made in the proposed Lip2Wav architecture to obtain natural speech in unconstrained settings. Below, the differences with the previous lip-to-speech approaches compared to the proposed one are highlighted.

3.3.1 Problem Formulation

In prior works on lip-to-speech synthesis, speech representation has been treated as a 2D-image [43, 47] for melspectrograms or as a single feature vector [43] in the case of LPC features. A 2D-CNN has been employed to decode these speech representations in both cases. This approach, however, violates the sequential ordering of speech data modeling, as future time steps are allowed to influence the prediction of the current time step. In contrast, the problem is formulated in this work within the standard sequence-to-sequence learning paradigm [126]. Specifically, each output speech time-step S_k is modeled as a conditional distribution of the previous speech time-steps $S_{<k}$ and the input face image sequence $I = (I_1, I_2, \ldots, I_N)$. The probability distribution of each output speech time step is given by:

$$P(S|I) = \Pi_k(S_k|S_{<k}, I) \tag{3.1}$$

Lip2Wav, as shown in Figure 3.2 consists of two modules: (i) Spatio-temporal face encoder and (ii) Attention-based speech decoder. The modules are trained jointly in an end-to-end fashion. The sequence-to-sequence approach enables the model to learn an implicit speech-level language model that helps it to disambiguate homophones.

3.3.2 Speech Representation

There are multiple output representations from which intelligible speech can be recovered, but each of them has its trade-offs. The LPC features are low-dimensional and easier to generate. However, they result in robotic, artificial-sounding speech. At the other extreme [47], one can generate raw waveforms, but the high dimensionality of the output (16000 samples per second) makes the network training process computationally inefficient. Inspired by previous text-to-speech works [51, 98] the proposed approach aims to generate melspectrograms conditioned on lip movements. The raw audio is sampled at 16kHz. The window-size, hop-size, and mel dimensions are $800, 200$, and 80, respectively. The algorithm to calculate melspectrograms is given in Algorithm 2 (Section 2.2). The ground-truth melspectrogram is represented by $Y_{1...T}$. Here T is the number of melspectrogram timesteps. In the setup, for 1 second of speech, there are $T' = 16000$ samples in the raw waveform while there are $T = 80$ mel timesteps.

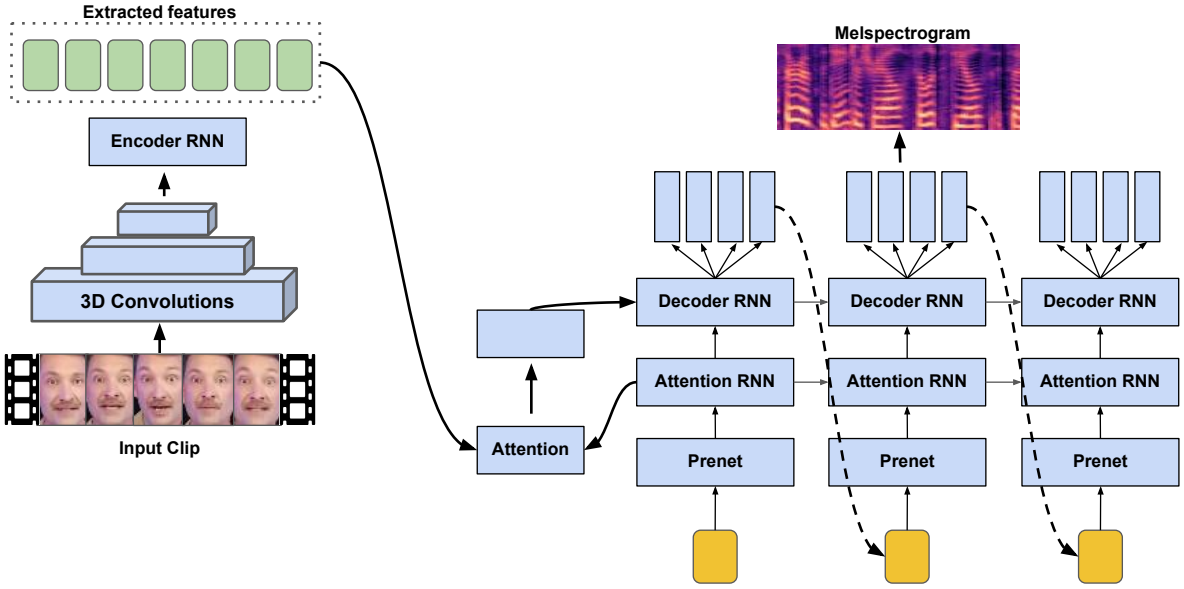

Figure 3.2 Lip2Wav model for lip-to-speech synthesis. The spatio-temporal encoder is a stack of 3D convolutions to extract the sequence of lip movements. This is followed by a decoder adapted from [51] for high-quality speech generation. The decoder is conditioned on the face image features from the encoder and generates the melspectrogram in an auto-regressive fashion.

3.3.3 Spatio-temporal Face Encoder

The visual input is presented as a short video sequence of face crops. The fine-grained sequence of lip movements must be extracted and processed by the model. 3D convolutional neural networks have been shown to be effective [47, 127, 128] in multiple tasks involving spatio-temporal video data. In this work, the spatio-temporal information of the lip movements is encoded using a stack of 3D convolutions followed by a stack of Bi-LSTM layers (Figure 3.2). The input to the network is a sequence of facial images of the dimension $N \times H \times W \times 3$, where N is the number of time-steps (frames) in the input video sequence, and H, W correspond to the spatial dimensions of the face image. The input is denoted as $I_{1...N}$. The corresponding speech melspectrogram to this segment is selected as the ground truth and is denoted by $Y_{1...T}$. The spatial extent of the feature maps is gradually down-sampled while the temporal dimension N is preserved. Residual skip connections [115] and batch normalization [129] are employed throughout the network. Each Conv3D block is followed by a ReLU activation. A single D-dimensional vector is output by the encoder for each of the N input facial images to obtain a set of spatio-temporal features $N \times D$. In the experimental setup, the spatio-temporal features are of shape 90×384. Information about future lip movements is also contained in each time step of the embedding generated from the encoder, which aids in the subsequent generation. The overall structure of the 3D CNN-based encoder is presented in Table 3.2.

Table 3.2 Structure of the Spatio-Temporal Encoder of the Lip2Wav Model

Layer	Kernel Size	Strides	#Filters / #Units	Residual	Input Size	Output Size
Conv3D	$5 \times 5 \times 5$	$1 \times 2 \times 2$	32	No	N $\times$ 96 $\times$ 96	N $\times$ 48 $\times$ 48
Conv3D	$3 \times 3 \times 3$	$1 \times 1 \times 1$	32	Yes	N $\times$ 48 $\times$ 48	N $\times$ 48 $\times$ 48
Conv3D	$1 \times 3 \times 3$	$1 \times 2 \times 2$	64	No	N $\times$ 48 $\times$ 48	N $\times$ 24 $\times$ 24
Conv3D	$1 \times 3 \times 3$	$1 \times 1 \times 1$	64	Yes	N $\times$ 24 $\times$ 24	N $\times$ 24 $\times$ 24
Conv3D	$1 \times 3 \times 3$	$1 \times 2 \times 2$	128	No	N $\times$ 24 $\times$ 24	N $\times$ 12 $\times$ 12
Conv3D	$1 \times 3 \times 3$	$1 \times 1 \times 1$	128	Yes	N $\times$ 12 $\times$ 12	N $\times$ 12 $\times$ 12
Conv3D	$1 \times 3 \times 3$	$1 \times 2 \times 2$	256	No	N $\times$ 12 $\times$ 12	N $\times$ 6 $\times$ 6
Conv3D	$1 \times 3 \times 3$	$1 \times 1 \times 1$	256	Yes	N $\times$ 6 $\times$ 6	N $\times$ 6 $\times$ 6
Conv3D	$1 \times 3 \times 3$	$1 \times 2 \times 2$	512	No	N $\times$ 6 $\times$ 6	N $\times$ 3 $\times$ 3
Conv3D	$1 \times 3 \times 3$	$1 \times 1 \times 1$	512	Yes	N $\times$ 3 $\times$ 3	N $\times$ 3 $\times$ 3
Conv3D	$1 \times 3 \times 3$	$1 \times 3 \times 3$	512	No	N $\times$ 3 $\times$ 3	N $\times$ 1 $\times$ 1
Bi-LSTM	-	-	256	-	N $\times$ 1 $\times$ 1	N $\times$ 256

The extracted spatio-temporal features are processed through a Bi-LSTM-based RNN encoder. These bidirectional LSTM layers excel in assimilating information from both antecedent and subsequent frames. This dual-directional processing endows the model with a comprehensive temporal context, which is crucial for accurately synthesizing speech from lip movements. To enhance the robustness of the model and mitigate the risk of overfitting, a zoneout regularization technique is integrated within these LSTM layers. This approach not only aids in preventing overfitting but also improves the model's generalization capabilities. The output from the LSTM layers is a concatenated representation of the

forward and backward sequence data, thereby enriching the model's interpretative capacity regarding complex spatio-temporal patterns inherent in visual speech data. Each of these layers comprises 256-dimensional Bi-LSTM units. The final encoded representation is denoted by Enc_N of the dimension 90×256.

3.3.4 Attention-based Speech Decoder

To achieve high-quality speech generation, multiple recent breakthroughs [51, 98] in text-to-speech generation are exploited. The Tacotron 2 [51] decoder is originally used to generate melspectrograms conditioned on text inputs. In this work, the decoder is conditioned on the encoded face embeddings from the previous module instead of text. The architecture of the decoder block comprises the following components -

- **Decoder Cell**: A decoder cell has multiple different components added consecutively, one after the other. They are listed below.

  - *Prenet*: The Prenet within the proposed architecture is a streamlined component consisting of two fully connected layers, each designed to function as an effective information bottleneck for the attention mechanism. These layers are identical in size, featuring 256 units each, which allows for a comprehensive yet efficient processing of input data. The activation function employed in both layers is ReLU (Rectified Linear Unit), known for its effectiveness in introducing non-linearity and mitigating the vanishing gradient problem in neural networks. A notable aspect of Prenet's design is the incorporation of dropout at a rate of 0.5 in both layers. This dropout rate plays a crucial role in enhancing the model's generalization capability and introducing variability in the generation process, which is particularly important during the inference phase.
  - *Attention Mechanism*: In the proposed model, the Location-Sensitive Attention mechanism is finely tuned with specific parameters to optimize its performance in speech synthesis. The query and key projection layers are crucial for aligning the dimensions of the decoder output and encoder outputs with the attention space; both are transformed to an attention dimension of 128. The Location Features Layer takes the previous alignment, processes it with 32 convolutional filters of a $31-$sized kernel, and projects the output to the same attention dimension. This alignment is essential for the mechanism to account for the sequence's positional context. The Score Calculation Layer then uses these $128-$dimensional vectors to compute attention scores for each encoder timestep. These scores are normalized by the Smoothing Normalization Layer, maintaining the sequence length. The context vector calculation Layer creates a context vector from the encoder outputs, weighted by the normalized attention scores. Finally, the Attention Integration Layer merges this context vector with the current decoder output, effectively combining content and context information. The attention mechanism accumulates weights over timesteps, facilitating a smoother transition

and focus across the input sequence, a key feature for accurate and coherent speech synthesis in the proposed system.

  - *Decoder LSTM*: The DecoderRNN in the proposed model is a vital component, structured with a series of LSTM layers specifically designed for effective decoding in speech synthesis. This decoder comprises two unidirectional LSTM layers, each equipped with the Zoneout regularization technique. The number of layers is set to 2. Each LSTM layer within the DecoderRNN contains 1024 units, providing substantial capacity to capture complex patterns in the data.
  - *Frame Projection*: The Frame Projection class in the proposed model serves as a vital projection layer, transforming the decoder's output to a specific dimensionality suited for subsequent processing stages in speech synthesis. This layer features a dense (fully connected) layer that transforms the input into a shape of 80 units. This transformation is crucial for ensuring that the decoder's output is in the correct format, enabling effective further processing and synthesis within the proposed speech generation model. The simplicity and precision of the Frame Projection layer make it an essential component in the model's architecture, ensuring consistency and accuracy in the output dimensionality.

- **Postnet**: The Postnet plays a critical role in refining the speech synthesis process in the proposed architecture. Functioning as a residual network, it takes the initial melspectrogram prediction from the Frame Projection layer and applies further processing to enhance its quality. The Postnet's output is effectively an adjustment or correction to the initial prediction. When added to the initial melspectrogram, this adjustment results in the final, enhanced melspectrogram. The addition of the Postnet's output to the initial prediction ensures a more accurate and higher-quality melspectrogram, crucial for generating natural-sounding speech. This mechanism demonstrates the importance of residual learning in speech synthesis, where slight modifications to an initial prediction can significantly improve the final output. In the proposed model, the Postnet is intricately designed with a series of convolutional layers to enhance the quality of the final melspectrogram. It consists of five convolutional layers, each contributing to refining the initial predictions from the decoder. These layers employ kernels of size 5, which are instrumental in processing the sequential data and extracting relevant temporal features. Each convolutional layer in the Postnet has 512 channels, indicating the number of filters used for feature extraction at each layer. A frame projection (dense layer) is then used at the end of the Postnet to ensure the final feature dimension of 80.

A summary of the different layers utilized in the decoder is presented in Table 3.3. The activation functions employed in various components of the decoder are as follows: The dense layers in the Prenet are equipped with the ReLU (Rectified Linear Unit) activation function. Within the Attention mechanism, the Conv1D layer and the Dense layer both incorporate the tanh (hyperbolic tangent) activation function. The LSTM layers in the DecoderRNN inherently integrate sigmoid and tanh functions within

Table 3.3 Decoder Components and Their Specifications

Decoder Component	Layer	Kernel Size	Strides	#Filters/#Units	Input Size	Output Size
Prenet	Dense	-	-	256	1×256	1×256
Prenet	Dense	-	-	256	1×256	1×256
Attention	Conv1D	(31,)	1	32	1×256	1×32
Attention	Dense	-	-	128	1×32	1×128
DecoderRNN	LSTM	-	-	1024	1×128	1×1024
DecoderRNN	LSTM	-	-	1024	1×1024	1×1024
FrameProjection	Dense	-	-	80	1×1024	1×80
Postnet	Conv1D	(5,)	1	512	1×80	1×512
Postnet	Conv1D	(5,)	1	512	1×512	1×512
Postnet	Conv1D	(5,)	1	512	1×512	1×512
Postnet	Conv1D	(5,)	1	512	1×512	1×512
Postnet	Conv1D	(5,)	1	512	1×512	1×512
FrameProjection	Dense	-	-	80	1×512	1×80

their architecture. The dense layer in the frame projection operates without an external activation function, serving as a linear projection. Finally, the Conv1D layers in the Postnet are implemented with the tanh activation function. This configuration of layers and activation functions has been carefully designed to optimize the decoder's performance in the lip-to-speech synthesis task.

In the model, the decoder operates in a sequence-to-sequence fashion, decoding one speech timestep at a time. For the i-th timestep of the melspectrogram, y_i, the decoder uses either Y_{i-1} or y_{i-1} as input, depending on whether teacher forcing is active. Specifically, if teacher forcing is on, Y_{i-1} (the ground truth speech's previous timestep) is used; otherwise, the previously generated timestep y_{i-1} is utilized as shown in Equation 3.2.

$$y_i = \text{Decoder\_cell}(\text{teacher forcing is on ? } Y_{i-1} : y_{i-1}, Enc_N) \tag{3.2}$$

To initiate the generation process, a start token is used in place of y_{i-1} when $i = 1$. This approach ensures that the decoding at each step is informed by the appropriate past output and the overall context provided by the encoded sequence Enc_N, leading to coherent and contextually accurate speech synthesis. The decoder cell is presented in more detail next.

In the sequence-to-sequence decoding process of the proposed model, the output from the Prenet for the i-th timestep is denoted as p_i. This output is determined based on whether teacher forcing is active. The Prenet processes either the ground truth speech's previous timestep Y_{i-1} or the previously generated timestep y_{i-1}, as described by the following equation:

$$p_i = \text{Prenet}(\text{teacher forcing is on ? } Y_{i-1} : y_{i-1}) \tag{3.3}$$

In the proposed sequence-to-sequence model, the context vector for the previous timestep is denoted as c_{i-1}. This context vector, calculated by the attention mechanism, plays a crucial role in decoding. For each timestep i, the output of the DecoderRNN denoted as d_i, is obtained by concatenating the Prenet

output p_i with the previous timestep's context vector c_{i-1} and then feeding this combined input into the DecoderRNN. This operation can be represented by the following equation:

$$d_i = \text{DecoderRNN}(p_i \| c_{i-1}) \tag{3.4}$$

where $\|$ denotes the concatenation operation. The context vector c_{i-1} is a result of the attention mechanism applied to the encoded sequence and the previous decoder outputs, providing the necessary contextual information for the current decoding step.

The context vector c_i for each decoder timestep i, where i ranges from 1 to T, is computed through an attention mechanism. In this mechanism, the set of keys is represented by Enc_N, where N is the number of encoder timesteps, and the query for each timestep is the output of the DecoderRNN, d_i. The following equations can represent the attention mechanism:

The calculation of the attention energy at each decoder timestep i and for each encoder timestep j is given by:

$$e_{i,j} = \text{AttentionEnergy}(d_i, Enc_j) \tag{3.5}$$

where j ranges from 1 to N, with N being the number of encoder timesteps. Here, $e_{i,j}$ represents the energy between the query d_i (the output of the DecoderRNN at timestep i) and the j-th encoder output Enc_j.

The alignment scores or attention weights are then computed using a softmax function over these attention energies:

$$\alpha_{i,j} = \frac{\exp(e_{i,j})}{\sum_{k=1}^{N} \exp(e_{i,k})} \tag{3.6}$$

where $\alpha_{i,j}$ represents the attention weight for the j-th encoder output at the i-th decoder timestep.

Finally, the context vector c_i for each decoder timestep is computed as a weighted sum of the encoder outputs:

$$c_i = \sum_{j=1}^{N} \alpha_{i,j} \cdot Enc_j \tag{3.7}$$

This context vector c_i captures the attended information from the encoder outputs relevant to the current decoding step and is utilized in conjunction with the Prenet output for further decoding operations.

The final step in the decoding process involves computing the i-th generated melspectrogram timestep, y_i. This is achieved by concatenating the output of the DecoderRNN at timestep i, d_i, with the context vector at the same timestep, c_i, and then passing this concatenated vector through the Frame Projection layer. The operation can be expressed as:

$$y_i = \text{FrameProjection}(d_i \| c_i) \tag{3.8}$$

where $\|$ denotes the concatenation operation. The Frame Projection layer reshapes this combined input to produce the final melspectrogram output for timestep i, denoted as y_i. This output y_i represents the synthesized mel chunk at that particular timestep, which in turn represents the synthesised speech. The above-mentioned steps are repeated T times to generate the full melspectrogram. The generated melspectrogram is denoted by y. This melspectrogram is then provided to the Postnet to calculate a residual and improve output quality. First, the residual is obtained by passing the decoder output y through the Postnet.

$$\text{residual} = \text{Postnet}(y) \tag{3.9}$$

Next, this residual is projected to match the dimensions of the melspectrogram:

$$\text{residual\_projection} = \text{FrameProjection}(\text{residual}) \tag{3.10}$$

Finally, the enhanced melspectrogram output, denoted as y', is computed by adding the initial melspectrogram output y to the projected residual:

$$y' = y + \text{residual\_projection} \tag{3.11}$$

This step, involving the Postnet and Frame Projection, effectively adds refinement and detail to the initial melspectrogram, resulting in y', the final generated high-quality melspectrogram. The generated melspectrogram is used to calculate the standard L_1 loss function as given in Equation 3.12.

$$L_1 \text{ loss} = ||Y - y'||_1 \tag{3.12}$$

A brief discussion on adapting Tacotron 2's decoder for Lip2Wav In adapting the Tacotron 2 decoder for the specific application of lip-to-speech synthesis, a significant alteration was made: the removal of the stop token prediction mechanism. This modification is driven by the nature of the input and output sequences, which are of a constant size. In the original Tacotron 2 model, the stop token prediction is crucial for determining the end of the speech generation process, especially since the length of the output speech could vary depending on the input text. However, in the proposed model, the lengths of the input video sequence and the corresponding output speech are predefined and fixed. As a result, the need for a dynamic stopping criterion, such as a stop token, is not required, leading to a more streamlined and efficient decoding process that runs for a set number of iterations based on the predetermined sequence length. This adjustment simplifies the model's architecture and aligns the decoding process more closely with the fixed structure of the lip-to-speech synthesis task.

3.3.5 Gradual Teacher Forcing Decay

In the initial stages of training, up to approximately 30K iterations, teacher forcing is employed, similar to the text-to-speech counterpart. It is hypothesized that this enables the decoder to learn an implicit speech-level language model to help disambiguate homophones. Similar observations have

been made in lip-to-text works [86], which employ a transformer-based sequence-to-sequence model. Over the course of the training, the teacher forcing is gradually decayed to enforce the model to attend to the lip region and to prevent the implicit language model from over-fitting to the train set vocabulary. The effect of this decay is examined in sub-section 3.5.3.

3.3.6 Context Window Size

A larger visual context window for inferring the current speech time-step is employed to help the model disambiguate homophones [43]. The context size used in this work is approximately 6 times larger than in prior works. It is shown in sub-section 3.5.1 that this design choice results in significantly more accurate speech. A context window size of 3 seconds is utilized to achieve the best results in the experiments conducted.

3.4 Benchmark Datasets and Training Details

3.4.1 Datasets

The primary focus of the proposed work is on single-speaker lip-to-speech synthesis in unconstrained, large vocabulary settings. The Lip2Wav model is also trained on the GRID corpus [60] and the TCD-TIMIT lip speaker corpus [61] to compare with previous works. Next, the model is trained on all five speakers of the newly collected speaker-specific Lip2Wav dataset. Unless specified, all the datasets are divided into $90 - 5 - 5\%$ train, validation and unseen test splits. In the Lip2Wav dataset, these splits are created using different videos, ensuring that no part of the same video is used for both training and testing. The train and test splits are also released with this work[2] for fair comparison in future works.

3.4.2 Training Methodology and Hyper-parameters

A training input example is prepared by randomly sampling a contiguous sequence of 3 seconds, which corresponds to $T = 75$ or $T = 90$, depending on the frame rate (FPS) of the video. The effect of various context window sizes is studied in Section 3.5.1. The face is detected and cropped from the video frames using the S^3FD face detector [64]. The face crops are resized to 48×48. The melspectrogram representation of the audio corresponding to the chosen short video segment is used as the desired ground truth for training. The hidden dimension is halved to prevent over-fitting for training on small datasets like GRID and TIMIT. The training batch size is set to 32, and training is continued until the mel reconstruction loss plateaus for at least $30K$ iterations. Convergence for unconstrained single-speaker experiments was achieved in about $200K$ iterations. Adam [130] is used as the optimizer with an initial learning rate of 10^{-3}. The model with the best performance on the validation set is chosen for testing and evaluation. More details, specifically a few minor speaker-specific hyper-parameter changes, can be found in the publicly released code[2].

[2] `https://github.com/rudrabha/Lip2Wav`

3.4.3 Speech Generation at Test Time

During inference, only the sequence of lip movements is provided, and the speech is generated in an auto-regressive fashion. It should be noted that speech can be generated for lip sequences of any length. Consecutive N frame windows are taken, the speech for each of them is generated independently, and they are concatenated together. A small overlap across the sliding windows is maintained to adjust for boundary effects. The waveform is obtained from the generated melspectrogram using the Griffin-Lim algorithm [107]. Griffin-Lim algorithm itself does not work directly on melspectrograms. It expects a normal linear frequency magnitude spectrogram. To apply the Griffin-Lim algorithm, we take the predicted melspectrogram and apply an approximate inverse of the mel filter bank using a pseudo inverse of the mel filter matrix, to recover an approximate linear magnitude spectrogram. Once we have this approximate linear magnitude, we can run the standard Griffin-Lim algorithm on it to iteratively estimate the phase and reconstruct the waveform. It was observed that neural vocoders [99, 116, 117] perform poorly in this case as the generated melspectrograms are significantly less accurate than state-of-the-art TTS systems. Finally, the ability to generate speech for lip sequences of any length is worth highlighting, as the performance of the current lip-to-text works trained at sentence level deteriorates sharply for long sentences that barely last over just 4 - 5 seconds [86].

3.4.4 Metrics used to measure the quality of the generated speech from different methods

In this section, the metrics used to measure the quality of generated speech are described. These metrics are not only employed in this chapter but also serve as a mainstay throughout the thesis, being utilized in subsequent chapters as well. The quality of generated speech from different methods is measured using three standard speech quality metrics: Short-Time Objective Intelligibility (STOI) [131], Extended Short-Time Objective Intelligibility (ESTOI) [132], and Perceptual Evaluation of Speech Quality (PESQ) [133].

3.4.4.1 Short-Time Objective Intelligibility (STOI)

STOI is designed to estimate the intelligibility of speech, which is particularly useful in assessing speech clarity. In this algorithm, both the reference and processed speech signals are divided into short time frames. These frames are transformed into a time-frequency representation. The correlation between corresponding time-frequency units of the reference and processed signals is calculated by STOI, and these correlations are averaged to yield an intelligibility score. A higher STOI score indicates better speech intelligibility. The STOI value ranges between $0 - 1$.

3.4.4.2 Extended Short-Time Objective Intelligibility (ESTOI)

ESTOI is developed as an extension of STOI, specifically tailored for non-stationary noise conditions. The STOI algorithm is modified by introducing a normalization step and different weighting of

the time-frequency units. A correlation-based score reflecting speech intelligibility is computed by the algorithm, making it more sensitive to variations in noisy environments. ESTOI is effectively used in assessing speech enhancement techniques and communication systems in challenging acoustic settings. The ESTOI value ranges between $0 - 1$.

3.4.4.3 Perceptual Evaluation of Speech Quality (PESQ)

PESQ is employed to measure the quality of speech, accounting for both distortion and noise. In this algorithm, a reference speech signal is compared with a degraded version to provide a quality score. Key steps include time alignment of the signals, psychoacoustic modeling, and calculation of a disturbance value quantifying the perceptual difference between the reference and degraded signals. PESQ is widely used in the evaluation of telecommunication systems and speech codecs, offering a comprehensive measure of speech quality as perceived by human listeners. The PESQ value ranges between $-0.5 - 4.5$.

3.4.5 Lip to Speech in Constrained Settings

The evaluation of the proposed approach against previous lip-to-speech works is initiated with constrained datasets, specifically the GRID [60] corpus and TCD-TIMIT lip speaker corpus [61]. The mean test scores for 4 speakers are reported for the GRID dataset, which are also presented in the previous works. The results for GRID and TIMIT datasets are summarized in Tables 3.4 and 3.5, respectively.

Method	STOI	ESTOI	PESQ	WER
Vid2Speech [43]	0.491	0.335	1.734	44.92%
Lip2AudSpec [46]	0.513	0.352	1.673	32.51%
GAN-based [47]	0.564	0.361	1.684	26.64%
Ephrat et al. [44]	0.659	0.376	**1.825**	27.83%
Lip2Wav (ours)	**0.731**	**0.535**	1.772	**14.08%**

Table 3.4 Objective speech quality, intelligibility and WER scores for the GRID dataset unseen test split.

Method	STOI	ESTOI	PESQ	WER
Vid2Speech [43]	0.451	0.298	1.136	75.52%
Lip2AudSpec [46]	0.450	0.316	1.254	61.86%
GAN-based [47]	0.511	0.321	1.218	49.13%
Ephrat et al. [44]	0.487	0.310	1.231	53.52%
Lip2Wav (ours)	**0.558**	**0.365**	**1.350**	**31.26%**

Table 3.5 Objective speech quality, intelligibility and WER scores for the TCD-TIMIT dataset unseen test split.

Method	Speaker	STOI	ESTOI	PESQ
GAN-based [47]	*Chemistry Lectures*	0.192	0.132	1.057
Ephrat et al. [44]		0.165	0.087	1.056
Lip2Wav (ours)		**0.416**	**0.284**	**1.300**
GAN-based [47]	*Chess Analysis*	0.195	0.104	1.165
Ephrat et al. [44]		0.184	0.098	1.139
Lip2Wav (ours)		**0.418**	**0.290**	**1.400**
GAN-based [47]	*Deep Learning*	0.144	0.070	1.121
Ephrat et al. [44]		0.112	0.043	1.095
Lip2Wav (ours)		**0.282**	**0.183**	**1.671**
GAN-based [47]	*Hardware Security*	0.251	0.110	1.035
Ephrat et al. [44]		0.192	0.064	1.043
Lip2Wav (ours)		**0.446**	**0.311**	**1.290**
GAN-based [47]	*Ethical hacking*	0.171	0.089	1.079
Ephrat et al. [44]		0.143	0.064	1.065
Lip2Wav (ours)		**0.369**	**0.220**	**1.367**

Table 3.6 In unconstrained single-speaker settings, Lip2Wav model achieves almost $4\times$ more intelligible speech than the previous methods.

A significant margin of improvement can be observed in the proposed approach compared to competing methods across all objective metrics. The difference is particularly noticeable in the TIMIT [61] dataset, where the test set is characterized by a high proportion of novel words unseen during training. This aspect of the test set is especially noteworthy, as it challenges the model's ability to generalize to new vocabulary. The superior performance of the proposed model in this context demonstrates that correlations across short phoneme sequences are effectively captured, leading to better pronunciation of new words compared to previous methods. This capability to handle unseen words is a crucial advancement in lip-to-speech synthesis.

3.4.6 Lip to Speech in Unconstrained Settings

The evaluation of the proposed approach is now extended to unconstrained datasets, which are characterized by a significant amount of head movements, much broader vocabularies, and substantial periods of silence or pauses between words and sentences. It is in this context that a more vivid difference between the proposed approach and previous approaches can be observed. Our model is independently trained on all 5 speakers of the newly collected Lip2Wav dataset. The training details are provided in sub-section 3.4.2. For comparison with previous works, the best performing models [44, 47] on the TIMIT dataset, based on STOI scores, are selected and their performance after training on the proposed Lip2Wav dataset is reported. The same metrics for speech intelligibility and quality that are used in Table 3.5 are computed. The scores for all five speakers for the proposed method and the two competing methods across all three metrics are presented in Table 3.6.

Much more intelligible and natural speech is produced by the proposed approach across different speakers and vocabulary sizes. Notably, more accurate pronunciation is achieved by the proposed model, as evidenced by the increased STOI and ESTOI scores compared to the previous works.

3.4.7 Human Evaluation

In addition to speech quality and intelligibility metrics, it is important to manually evaluate the speech as these metrics are not perfect [44] measures.

3.4.7.1 Objective Human Evaluation

In this study, human participants are asked to manually identify and report (A) the percentage of mispronunciations, (B) the percentage of word skips, and (C) the percentage of mispronunciations that are homophones. Word skips denote the number of words that are either completely unintelligible due to noise or slurry speech. Ten predictions from the unseen test split of each speaker in the Lip2Wav dataset are chosen to get a total of 50 files. The mean numbers of (A), (B), and (C) are reported in Table 3.7.

Model	(A)	(B)	(C)
GAN-based [47]	36.6%	24.3%	63.8%
Ephrat et al [44]	43.3%	27.5%	60.7%
Lip2Wav (ours)	**21.5%**	**8.6%**	**49.8%**

Table 3.7 Objective Human evaluation results. The participants manually identified the percentage of (A) Mispronunciations, (B) Word skips and (C) Homophene-based errors in the test samples.

Far fewer mispronunciations are made by the proposed approach compared to the current state-of-the-art method. Words are also skipped $3\times$ less frequently; however, the key point to note is that the issue of homophones remains a dominant cause for errors in all cases, indicating that there is still scope for improvement in this area.

3.4.7.2 Subjective Human Evaluation

Human participants, 15 in number, are asked to rate the different approaches for unconstrained lip-to-speech synthesis on a scale of $1-5$ for each of the following criteria: (i) *Intelligibility* and (ii) *Naturalness* of the generated speech. Using 10 samples of generated speech for each of the 5 speakers from the Lip2Wav dataset, the following approaches are compared: (i) The proposed Lip2Wav model (ii) Current state-of-the-art lip to speech models [44, 47] (iii) Manually transcribed text followed by a multi-speaker TTS [51, 55] to demonstrate that even with the most accurate text, lip to speech is not a concatenation of lip-to-text and text-to-speech. And finally, (iv) Human speech is also added for reference. In all cases, the speech is overlaid on the face video before being presented to the rater. The mean scores are reported in Table 3.8.

It is important to note that, ultimately, the generated speech is intended for human consumption. Thus, human evaluation is considered absolutely necessary for any of these works, as it provides crucial insights into the perceived quality and naturalness of the synthesized speech.

Approach	Intelligibility	Naturalness
GAN-based [47]	1.56	1.71
Ephrat et al. [44]	1.34	1.67
Lip2Wav (ours)	**3.04**	**3.63**
MTT + TTS [51]	3.86	3.15
Actual Human Speech	4.82	4.95

Table 3.8 Mean human evaluation scores based on speech quality and intelligibility for various approaches for lip to speech. MTT denotes "manually-transcribed text". The penultimate row simulates the best possible case of an automatic lip-to-text followed by a state-of-the-art text-to-speech system. In this case, the drop in naturalness score illustrates the loss in speech style and prosody.

In line with the previous evaluations, it can be observed that significantly higher quality and legible speech is produced by the proposed approach compared to the previous state-of-the-art [47]. It is observed that a baseline approach using lip-to-text followed by plain text-to-speech (TTS) is implemented in this evaluation. While this method produces linguistically accurate content, it is noted that the generated speech is completely out of sync with the lip movements. This lack of synchronization is due to the TTS component operating independently of the visual input. As a result, this approach is not considered a viable solution for lip-to-speech synthesis. It is worth noting that in Chapter 5 of this thesis, an improved approach is proposed that addresses this limitation. In the advanced method, the TTS component is conditioned on lip movements, thereby maintaining synchronization between the generated speech and the visual input. This development represents a significant improvement over the baseline method evaluated here, highlighting the importance of integrating visual information throughout the speech synthesis process.

3.4.7.3 Qualitative results

Qualitative results are presented in the form of a video linked in Figure 3.3, providing a dynamic and illustrative demonstration of the research findings. This video offers an auditory representation of the results, complementing the textual and quantitative analyses provided elsewhere in this thesis. However, it should be noted that to access this feature, opening the document in Adobe Acrobat Reader is essential, as other PDF viewers may not support the interactive multimedia functionality.

It is important to mention that this region may appear empty in the printed version of the thesis. Readers are strongly encouraged to refer to the online version of the thesis to access the video. It is emphasized that listening to this video is extremely important for a thorough understanding of the generation quality achieved by the proposed method. The auditory aspects of the lip-to-speech synthesis can be fully appreciated only through this multimedia presentation, providing crucial insights into the qual-

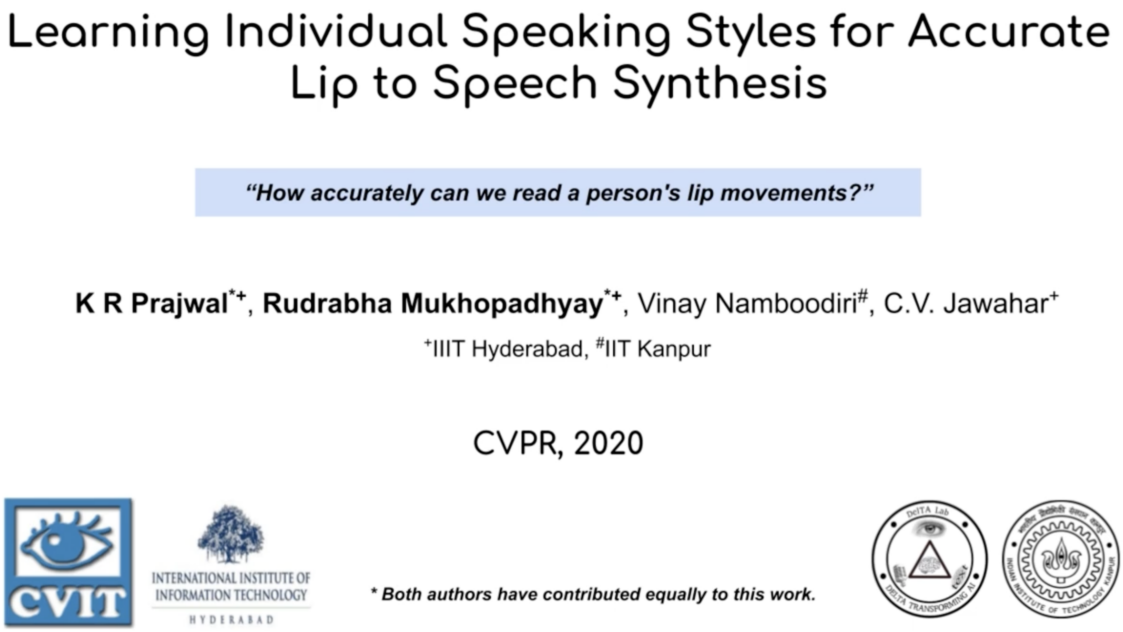

Figure 3.3 A video is presented in this link: `https://youtu.be/8VnEHcRa2l4`. This video contains qualitative results and comparisons from the proposed proposed Lip2Wav model. This image is presented as the thumbnail for this video.

ity and naturalness of the generated speech that may not be fully conveyed through text or quantitative metrics alone.

The experimental section concludes here, demonstrating significant improvements over previous lip-to-speech works. In the next section, ablation studies on the model are conducted.

3.5 Ablation Studies

In this section, different aspects of the Lip2Wav approach are probed. All results presented are calculated using the unseen test predictions on the "Hardware Security" speaker of the Lip2Wav dataset.

3.5.1 Larger context window helps in disambiguation

As stated before, the lip-to-speech task is highly ambiguous and can be inferred solely from lip movements. To combat this, reasonably large context information is provided to the model to disambiguate a given viseme. Previous works, however, have used only about $0.3 - 0.5$ seconds of context. In this work, close to 6 times more context is used, providing a context of 3 seconds. This allows the model to disambiguate by learning co-occurrences of phonemes and words, and the resulting improvement is evident in Table 3.9.

Context Window size	STOI	ESTOI	PESQ
0.5 seconds	0.264	0.193	1.062
1.5 seconds	0.321	0.226	1.080
3 seconds	**0.446**	**0.311**	**1.290**

Table 3.9 Larger context information consistently results in more accurate speech generation. The window size is limited to 3 seconds due to memory constraints.

3.5.2 Model is highly attentive to the mouth

A simple test is performed to empirically show that the model is highly attentive towards the lip and mouth region despite large changes in face pose and expressions. The top half of the face is blurred in one case, and the bottom half containing the mouth region is blurred in the second case. The blurred regions are virtually incomprehensible to the human eye. In Table 3.10, a drastic drop in the metrics is observed when the mouth region is blurred, which is not the case when the top half of the face is blurred.

Blur Type	STOI	ESTOI	PESQ
No blur	0.446	0.311	1.290
Top Half	0.312	0.139	1.195
Bottom half	**0.226**	**0.068**	**1.164**

Table 3.10 Blurring the mouth region drastically affects the generated speech compared to blurring the top half of the face.

The activations of the penultimate layer of the spatio-temporal face encoder are plotted in Figure 3.4 to show that the encoder is highly attentive towards the mouth region of the speaker. The attention alignment curve in Figure 3.5 shows that the decoder conditions on the appropriate video frame's lips while generating the corresponding speech.

3.5.3 Teacher Forcing vs Non-Teacher Forcing

To accelerate the training of a sequence-to-sequence architecture, the previous time step's ground truth (instead of the generated output) is typically given as input to the current time step. While this is highly beneficial in the initial stages of training, it was observed that gradually decaying the teacher forcing from $\approx 30K$ iterations significantly improves results and prevents over-fitting to the training vocabulary. A similar improvement was also observed in lip-to-text works [86]. In Table 3.11, a significant improvement in test scores is shown by gradually decaying teacher forcing.

3.5.4 Effect of different spatio-temporal encoders

While a 3D-CNN was used as the spatio-temporal encoder in Section 3.3.3, achieving the best results in the experiments to capture both spatial and temporal information in unconstrained settings, the effect of using different kinds of encoders is also reported in Table 3.12. The encoder module was replaced

Figure 3.4 The activations of the penultimate layer of the face encoder and the attention alignment from the decoder are plotted. It is observed that the face encoder is highly attentive towards the mouth region.

Teacher-forcing	STOI	ESTOI	PESQ
Always forced	0.221	0.162	1.141
Gradual decay	**0.446**	**0.311**	**1.290**

Table 3.11 Gradually decaying the teacher forcing enables the model to generalize to unseen vocabulary by forcing it to look at the visual input and not just predict from the previously uttered speech.

while keeping the speech decoder module intact. It is observed that the best performance is obtained with a 3D-CNN encoder.

Encoder	STOI	ESTOI	PESQ
2D-CNN	0.291	0.211	1.112
2D-CNN + 1D-CNN	0.298	0.223	1.170
3D-CNN (ours)	**0.446**	**0.311**	**1.290**

Table 3.12 Lip2Wav employs a 3D-CNN encoder to capture the spatio-temporal visual information and is the superior choice over the other alternatives.

3.5.5 Effect of resolution of the input face crops

The input resolution for the face crops to the network was selected to be 96×96. However, this resolution was inspired by other parallel works like [134]. Therefore, the input resolution is varied to understand the effect of this property. The visual encoder is modified accordingly to handle the different input dimensions. The output of the visual encoder is unchanged and thus the rest of the network is not modified. The results from this experiment are given in Table 3.13.

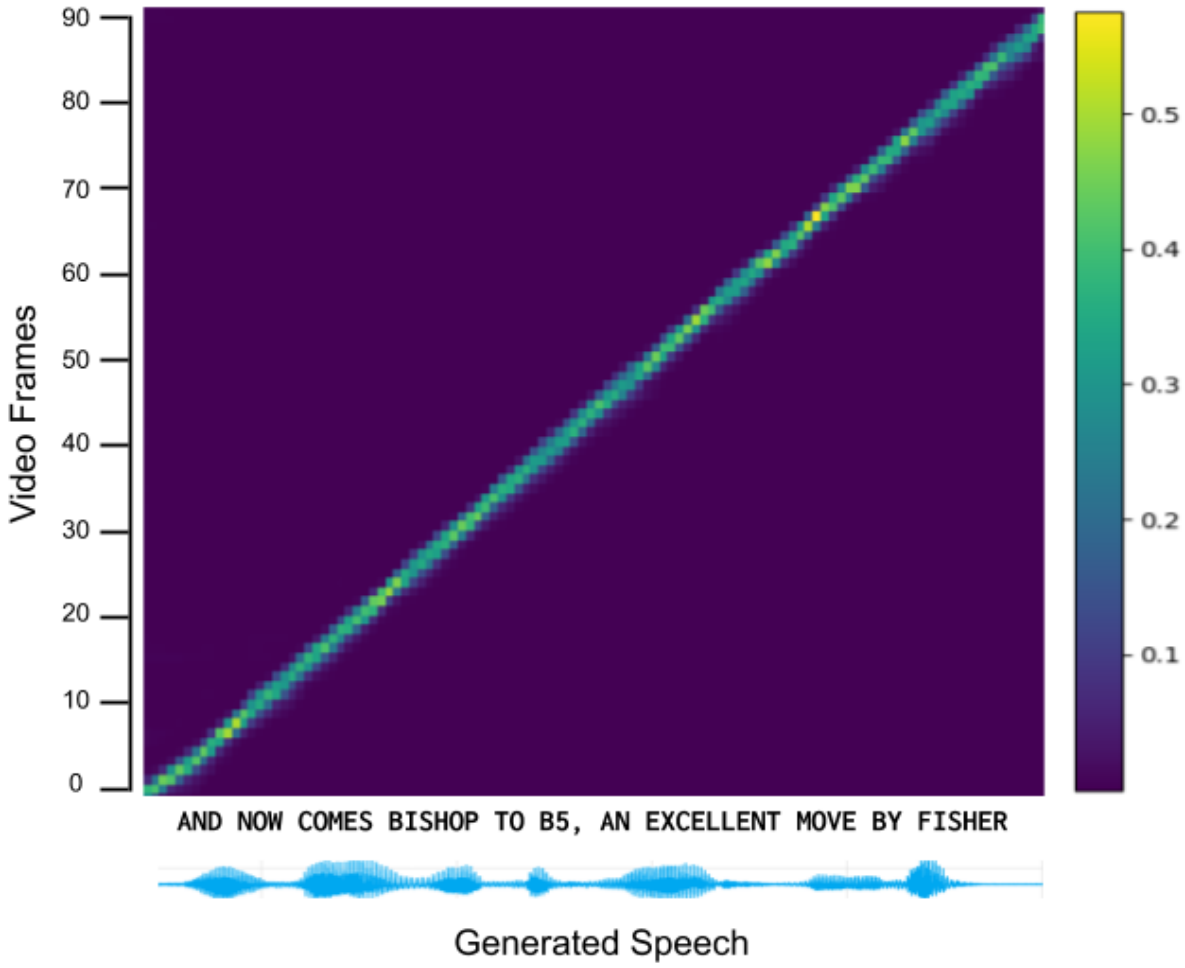

Figure 3.5 The decoder alignment curve illustrates that the model is generating speech by strongly conditioning on the corresponding lip movements.

Resolution	STOI	ESTOI	PESQ
48×48	0.442	0.308	1.285
96×96	0.446	0.311	1.290
128×128	0.449	0.314	1.293
256×256	0.389	0.275	1.198

Table 3.13 Analysis of different input resolutions. It is to be noted that, in fact, 128×128 shows marginally better performance. However, there is a drop in performance as the resolution is increased to 256×256

3.6 Multi-speaker Word-level Lip-to-Speech

In the field of speech synthesis, the concept of single-speaker models involves training on an extensive dataset from one individual speaker. While this approach allows for highly accurate lip-to-speech synthesis for that specific speaker, it inherently limits the model's effectiveness when applied to individuals with different vocal characteristics and lip movements. Consequently, a model trained on a single speaker typically fails to generalize to other identities, presenting a significant constraint in broader, real-world applications. Recognizing this limitation, the scope of the proposed Lip2Wav approach is extended to address the challenge of multi-speaker lip-to-speech synthesis. The objective is to train models capable of effectively handling speech synthesis for any arbitrary identity encountered in diverse settings. This expansion aims to create a more versatile and universally applicable model, transcending the constraints of single-speaker dependencies.

This chapter introduces the first multi-speaker adaptation of the Lip2Wav model, marking a significant transition in the research focus of this thesis. As a natural and logical progression from the single-speaker setup explored till now, this preliminary extension into multi-identity scenarios represents an important step towards more versatile and generalizable lip-to-speech synthesis. The multi-speaker approach aims to capture not only the general principles of lip movement and speech correlation but also to incorporate the unique vocal characteristics and articulation styles of different speakers. This initial foray into multi-speaker synthesis lays the groundwork for the more advanced multi-speaker approaches that will be explored in depth in the subsequent two chapters, underscoring the evolving complexity and broader applicability of lip-to-speech technology.

The formulation of the problem is modified for the multi-speaker scenario. The goal is to generate a speech segment $S = (S_1, S_2, ..., S_{T'})$ corresponding to a sequence of lip movements $I = (I_1, I_2, ..., I_N)$. Additionally, a speaker identity vector V is taken to generate S in the voice of V.

The approach presented in [55] is adapted, and a speaker embedding is fed as input to the model. The process for generating this speaker embedding is detailed in Section 2.2.2.2.1 from the previous chapter. During training, the speaker embedding V is generated using a $1-$second speech segment from the ground-truth video. This approach enables the unique characteristics of the speaker's voice to be captured by the model. During inference, the flexibility to generate speech in different voices is offered by the model by utilizing a speaker embedding from any arbitrary speech segment. This adaptability

enhances the model's utility for diverse voice synthesis applications. A visual reference of the changed model is provided in Figure 3.6.

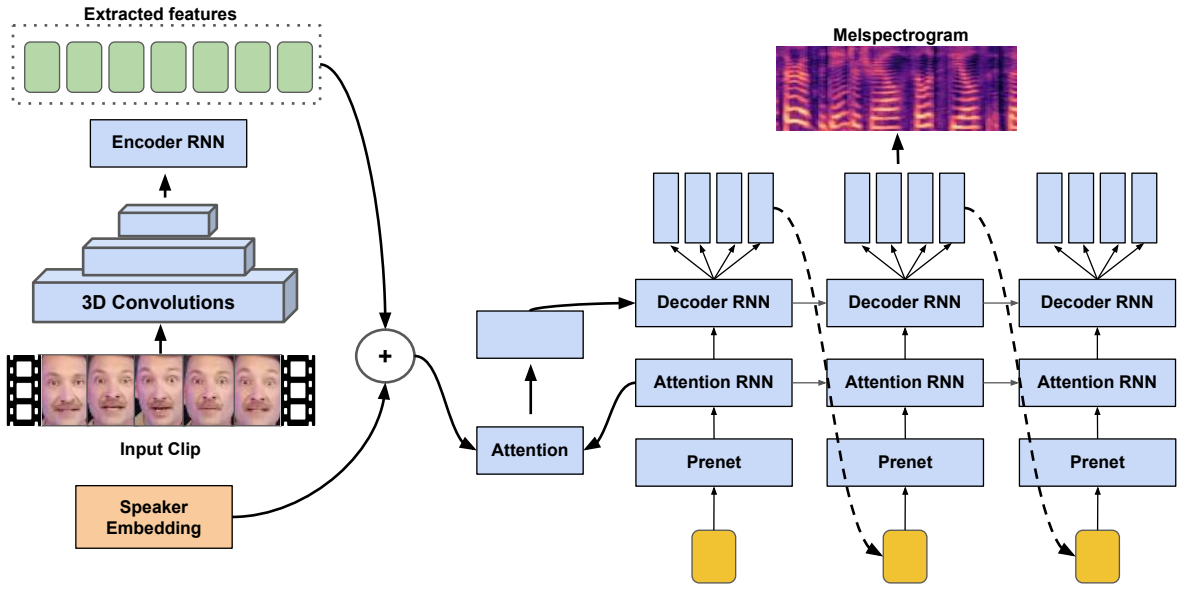

Figure 3.6 Lip2Wav model for the multi-speaker lip-to-speech synthesis scenario. A speaker embedding is added to the network to provide additional target voice information to the network.

Baseline results on the LRW [30] dataset, intended for word-level lip-reading, are reported, i.e., it is used to measure the performance of recognizing a single word in a given short phrase of speech. Demonstration on the LRS2 dataset [28] is not provided as its clean train set contains just 29 hours of data, which is quite small for multi-speaker speech generation. For instance, multi-speaker text-to-speech generation datasets [135] containing a similar number of speakers contain several hundreds of hours of speech data.

In Table 3.14, the speech quality and intelligibility metrics achieved by the multi-speaker Lip2Wav model on the LRW test split are reported. As none of the previous works in lip-to-speech tackle the multi-speaker case, comparisons with them are not made. The WER is also reported by getting the text using the Google ASR API and comparing the WER of the baseline lip-to-text work on LRW [30] as well as a current word level state-of-the-art lip-to-text work [136]. Note that the speech metric scores shown in Table 3.14 for word-level lip-to-speech cannot be directly compared with the single-speaker case, which contains word sequences of various lengths along with pauses and silences.

3.7 Summary

In this study, the challenge of synthesizing speech from lip movements with a focus on single speakers was tackled. The approach hinged on a data-driven learning method, underpinned by the creation

Method	STOI	ESTOI	PESQ	WER
Lip2Wav (Ours)	0.543	0.344	1.197	34.2%
Chung et al. [30]	NA	NA	NA	38.8%
TCN (Current SOTA) [136]	NA	NA	NA	11.64%

Table 3.14 Objective speech quality and intelligibility scores on the LRW dataset. WER is also calculated after using an ASR on the generated speech. Our model outperforms the baseline method proposed in [30] without any text-level supervision. The speech metrics are not applicable for [30, 136] as they are lip-to-text works.

of a comprehensive benchmark dataset tailored for single-speaker, unconstrained, large vocabulary lip-to-speech synthesis. The task was approached as a sequence-to-sequence problem, resulting in speech synthesis that surpasses previous methods in both accuracy and naturalness. The model was rigorously evaluated using a suite of quantitative metrics along with human studies. In the spirit of open research and collaboration, all code and data from the study have been made publicly accessible[2]. This chapter fully focused on single-speaker scenarios, adapting a single-speaker TTS to the lip-to-speech task. More practical scenarios involve multi-speaker tasks. Word-level multi-speaker synthesis was achieved; however, it is shown in the next chapter that this approach does not scale to sentence level, necessitating further work. Therefore, the next two chapters focus on multi-speaker lip-to-speech synthesis. The next chapter addresses the sentence-level version of multi-speaker lip-to-speech, and the chapter after that presents a much-improved version that is practically usable.

Chapter 4

Towards Lip-to-Speech Synthesis for Arbitrary Identities in the Wild

4.1 Introduction

Following the in-depth exploration of single-speaker lip-to-speech synthesis in Chapter 3, the focus of this thesis now shifts to the more complex domain of multi-speaker lip-to-speech synthesis. While the previous chapter laid the groundwork with individual speaker models, this chapter expands the scope to address the challenges of synthesizing speech from lip movements across multiple speakers. It's worth noting that near the conclusion of the last chapter, a preliminary discussion on multi-speaker lip-to-speech at the word level was introduced. Building upon that initial exploration, this chapter presents the first comprehensive approach to multi-speaker lip-to-speech synthesis at a sentence level. The core challenge addressed here is the development of a universal model capable of interpreting and synthesizing speech from a diverse array of lip movements and speech patterns, accommodating the unique characteristics of various individuals. This transition from single to multi-speaker synthesis represents a significant advancement in the field, aiming to create more versatile and broadly applicable lip-to-speech technology.

4.1.1 Lip2Wav fails to learn the attention alignment

Lip2Wav, the sequence-to-sequence model with attention proposed for the *single-speaker* lip-to-speech task in the previous chapter. The attention mechanism lets the decoder look at the correct frame's lip movements while decoding. The model learns diagonal attention upon convergence.

We observed that when training on challenging sentence-level, multi-speaker datasets such as LRS2 [28] with a vast number of voices and vocabulary, it fails to learn the temporal attention alignment. We add the final alignment plots of the trained model in Fig 4.2 to clearly show the failure of this model to learn audio-visual correspondence in such unconstrained settings.

4.1.2 Overcoming Challenges in Lip-to-Speech Synthesis

Building upon the limitations observed in Lip2Wav, particularly its inability to learn proper attention mechanisms for multi-speaker scenarios, the broader challenges inherent in lip-to-speech synthesis are

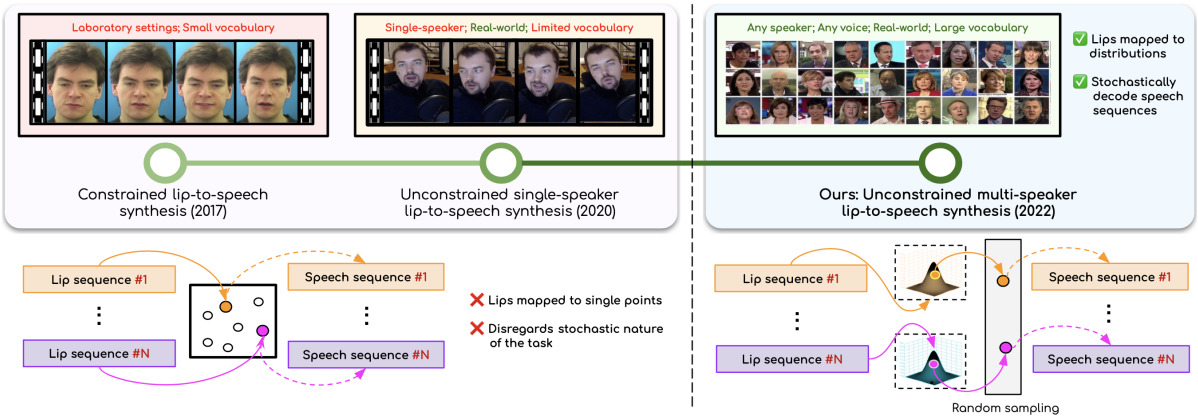

Figure 4.1 The problem of generating speech from silent lip videos for any speaker in the wild is addressed in this work. In previous works, training was conducted either on large amounts of data of isolated speakers or in laboratory settings with a limited vocabulary. Conversely, in this approach, speech can be generated for the lip movements of arbitrary identities in any voice without additional speaker-specific fine-tuning. A new VAE-GAN approach is introduced, which allows strong audio-visual associations to be learned despite the ambiguous nature of the task.

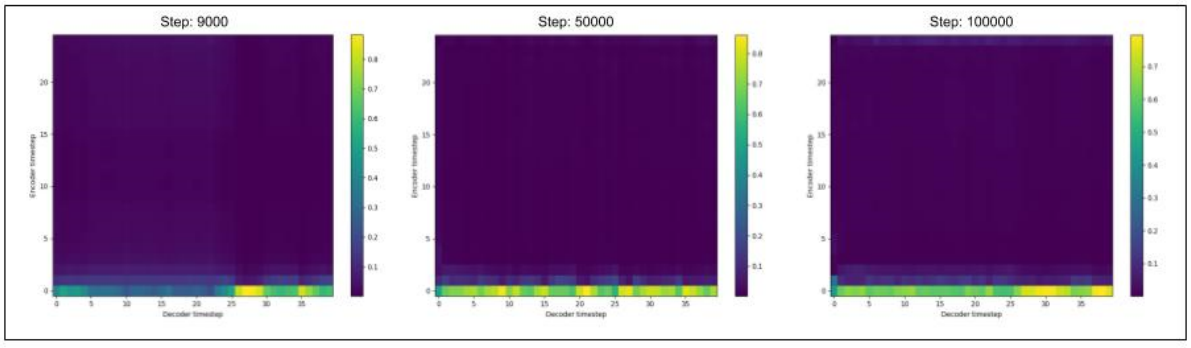

Figure 4.2 Attention alignment plots at various training stages indicating that Lip2Wav fails to learn temporal attention in the unconstrained multi-speaker setting.

now addressed. The failure of Lip2Wav to learn alignment in multi-speaker contexts underscores the complexity of the task at hand. The challenges in lip-to-speech synthesis are multifaceted. Foremost among these is the inherent ambiguity in lip movements, where homophones — different sounds sharing visually similar lip shapes — present a substantial barrier to accurate speech generation. This ambiguity is further compounded in a multi-speaker context, where variations in voices, accents, and speaking styles add layers of complexity. While commendable results are achieved in controlled environments, existing single-speaker methods, including Lip2Wav, fail to address these stochastic variations and scale to the diverse identities encountered in real-world scenarios. In Table 4.1, the task at hand is contrasted against the previous approaches. Most of the earlier works function under one or more constraints. As it will be shown later, these methods do not scale to the case of "generating speech for any identity in any voice". The reasons are discussed, and how the proposed novel approach addresses these issues, overcoming the alignment challenges faced by methods like Lip2Wav, is described.

Table 4.1 Major differences between the proposed approach and the existing approaches. As shown in this table, this work deals with the most challenging task in this space.

Approach	**vocab. size**	**natural setting?**	**training data per spkr (in mins.)**	**zero-shot gen. (unseen spkrs.)**
Vid2Speech [43]	56	×	48	×
Ephrat et.al [44]	82	×	30	×
GAN based [47]	82	×	30	×
Lip2AudSpec [46]	56	×	48	×
Lip2Wav (Presented in Chapter 3)	$\approx 5K$	✓	1200	×
Ours	**50K+**	✓	**3**	✓

This chapter introduces a novel approach to multi-speaker lip-to-speech synthesis. The goal is to generate speech for any identity from silent videos, in any voice, without speaker-specific training. This approach addresses the question: "How accurately can an individual's speech style and content be inferred from their lip movements?"

A VAE-GAN architecture is proposed to capture lip and speech sequence correlations across various speakers. This model is designed to handle the ambiguities in lip-to-speech synthesis and represents a significant advancement in speech synthesis for unconstrained settings.

Key contributions of this chapter include:

- Formulation of lip-to-speech synthesis for unconstrained settings, without limits on speaker numbers or vocabulary.
- Introduction of a VAE-GAN architecture to align speech content with lip movements.
- Establishment of a new benchmark in multi-speaker lip-to-speech synthesis.
- Demonstration of the model's ability to fine-tune to specific speakers efficiently.

The proposed model's performance is evaluated through experiments and compared with existing baselines. This work sets a new standard in the field of multi-speaker lip-to-speech synthesis.

4.2 Essential Background

This section provides a brief overview of key concepts used in the proposed approach.

4.2.1 Variational Autoencoders (VAEs)

VAEs consist of an encoder and a decoder. The encoder maps input data $\mathbf{x}$ to a latent space distribution:

$$\boldsymbol{\mu}, \log(\boldsymbol{\Sigma}) = \text{Encoder}(\mathbf{x}) \tag{4.1}$$

Here, $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ denote the mean and (typically diagonal) covariance of the approximate posterior over the latent variable $\mathbf{z} \in \mathbb{R}^d$.

The decoder reconstructs the input from a latent sample:

$$\hat{\mathbf{x}} = \text{Decoder}(\mathbf{z}), \quad \mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{\epsilon} \tag{4.2}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\hat{\mathbf{x}}$ is the reconstructed version of $\mathbf{x}$.

4.2.2 Generative Adversarial Networks (GANs)

GANs consist of a Generator G and a Discriminator D trained adversarially:

$$\min_G \max_D V(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))] \tag{4.3}$$

Here, $p_{\text{data}}(\mathbf{x})$ denotes the distribution of real data, $p_{\mathbf{z}}(\mathbf{z})$ is a prior distribution over latent noise (e.g., a standard normal), and $V(D, G)$ is the standard GAN value function.

4.2.3 Wasserstein GAN (WGAN)

WGAN uses the Wasserstein distance as the loss function:

$$\min_G \max_{D \in \mathcal{D}} V_W(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})}[D(\mathbf{x})] - \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})}[D(G(\mathbf{z}))] \tag{4.4}$$

where $\mathcal{D}$ is the set of 1-Lipschitz functions. The model (or generator) distribution is denoted by p_g, induced by sampling $\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})$ and mapping through G.

4.2.4 Wasserstein GAN with Gradient Penalty (WGAN-GP)

WGAN-GP adds a gradient penalty term to enforce the Lipschitz constraint:

$$L = V_W(D, G) + \lambda \mathbb{E}_{\hat{\mathbf{x}} \sim p_{\hat{\mathbf{x}}}}[(\|\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}})\|_2 - 1)^2] \tag{4.5}$$

where $\hat{\mathbf{x}}$ is sampled uniformly along straight lines between pairs of points from p_{data} and p_g, $p_{\hat{\mathbf{x}}}$ denotes the distribution of these interpolated samples, and $\lambda > 0$ is the penalty coefficient.

4.2.5 VAE-GAN

VAE-GAN combines VAE and GAN architectures. The objective function is:

$$L_{\text{VAE-GAN}} = L_{\text{VAE}} + \alpha \cdot L_{\text{GAN}} \tag{4.6}$$

where L_{VAE} includes reconstruction and KL divergence losses, L_{GAN} is the adversarial loss, and $\alpha > 0$ is a scalar weight balancing the two terms.

4.2.6 KL Divergence for Multivariate Gaussians

For multivariate Gaussian distributions $p(\mathbf{z}) = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $q(\mathbf{z}) = \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, the KL divergence is:

$$D_{KL}(q(\mathbf{z})\|p(\mathbf{z})) = \frac{1}{2}\left[\log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} + \text{tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) - d\right] \tag{4.7}$$

where d is the dimensionality of the distributions, $|\cdot|$ denotes the determinant, and $\text{tr}(\cdot)$ is the trace of a matrix.

4.3 VAE-GAN architecture for multi-speaker lip-to-speech

4.3.1 Formulating the task

As defined in Chapter 3 (Section 3.6), the goal is to take a sequence of lip movements $I = (I_1, I_2, ..., I_N)$ to generate speech segment $S = (S_1, S_2, ..., S_{T'})$ corresponding to the lip movements I. We also take a speaker identity vector V to generate S in the voice of V. The issues of the previous approaches are examined and appropriate changes to enable learning in a significantly more unconstrained multi-speaker setting are proposed. As mentioned previously in Chapter 3, melspectrograms are used to represent speech. The melspectrogram is represented as $Y = (Y_1, Y_2, ..., Y_T)$.

4.3.2 Fundamental issues in Previous Works

4.3.2.1 Stochastic Nature in Lip-to-Speech Synthesis

All of the previous works aim to map the input lip sequence to a single speech sequence, i.e., they do not account for the stochastic nature of the task. The stochasticity arises due to inadequate priors, i.e., the speech cannot be entirely inferred from the lip movements due to the homophone ambiguity. But additional ambiguities are introduced as the overall set up is changed from laboratory settings to utterances

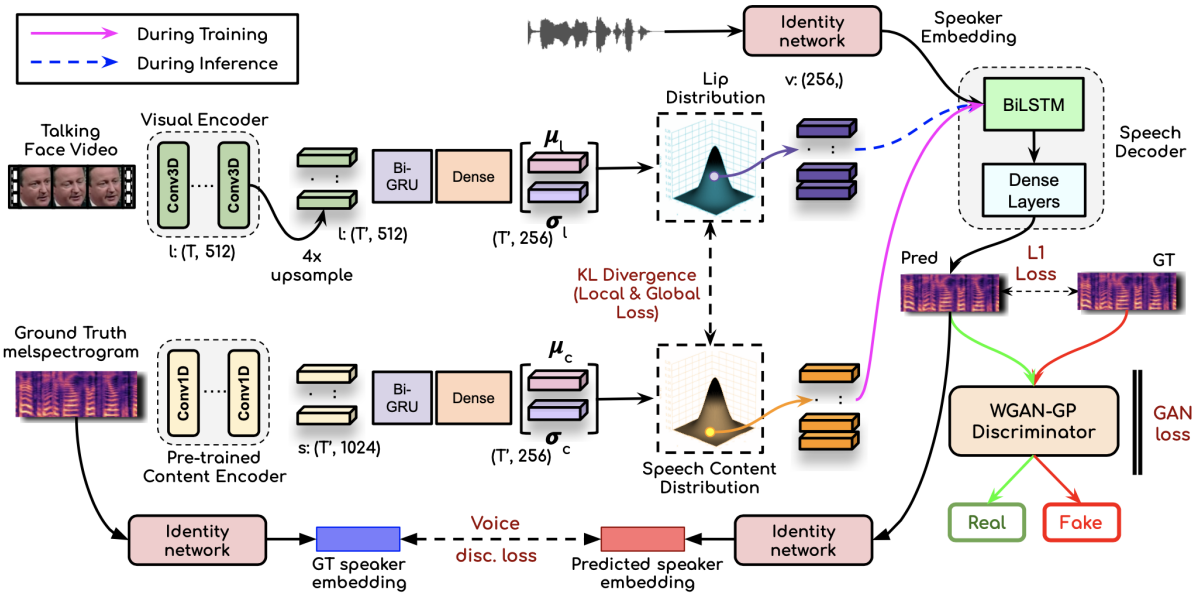

Figure 4.3 A novel VAE-GAN architecture is proposed for this task. Unlike previous approaches that enforce a one-to-one mapping between lip and speech sequences, this model addresses the task's ambiguities by mapping the speech content (ASR representations) and lip sequence to similar distributions. A decoder then generates realistic speech outputs from this latent space. Additional discriminators are used to enable high-fidelity generation in unconstrained settings.

in real-world videos, as shown in Chapter 3, where even the single-speaker case becomes challenging when the speech is "freely uttered": it can have varying decibel levels (no concrete correlation to lips), stress on particular phonemes, and even transient lip motion during pauses.

4.3.2.2 Scaling to Multi-speaker Lip-to-Speech

Moving further into the multi-speaker case, the task becomes severely ill-posed and extremely challenging. Given only the lip movements and a voice token, there are many stochastically varying factors that cannot be inferred from either of the inputs. In addition to the ambiguities mentioned in Section 4.3.2.1, each speaker can have distinct speaking styles and lip shapes in the multi-speaker setting. The large variation in voices and accents also influences how the phonemes are uttered. Such variations cannot be adequately captured in the voice token input. None of the existing models handle these issues, even in the single-speaker case, so they do not scale well to multi-speaker lip-to-speech.

4.3.2.3 What "Space" is Right to Learn these Ambiguous Audio-Visual Correspondences?

In the unconstrained multi-speaker scenario, existing models struggle to learn effectively due to their reliance on learning signals from raw spectrograms. This limitation is attributed to the prevalent use of a *visual encoder - speech decoder* approach with only an L_1 reconstruction loss in most current models. Given the large number of stochastic variations in both visual and speech modalities, it is argued that

learning speech-lip correspondences in the feature space would be beneficial, a concept well-studied in the literature [16, 137–140]. The intuition behind this approach is that low-level variations are more meaningfully represented in the feature space. For instance, matching the lip shapes of "ma" with its instances in speech across different voices would be considerably easier in a feature space that is voice-invariant and contains only the content information from the speech sequences. This intuition serves as the foundation for the core idea presented in this work.

4.3.3 The core idea of this chapter

The core idea presented in this work is two-fold. Firstly, the distributions (Figure 4.1) of (i) the lip sequence and (ii) the content from the speech sequence are matched in the latent feature space to allow the model to handle the aforementioned stochastic variations. Secondly, a decoder is learned that decodes meaningful speech samples from this latent space while also conditioning on a speaker identity embedding that provides the voice information.

Concretely, each input lip sequence is first represented as a distribution (instead of a single vector) and matched to the corresponding speech content distribution. The intuition for matching at the level of distributions is that it allows the ambiguities to be meaningfully represented by allowing a "one-to-many" correspondence. Once such a shared latent space is established, the second step involves decoding samples from this latent distribution to generate meaningful speech.

These ideas are realized in the following manner: A standard automatic speech recognition (ASR) model is used to extract content information from the input speech sequences. A variational auto-encoder [105] is then employed to map the speech content information to a shared latent space and decode the samples from this latent space to real speech sequences. An additional visual encoder is utilized to map the lip sequences to the same shared latent space. These two latent distributions are tied together using the KL divergence loss [141], as illustrated in Figure 4.3. Finally, points are sampled from these distributions and fed to a speech decoder along with the speaker identity embedding to generate intelligible speech sequences. Each of the modules is discussed in detail in the subsequent sections.

4.3.4 Key changes to melspectrogram formation

In the transition from Chapter 3 to this chapter, a significant modification was implemented in the melspectrogram parameters, driven by the need to achieve a more synchronous alignment between audio and video data. In Chapter 3, the audio processing parameters were configured such that one second of audio, equating to 16000 raw wave samples (sampled at 16000 Hz), was converted into a melspectrogram of dimensions 80×80. However, given the video data's frame rate of 25 FPS, this setup resulted in a less optimal relationship between the audio and video frames. To address this issue, key changes were introduced in this chapter: the `n_fft` parameter was increased from 512 to 800, `hop_size` was adjusted from 160 to 200, and `win_size` was altered from 400 to 800. As a result, the melspectrogram for each second of audio in this chapter was resized to 100×80, establishing a more congruent relationship with the 25 FPS video frame rate, where a grouping of 100 frames in the spectrogram corresponds

more effectively to 25 video frames. These adjustments in the audio processing parameters were deemed critical for achieving better synchronization between audio and visual data and were instrumental for the fully convolutional architecture adopted in this chapter. This architectural shift marks a departure from the sequence-to-sequence learning paradigm employed in the previous chapter, underscoring the necessity of such key parameter adaptations to align with the model's evolving structural and functional requisites.

4.3.4.1 Visual Encoder

The visual encoder used in several previous models that aim to learn audio-visual correspondence [15, 18, 120, 142] is adopted. This encoder consists of 3D convolutional layers, with only the first layer having a temporal receptive field of 5 frames. A good trade-off between speed and capturing short-range temporal information is provided by this configuration. A spatio-temporal volume $I : (N, 96, 96, 3)$ is input to the visual encoder, and 1D embeddings $l : (N, 512)$ are output at each time-step. $4\times$ temporal upsampling is performed using nearest-neighbor interpolation on N to match the speech time-steps T. This is made possible due to the reasons mentioned in Section 4.3.4. The visual encoder's architecture is engineered to process the input I through a series of Conv3D blocks, each of which is comprised of a 3D convolutional layer and a batch normalization layer, followed by a ReLU non-linearity. Residual connections are selectively employed in these blocks to facilitate gradient flow and improve training efficiency. A time upsampling layer is also incorporated into the architecture, which quadruples the time dimension from N to T, enhancing the temporal resolution of the data. Following the convolutional stages, the architecture converges into a dense layer. The spatial dimensions are retained by this layer, but the feature depth is transformed to 512 units, effectively consolidating the extracted features into a compact and representative form. The network's final output is l, with dimensions $T \times 512$. This configuration, detailed in Table 4.2, is optimized for robust spatio-temporal data processing, balancing feature extraction, temporal resolution, and feature consolidation.

4.3.4.2 Speech Content Encoder

Lip movements are considered to primarily represent the content information present in a speech sequence. Thus, before matching with the lip distribution, the content from the speech segment needs to be distilled. This is achieved using a standard pre-trained ASR network [54]. The melspectrogram is passed through this frozen encoder to generate a $T \times 1024$ dimensional embedding denoted by c. In this way, the voice information is separated from the speech representations, which is crucial to the training strategy, as will be seen later. The architecture of this content encoder begins with a GRU (Gated Recurrent Unit) layer containing 256 units, which is designed to handle sequential data. It processes an input of size $T \times 1024$ and produces an output of size $T \times 512$. Following the GRU, a Dense layer with 512 units is employed. This layer maintains the input-output size from $T \times 512$ to $T \times 512$, ensuring consistent information flow. The final component is a ReLU activation layer, which applies a

Table 4.2 Architectural details for the visual encoder. Each row represents a layer in the network, showing how the input size is transformed into the output size. Notably, the time upsampler layer upsamples the time dimension from N to $4N$, denoted as T in the presented setup

Layer	Kernel Size	Stride	#Filters	Residual	Input Size	Output Size
Conv3D	(5, 5, 5)	(1, 2, 2)	32	×	N × 96 × 96	N × 48 × 48
Conv3D	(3, 3, 3)	1	32	✓	N × 48 × 48	N × 48 × 48
Conv3D	(1, 3, 3)	(1, 2, 2)	64	×	N × 48 × 48	N × 24 × 24
Conv3D	(1, 3, 3)	1	64	✓	N × 24 × 24	N × 24 × 24
Conv3D	(1, 3, 3)	(1, 2, 2)	128	×	N × 24 × 24	N × 12 × 12
Conv3D	(1, 3, 3)	1	128	✓	N × 12 × 12	N × 12 × 12
Conv3D	(1, 3, 3)	(1, 2, 2)	256	×	N × 12 × 12	N × 6 × 6
Conv3D	(1, 3, 3)	1	256	✓	N × 6 × 6	N × 6 × 6
Conv3D	(1, 3, 3)	(1, 2, 2)	512	×	N × 6 × 6	N × 3 × 3
Conv3D	(1, 3, 3)	1	512	✓	N × 3 × 3	N × 3 × 3
Conv3D	(1, 3, 3)	(1, 3, 3)	512	×	N × 3 × 3	N × 1 × 1
Conv3D	(1, 1, 1)	1	512	×	N × 1 × 1	N × 1 × 1
Time Upsampler	-	-	-	×	N × 1 × 1	4N × 1 × 1
Conv1D	3	1	512	×	N × 1 × 1	T × 1 × 1
Dense	-	-	512	×	T × 1 × 1	T × 512

non-linear transformation to the data without altering its dimensions, keeping the size at $T \times 512$. A tabular overview of this architecture is presented in Table 4.3.

Table 4.3 The architecture for the speech content encoder

Layer	#Units	Input Size	Output Size
GRU	256	$T \times 1024$	$T \times 512$
Dense	512	$T \times 512$	$T \times 512$
ReLU	-	$T \times 512$	$T \times 512$

4.3.4.3 Variational Auto Encoder Based Approach & Latent Distribution Matching

Both lip and speech content embeddings, l and c, are now mapped to Gaussian distributions with a diagonal covariance matrix: $\mathcal{N}(\mu_l, \sigma_l)$ and $\mathcal{N}(\mu_c, \sigma_c)$, where $(\mu_l, \sigma_l), (\mu_c, \sigma_c)$ are obtained using two projection modules P_l and P_c. Both these modules contain a bi-directional GRU [143] followed by a ReLU-activated fully-connected layer. Contextual information in both directions at each time step is captured by the bi-directional GRU. Now that two distributions have been obtained, one corresponding to the speech content and another corresponding to the lips, random points c_p and l_p are sampled from these distributions using the re-parametrization trick [105]. Brief information about the re-parametrization trick is provided in Section 4.2.1. It can be clearly seen that a "single value" for each input lip or speech sequence is no longer present, but rather two probability distributions for these inputs are now available.

The final step is to tie these distributions together, i.e., to ensure that the lip distribution $\mathcal{N}(\mu_l, \sigma_l)$ is close to the speech content distribution $\mathcal{N}(\mu_c, \sigma_c)$. By doing so, a decoder can be trained to decode speech samples from the content distribution $\mathcal{N}(\mu_c, \sigma_c)$ and also to decode from points in the lip distribution. Therefore, the Kullback-Leibler Divergence (KL) loss [141] between these two distributions is minimized.

$$L_{\text{kl}_{\text{global}}} = \frac{1}{N}\sum_{i=1}^{N} KL[\mathcal{N}(\mu_{c,i}, \sigma_{c,i})||\mathcal{N}(\mu_{l,i}, \sigma_{l,i})] \tag{4.8}$$

The term $L_{kl_{global}}$ is referred to as the global KL-divergence loss since the distributions are created by considering the complete speech and lip movements sequence. To further improve the alignment, and inspired by [144], random corresponding temporal segments of the distributions are taken and aligned by minimizing a "local" KL-divergence loss (Figure 4.4). For each batch sample, $R = 10$ small temporal segments are chosen, and their (μ^r, σ^r) are used to minimize Equation 4.9. This approach marks a small but significant contribution to enhancing alignment.

$$L_{\text{kl}_{\text{local}}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{r=1}^{R} KL[\mathcal{N}(\mu^r_{c,i}, \sigma^r_{c,i})||\mathcal{N}(\mu^r_{l,i}, \sigma^r_{l,i})] \tag{4.9}$$

Here, $\mathcal{N}(\mu^r, \sigma^r)$ are r^{th} random patches sampled from the lip and content distributions along the temporal dimension. Binding the distributions at the local level is crucial as phoneme-viseme mappings occur locally rather than globally. The importance of employing both local and global KL-divergence losses shown in Table 4.9. Since the two distributions are aligned using the KL-divergence loss, it is possible to sample from the lip distribution during inference while sampling from the speech content one during training.

4.3.4.4 Speaker Embedding

While the visual/content encoder specifies "what to utter," input for "which voice to utter in" is also needed. Representing each speaker in the dataset with a one-hot vector does not generalize to new speakers during inference. Instead, a recent advancement [55] in training multi-speaker text-to-speech models is adopted, where a pre-trained identity network[1] containing the embedding with voice information is used. The speaker embedding can be obtained for any voice, given just one second of the voice sample. For each video in the dataset, a 256-dimensional speaker embedding is generated using a random one-second segment of the audio. A ReLU-activated fully-connected layer is applied to this pre-trained speaker embedding input V before it is fed to the decoder. For more information about the speaker embedding, please refer to Section 2.2.2.2.1 in Chapter 2.

[1] `github.com/CorentinJ/Real-Time-Voice-Cloning`

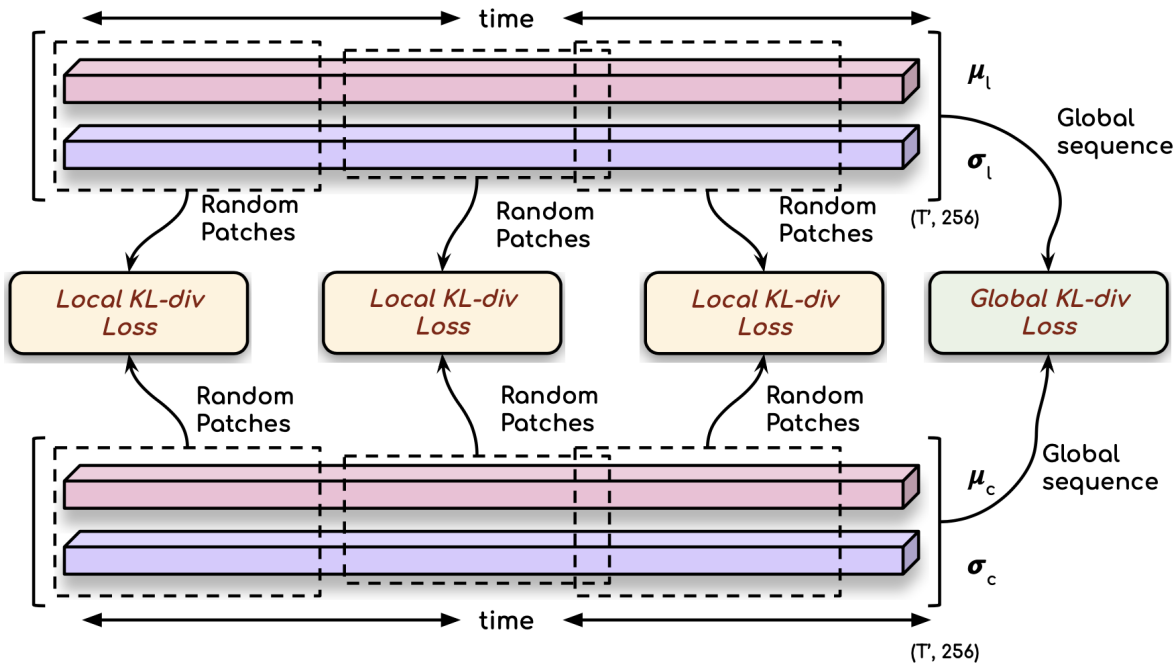

Figure 4.4 In addition to using a global KL-divergence loss to tie the lip and speech content distributions, these distributions are also enforced to be temporally aligned by minimizing a local KL-divergence loss on random smaller time segments. The intuition is that lips and speech are locally aligned in time in the form of visemes and phonemes.

4.3.4.5 Speech Decoder

The final step involves training a module to generate speech segments given points sampled from the previously created joint latent space. At test time, points from the lip distribution need to be fed, as the speech is not available. During training, three ways of sampling the points are utilized: (i) only from the lip distribution, (ii) only from the speech content distribution, and (iii) alternately sampling from both distributions. It is hypothesized that learning with points from (ii) is far easier and allows the network to learn excellent latent representations of the speech content. As the distributions are matched in the latent space, learning accurate, meaningful representations of one of them is quite beneficial for learning the joint space. Indeed, good convergence and intelligible speech, both at training and test time, were observed only when the decoder was trained on points sampled from the speech content distribution. The points are sampled using the re-parameterization trick [105] and are of the shape $(T, 256)$. Along with the sampled points from the content (c_p) or lip distribution (l_p), the speaker embedding input V is ingested by the decoder, which concatenates the sampled points. The concatenated content-voice feature vectors $[(c_p|l_p); V] : (T, 512)$ contain information on both "what to utter" and "the voice to utter in." A bi-directional LSTM layer followed by 3 dense layers is used to decode the melspectrogram segment $(T, 80)$ from the concatenated feature vector. The architecture of the decoder is summarized in Table 4.4.

Table 4.4 Architecture of the speech decoder

Layer	#Units	Input Size	Output Size
LSTM	128 ($\times$2 for bidirectionality)	T x 512	T x 256
Dense - 1st Layer	256	T x 256	T x 256
Dense - 2nd Layer	128	T x 256	T x 128
Dense - 3rd Layer	80	T x 128	T x 80

During training, the generator G ingests speech content c and speaker embedding V and minimizes the L_1 reconstruction loss between the generated speech and the ground-truth speech melspectrogram Y:

$$L_{\mathrm{r}} = \frac{1}{N}\sum_{i=1}^{N} ||G(c_i, V_i) - Y_i||_1 \tag{4.10}$$

Note that during training, the generator network is essentially a VAE for the speech with an additional KL loss constraint on its latent space. Because of the KL loss, sampling from the speech distribution during training is possible, and during inference, when the speech is absent, points can be sampled and decoded from the lip distribution. Therefore, $G(I, V)$ is predicted as the output during inference. Since a content encoder is employed during training, the decoder is forced to condition on the speaker embedding for the voice information. The content encoder distills only the content information from a speech sequence and does not leak the voice information, allowing good voice quality to be maintained even at test times when decoding from the lip distribution. Additional discriminators that will be used along with the generator to improve the quality and accuracy of the generated speech outputs are now described.

4.3.4.6 Enforcing Realism with a VAE + GAN

In the experiments, it was observed that generating realistic samples for such diverse voices, accents, and speaking styles, using a plain L_1 reconstruction loss, produced unrealistic, unintelligible samples (Table 4.8). It is hypothesized that this occurs because of the known issue of the L_1 loss regressing to the mean. Multiple works in the past [145, 146] also point to the benefits of using a GAN along with a VAE. It was found that it is highly beneficial to train a WGAN-GP [113] critic in a GAN setup along with the VAE architecture. The critic consists of a series of 1D convolutional layers that take an audio spectrogram segment of shape $T \times 80$ as input and output a single number as the score. The generator G and the critic D optimize the Wasserstein objective [147] along with the gradient penalty [113] in the equations below, where $\hat{Y}$ contains all linear interpolates between Y and $G(I, V)$:

$$L_{\mathrm{adv}} = \mathbb{E}_{x\sim Y}[D(x)] - \mathbb{E}_{x'\sim G(I,V)}[D(x')] \tag{4.11}$$

$$L_{\mathrm{gp}} = \mathbb{E}_{\hat{x}\sim\hat{Y}}[(\|\nabla_{\hat{x}} D(\hat{x})\| - 1)^2 \tag{4.12}$$

4.3.4.7 Improving Voice of the Generated Speech

To ensure that the model learns the voice and other style attributes, the pre-trained identity network described in Section 4.3.4.4 is used to penalize the generated speech segments if they do not match the voice/style attributes of the ground-truth speech segment. The discriminator network is trained to maximize the cosine similarity L_{voice} between the embeddings of the generated (V_{gen}) and the ground-truth (V_{GT}) speech segments.

$$L_{\text{voice}} = \frac{1}{N}\sum_{i=1}^{N}\frac{V_{gen,i}\cdot V_{GT,i}}{max(\|V_{gen,i}\|_2\cdot\|V_{GT,i}\|_2,\epsilon)} \tag{4.13}$$

4.3.5 Training Settings & Inference

The complete loss function to train the network is the weighted summation of all the above losses:

$$L_{\text{G}} = \lambda_r L_{\text{r}} + \lambda_{k_{global}} L_{\text{KL}_{\text{global}}} + \lambda_{k_{local}} L_{\text{KL}_{\text{local}}} + \lambda_{\text{voice}} L_{\text{voice}} \tag{4.14}$$

$$L_{\text{GAN}} = \min_{G}\max_{D} L_{\text{adv}} + \lambda_{\text{g}} L_{\text{gp}} \tag{4.15}$$

Here, L_{r} is the reconstruction loss, $L_{\text{KL}_{\text{global}}}$ and $L_{\text{KL}_{\text{local}}}$ are the global and local KL divergence terms on the latent distributions, and L_{voice} encourages the correct speaker identity. The term L_{adv} denotes the adversarial (Wasserstein) loss and L_{gp} is the gradient penalty term, with G and D representing the generator and discriminator (critic) networks, respectively. The scalars λ_r, $\lambda_{k_{global}}$, $\lambda_{k_{local}}$, λ_{voice}, and λ_{g} control the relative weight of each loss component.

In the experiments, we set $\lambda_r = 10$, $\lambda_{k_{global}} = 5$, $\lambda_{k_{local}} = 5$, and $\lambda_{voice} = 5$. The pre-processing procedures of Lip2Wav [50] were followed to detect and extract face crops from training videos. Video inputs were created by randomly sampling a window of $N = 25$ contiguous face crops (corresponding to 1 second of video at 25 fps), each resized to 96×96 pixels. The corresponding audio segment was sampled at 16,000 Hz, resulting in $T' = 16{,}000$ waveform samples for 1 second of audio. The short-time Fourier transform (STFT) was computed with a hop length of 10 ms and a window length of 25 ms. From this, melspectrograms with 80 mel bands and $T = 100$ mel time steps per second were obtained.

We used a batch size of 32 and the RMSProp [148] optimizer with an initial learning rate of 0.00005 for both the generator and the discriminator, following the recommendations for WGAN-GP training [113]. The generator was updated once every five discriminator updates, as in [113]. Since this is a WGAN-style setup, the discriminator (critic) loss was used to monitor training progress and was found to correlate well with the perceptual quality of generated samples. Training was stopped once the discriminator loss did not improve for 10 consecutive epochs.

At inference time, the decoder is conditioned on the speaker embedding and the lip latent distribution instead of the content distribution. Because the model accepts a variable number of time steps as input, it can generate speech for arbitrary-length video sequences without any further architectural changes.

4.3.5.1 Datasets and Training Strategy

The primary focus is on synthesizing speech for silent lip videos in unconstrained settings, with the intention to make the model identity-agnostic and capable of working with a larger vocabulary. However, to compare with previous works, the model was also trained on the lab-recorded constrained GRID [60] and TCD-TIMIT [61] datasets. The speaker-independent train-test setting as described by [47] for GRID and the single-speaker lip-to-speech setting as used in the previous chapter by the Lip2Wav model for the TCD-TIMIT dataset were utilized. For unconstrained evaluation, the model was first trained on the word-level LRW data [30]. Next, the complete LRS2 dataset [28] (both train and pre-train sets) was used, which contains sentences and phrases as opposed to specific words. The LRS2 data comprises thousands of speakers from BBC programs with a vocabulary of $59k$ and $2M$ word instances. The large number of speakers and the vast vocabulary covered in both of these datasets encourage the model to be speaker-agnostic and pose no limitations on the vocabulary size.

4.3.5.2 Computation cost

The network is trained using 4 NVIDIA 2080 Ti GPUs. The network consists of 18M parameters and takes 0.5-seconds to generate 1-second of speech.

4.4 Experiments

The model is evaluated against various baseline methods in (i) laboratory-setting videos and (ii) in-the-wild videos.

4.4.1 Evaluation in Constrained Settings

4.4.1.1 Baselines

The model is compared with the following existing lip-to-speech methods: (i) Improved Vid2Speech [44], (ii) GAN-based [47], and (iii) the Lip2Wav model proposed in Chapter 3. Note that the same settings as Lip2Wav for the TCD-TIMIT dataset were used for training, so the paper scores for all the comparison methods are reported. Similarly, the scores from [47] for the GRID dataset (speaker-independent training settings) are used.

4.4.1.2 Metrics

The model is evaluated using the standard speech metrics as discussed in the previous chapter, Section 3.4.4: Perceptual Evaluation of Speech Quality (PESQ) and short-time objective intelligibility measure (STOI). PESQ measures the overall perceptual quality of speech, and STOI correlates with the intelligibility of speech. Additionally, to specifically evaluate the voice quality of the generated sam-

ples, the distance (L_1) between the speaker embeddings of the generated and the ground-truth samples (termed speaker embedding distance (SED)) is measured.

4.4.1.3 Results

Table 4.5 shows the results of different models on GRID and TCD-TIMIT datasets. It is observed that the approach, designed specifically for the unconstrained scenario, performs slightly better or is comparable to other methods when used in constrained settings. Additionally, it can be observed that in terms of voice quality (SED metric), the method outperforms existing approaches, indicating that the voice of the identity is preserved to a large extent.

Table 4.5 Quantitative results on the constrained GRID [60] and TCD-TIMIT [61] datasets.

Dataset	GRID [60]			TCD-TIMIT [61]		
Method	PESQ↑	STOI↑	SED↓	PESQ↑	STOI↑	SED↓
Imp. Vid2Speech [44]	n/a	n/a	n/a	1.23	0.49	n/a
GAN-based [47]	1.24	0.44	n/a	1.22	0.32	n/a
Lip2Wav	1.20	0.38	4.38	1.35	**0.56**	4.64
Ours	**1.28**	**0.45**	**3.76**	**1.35**	0.55	**4.36**

Table 4.6 All models are pre-trained on the LRW dataset and then trained on LRS2. All the comparative methods are outperformed, especially on the challenging LRS2 data, which contains unseen speakers, words, poses and a large vocabulary.

Dataset	LRW [30]							LRS2 [28]						
Method	PESQ↑	STOI↑	SED↓	FDSD↓	KDSD↓	LSE-C↑	LSE-D↓	PESQ↑	STOI↑	SED↓	FDSD↓	KDSD↓	LSE-C↑	LSE-D↓
Imp. Vid2Speech [44]	0.65	0.09	6.01	5.645	10.2	1.782	10.43	0.59	0.30	6.25	4.275	3.1	2.009	8.424
GAN-based [47]	0.72	0.10	5.90	5.189	9.1	1.983	9.426	0.80	0.40	6.13	3.626	1.8	2.503	8.489
Lip2Wav	1.19	0.54	5.73	1.831	1.1	2.526	8.286	0.58	0.28	6.22	10.71	15.5	1.874	11.48
Seq2seq baseline	1.01	0.50	6.16	4.306	7.7	2.396	8.412	0.97	0.43	6.47	3.840	2.8	1.991	8.532
Non seq2seq baseline	1.05	0.49	6.17	4.112	7.1	2.282	8.441	0.96	0.44	6.43	3.803	2.3	2.078	8.536
Ours w/o Content Encoder	0.53	0.14	5.89	2.941	3.4	2.531	8.205	0.45	0.32	6.02	2.856	1.2	2.385	8.230
Lip-to-text [71] + TTS [55]	0.60	0.09	5.91	1.056	0.5	2.181	14.160	0.48	0.11	6.17	0.984	0.1	2.024	19.012
Ours	0.78	0.15	**5.65**	**1.638**	**0.8**	**2.538**	**8.173**	0.60	0.34	**5.95**	**1.273**	**0.2**	**2.507**	**8.155**

4.4.2 Evaluation in Unconstrained Settings

4.4.2.1 Baselines

As no prior works in multi-speaker lip-to-speech synthesis train on such unconstrained datasets, previous models [44, 47, 50] were extended with the same speaker embedding used in the proposed model and all of them were trained on the same dataset as ours. On the LRW dataset, the publicly released multi-speaker Lip2Wav model is evaluated. Additionally, to highlight the importance of the novel modules and facilitate more direct comparison, the following baselines were implemented: (i) a non-sequence-to-sequence encoder-decoder architecture, (ii) a sequence-to-sequence model with only L_1 reconstruction loss and without the VAE-GAN setup, (iii) a standard speech encoder trained from

scratch instead of the pre-trained content encoder, and (iv) Lip-to-text [71] followed by text-to-speech (TTS) [55] model.

4.4.2.2 Additional metrics

Explicitly modeling the problem's stochastic nature is one of the major contributions of this work. Naturally, this allows the model to generate speech samples that differ from the original ground-truth. Thus, along with the standard speech evaluation metrics (PESQ, STOI) and the voice quality metric (SED), which directly evaluate the generated speech against a fixed ground-truth, the model is also evaluated using perceptual metrics. Specifically, following the recent GAN-based TTS systems [149], it is proposed to use: *Frechet DeepSpeech Distance (FDSD)* and *Kernel DeepSpeech Distance (KDSD)* to evaluate the perceptual quality and the linguistic aspect of the generated speech. Note that KDSD scores are multiplied by 10^3 for better readability. Further, it is also evaluated whether the output speech matches the lip movements using LSE-C (measures the confidence of lip-syncing) and LSE-D (measures an embedding level distance between the speech and lip movements) metrics of [27]. The public implementations of these metrics are used for reliable comparison and reproducibility.

4.4.2.3 Results

Table 4.6 compares the model with different methods on the LRW and LRS2 datasets. Existing approaches [44, 47] and the baseline methods are outperformed by a significant margin in perceptual metrics. Thus, although underperformance in standard speech metrics (PESQ and STOI) is observed, it is argued that the method is superior because perceptual metrics are more correlated with human judgment of intelligibility and speech quality. This fact is further supported by providing qualitative results in the demo video on the project website and conducting a human evaluation (Table 4.7). The standard metrics PESQ and STOI enforce one-to-one mapping and thus are not ideal for evaluating this method. Additionally, GRID and TIMIT are constrained datasets with very few variations. On the other hand, LRW and LRS2 are more challenging and unconstrained datasets, and this method is more effective on such challenging data. State-of-the-art perceptual metric scores are achieved on the LRS2 dataset where single-speaker methods such as Lip2Wav fail to learn the audio-visual alignment. The reader is encouraged to view the demo video for qualitative comparisons demonstrating the superiority of this approach.

4.4.2.3.1 Lip-to-text + TTS baseline An additional baseline involves using a state-of-the-art lip-to-text model [71] and converting the predicted text transcripts to speech using a multi-speaker TTS model [55]. This type of baseline was also shown in the previous chapter. From the scores reported in Table 4.6, several deductions can be made. The lip-to-text model trained on text transcripts is naturally far more accurate in predicting the word tokens than any lip-to-speech model, achieving the best results regarding intelligibility and perceptual quality metrics such as FDSD and KDSD. The fact that the lip-

to-speech model comes close to the lip-to-text baseline for the same metrics shows that the approach captures the speech content most accurately.

However, for other metrics like LSE-D that measure if the generated speech is in sync with the video, it is observed that the output of the lip-to-text baseline is not in sync with the lip movements. This issue is critical, as the speech is high quality but not in sync, a key aspect of the task being solved. The same content can be uttered differently (speeds, accent, prosody, voice), and the lip-to-text + TTS baseline cannot capture this. All lip-to-speech models achieve this to different extents, which is an essential condition for the lip-to-speech task. Thus, the approach presented is the best for the task of lip-to-speech synthesis.

To address the synchronization issue, one potential strategy is to condition the TTS on lip movements such that it gets the content from text but lip movements guide it to utter the text at a certain time. This strategy is extensively used in the next chapter to improve multi-speaker lip-to-speech synthesis.

4.4.2.3.2 Qualitative results Qualitative results for the VAE-GAN-based approach are presented through a linked video, in Figure 4.5. This dynamic visualization offers an illustrative demonstration of the research findings, showcasing the effectiveness of the VAE-GAN model in speech synthesis. To fully appreciate the nuances of the synthesized speech, it is recommended to use headphones while viewing the video, as it will allow for a clearer and more detailed auditory experience.

Since the results are ultimately meant for human consumption, providing them in a format that allows for proper judgment of the work's quality is necessary. For the printed version, this area may appear blank. It is requested that the reader check the online version for the final results from this work.

4.4.3 Human Evaluations

Human evaluations were performed with the help of 20 participants. The participant group consisted of members aged $22 - 40$ years with an almost equal male-female ratio. 15 random samples from the LRS2 dataset [28] were chosen, and the results for all the comparison models were generated. Participants rated the speech segments on a scale of $1 - 5$ based on: (A) Intelligibility (is the speech meaningful?), (B) Perceptual Quality, (C) Sync Accuracy (is the generated speech in sync with lip movements?), and (D) Voice Match. The participants' mean scores are summarized in Table 4.7. In line with the quantitative evaluations, the speech generated by the approach is of considerably higher quality and is more legible and natural. A Student's T-Test was also performed for Table 4.7, and the p-value was computed to be ≈ 0.035, indicating that the differences are statistically significant.

4.5 Adapting to Single-Speaker Lip-to-Speech

The previous sections in this chapter discussed the development of a multi-speaker lip-to-speech model, highlighting its benefits and applications. Despite the advantages of a multi-speaker model, some applications still require single-speaker models. Obtaining the necessary 20 hours of single-speaker data

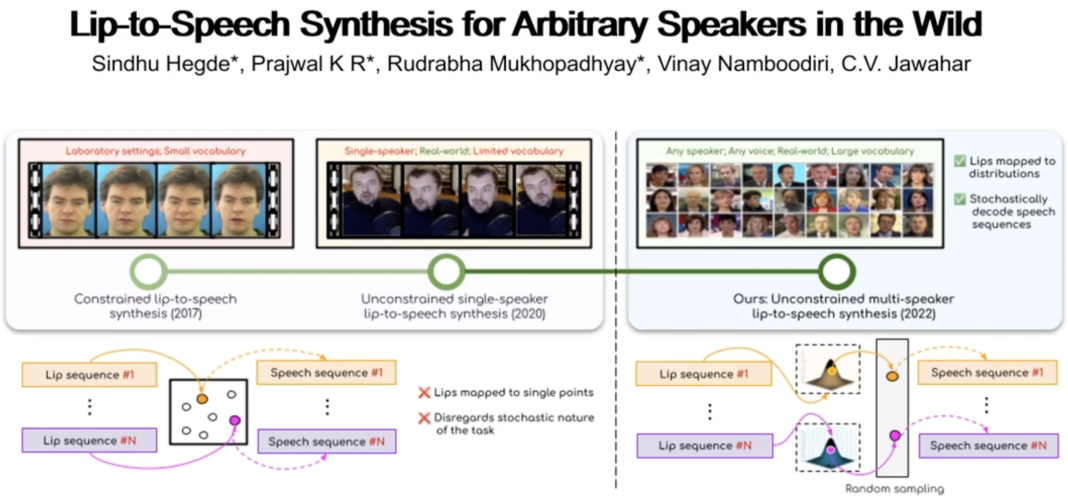

Figure 4.5 A video is presented in this link: `https://youtu.be/iYehW3sd33k`. This video contains qualitative results and comparisons from the proposed VAE-GAN architecture. This image is presented as the thumbnail for this video.

Table 4.7 (A) Intelligibility (is the speech meaningful?), (B) Perceptual Quality, (C) Sync Accuracy, (D) Voice Match. The proposed approach outputs meaningful, intelligible speech that matches lip movements and voice of the target person.

Method	(A)	(B)	(C)	(D)
Imp. Vid2Speech [44]	2.02	1.98	1.74	1.13
WGAN-based [47]	2.17	2.43	2.19	2.01
Lip2Wav	1.07	1.02	1.25	1.03
Seq2seq baseline	1.98	2.10	1.86	1.83
Non seq2seq baseline	2.01	2.23	1.92	1.84
Ours w/o Content Encoder	2.51	2.62	2.01	1.76
Ours	**3.22**	**2.98**	**2.28**	**2.69**

to produce impressive results remains a significant challenge. Therefore, it is imperative to reduce the amount of data needed even for single-speaker models.

The VAE-GAN model proposed in this chapter is capable of generating speech for arbitrary speakers. However, in some cases, obtaining a small amount of data on a target speaker for fine-tuning is possible. Therefore, in this section, it is shown that the pre-trained multi-speaker model can be fine-tuned on a small amount of speaker-specific data to achieve impressive personalized results. By using only 25% of the training data (5 hours), the performance can nearly match that of the single-speaker model trained with 20 hours.

The network was fine-tuned on the speakers in the Lip2Wav dataset. The number of hours in the train set was varied, and the current single-speaker state-of-the-art Lip2Wav model and the fine-tuned multi-speaker model were trained. The variation of the FDSD metric with the training data size is plotted in Figure 4.6. Pre-training on multi-speaker data in the low data regime vastly outperforms the best single-speaker model trained from scratch.

4.6 Ablation Studies

Multiple ablation studies are performed as a part of this work. These are described and listed in this section.

4.6.1 Impact of each discriminator

The final model uses two discriminators: one for enforcing better voice and style attributes and another for enforcing realistic speech. The importance of using each of them is assessed in Table 4.8. It can be observed that, despite achieving a minor improvement in lip-sync metrics, both discriminators enforce better overall speech generation as indicated by the speech metrics.

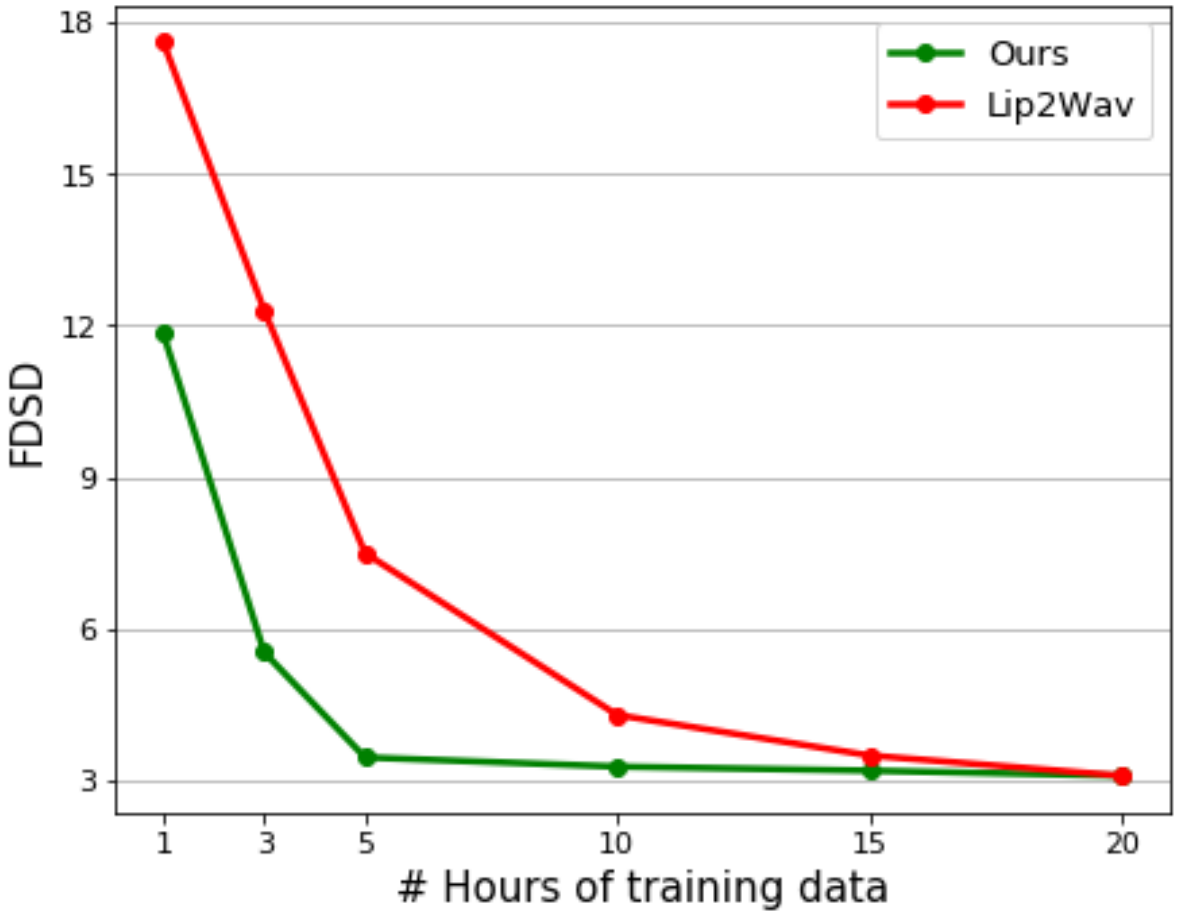

Figure 4.6 Fine-tuning the proposed pre-trained multi-speaker model consistently outperforms the current best single-speaker model (FDSD lower is better) in the low data regime.

4.6.2 Importance of Local and Global Alignment

Table 4.9 shows that optimizing both local and global KL-divergence together improves the alignment between lip and content distributions, thus improving the overall performance. Training with either of the losses in isolation leads to inferior results.

4.6.3 Additional comparisons on LRS3 dataset

The work was compared on the LRS3 [29] test split. The LRS3 dataset, collected from TedX videos, consists of a large vocabulary and mostly profile face videos. Additionally, the videos have very different lighting conditions and extreme head motions compared to the LRS2 [28] dataset. It should be noted that the model was not fine-tuned on the LRS3 dataset, thus evaluating it highlights the model's generalization ability and robustness on completely unseen speakers. As seen in Table 4.10, the method per-

Table 4.8 The discriminators enforce the proposed model to produce meaningful and realistic speech outputs.

Method	FDSD↓	KDSD↓	LSE-C↑	LSE-D↓
Ours w/o both Discs	4.055	2.9	2.188	8.199
Ours w/o WGAN	3.916	2.7	2.294	8.194
Ours w/o Voice Disc	4.310	3.6	2.319	8.189
Ours	**1.273**	**0.2**	**2.507**	**8.155**

Table 4.9 Optimizing both the global and local KL-divergence loss improves the overall quality of the results.

Method	FDSD↓	KDSD↓	LSE-C↑	LSE-D↓
Ours w/o local KL div	2.883	3.2	2.340	8.249
Ours w/o global KL div	5.040	6.8	2.003	8.937
Ours	**1.273**	**0.2**	**2.507**	**8.155**

forms well on LRS3 videos containing unseen vocabulary, identities, voices, and profile views, clearly indicating the robustness of the approach.

Table 4.10 Quantitative comparison on the LRS3 dataset [29]. It can be seen that all competitive methods are outperformed, even in the very different setting of LRS3, which contains unseen speakers, words, and a large number of profile views. Note that the model is not fine-tuned on the LRS3 dataset.

Dataset	LRS3 [29]			
Method	FDSD↓	KDSD↓	LSE-C↑	LSE-D↓
Imp. Vid2Speech [44]	5.286	5.4	1.929	8.435
GAN-based [47]	4.589	4.2	2.031	9.372
Lip2Wav	12.663	16.2	1.632	11.465
Seq2seq baseline	4.821	4.8	1.880	8.568
Non seq2seq baseline	4.766	4.2	1.874	8.579
Ours w/o Content Encoder	4.054	2.9	2.041	8.312
Ours	**3.148**	**1.8**	**2.063**	**8.256**

4.6.4 Near Frontal vs. Non-frontal Videos

The extent to which the performance deteriorates if the face view moves towards a non-frontal profile view is studied in this section. As expected and also observed in past lip-reading works [86], room for improvement when handling non-frontal talking faces is noted for the proposed VAE-GAN model.

4.6.5 What Kind of Visual Input is the best?

Different forms of visual inputs, such as feeding only the lower half of the face or even pre-trained face embeddings [69], are compared. As shown in Table 4.12 and reflected by activation maps near the eye regions in Figure 4.8, providing the full face crop performs the best. Another pre-trained

Table 4.11 Comparison of performance between frontal and non-frontal views. Similar to other lip-reading models [86], a drop in performance is observed in non-frontal views, indicating room for improvement in handling such cases.

Method	LSE-C↑	LSE-D↓	FDSD↓	KDSD↓
Near frontal	2.608	8.016	1.351	0.3
Non-frontal	2.491	8.058	3.473	1.0

transformer-based embedding [150] was tried, but no legible results were obtained. Therefore, the results are not discussed.

Table 4.12 Feeding the full face crop produces the best results.

Method	LSE-C↑	LSE-D↓	FDSD↓	KDSD↓
Facenet emb [69]	1.931	**7.664**	8.641	8.7
Lower half (ours)	2.338	8.173	3.224	2.8
Full face (ours)	**2.507**	8.155	**1.273**	**0.2**

4.6.6 Sampling strategy of VAE at train-time

In this work, it has been previously mentioned in Section 4.3.4.5 that there are three ways of sampling the points to decode from during the training: (i) only from the lip distribution, (ii) only from the speech content distribution, and (iii) alternately sample from both the distributions. It was stated that good convergence and intelligible results are obtained only by following (ii) because it allows the network to learn excellent latent representations of the speech, which can help overall learning. This is experimentally verified in Table 4.13 by demonstrating that sampling from other distributions, i.e., (i) Lip distributions and (iii) Alternately sampling from both distributions, leads to far worse results.

Table 4.13 Sampling solely from the speech distribution during training enables the decoder to learn to generate realistic, accurate outputs.

Method	LSE-C↑	LSE-D↓	FDSD↓	KDSD↓
Lip dist.	2.261	8.195	4.036	2.7
Speech content dist.	**2.507**	**8.155**	**1.273**	**0.2**
Alternate sampling	2.241	8.320	3.487	2.4

4.6.7 Auto-encoder vs. VAE

The importance of mapping the lip and content distributions using a VAE is assessed. In Table 4.14, it is shown that the removal of the variational aspect and the use of a naive auto-encoder approach results in poor speech generation. Interestingly, while the network learns comparable audio-visual correspondence, the speech is found to be neither intelligible nor meaningful, as indicated by the speech metrics.

Table 4.14 Using a VAE enables the model to generate meaningful, high-fidelity speech outputs.

Method	FDSD↓	KDSD↓	LSE-C↑	LSE-D↓
Auto-encoder	6.015	5.3	2.002	8.431
Ours (VAE)	**1.273**	**0.2**	**2.507**	**8.155**

4.6.8 Model's variation across speaker attributes

In Table 4.15, the performance of the proposed model is evaluated across the gender of the identities. The LRS2 test set is automatically classified into male and female speakers using a gender detection tool [151]. From the table, it can be clearly observed that there is no significant variation in performance across the genders of the identities.

Table 4.15 There is no distinctive variation of performance across the genders of the speakers.

Gender	LSE-C↑	LSE-D↓	FDSD↓	KDSD↓
Female	2.549	8.138	1.633	0.8
Male	2.424	8.233	1.703	0.8

4.6.9 Generative Strength of the proposed Lip-to-Speech Model

Lip-to-speech synthesis is a highly ambiguous task, with multiple possible speech outputs for the same input lip sequence. Variations in voice, speech amplitude, intonation, prosody, and emotion are not clearly correlated with lip movements. Additionally, the content to be generated is also ambiguous due to the presence of homophones. The VAE-GAN model presented is the first architecture capable of modeling these variations in the latent space. Different output speech sequences can be generated for the same input lip sequence by sampling different points from the lip distribution.

The evaluation of the generative capabilities of a VAE and GAN has been explored in previous work [152], and the same metric is adopted here. The "Generative Strength" metric is used to determine the average percentage of unique speech samples generated for every input lip sequence. To compute this metric, N points ($N = 100, 300, 500$) are sampled for each input lip video from the test set of LRS2, resulting in N speech outputs for every LRS2 test video. A generated spectrogram output is considered "unique" if its L_2 distance from the remaining $N - 1$ spectrograms is at least $\delta = 0.5$. It is noted that this is a sufficiently high L_2 threshold, as the generated audio samples are clearly distinct to hear.

In Figure 4.7, the average count of unique speech outputs per input lip video is plotted at different stages of the model training. It is observed that the model captures more variations for the same input lip sequence as the training progresses.

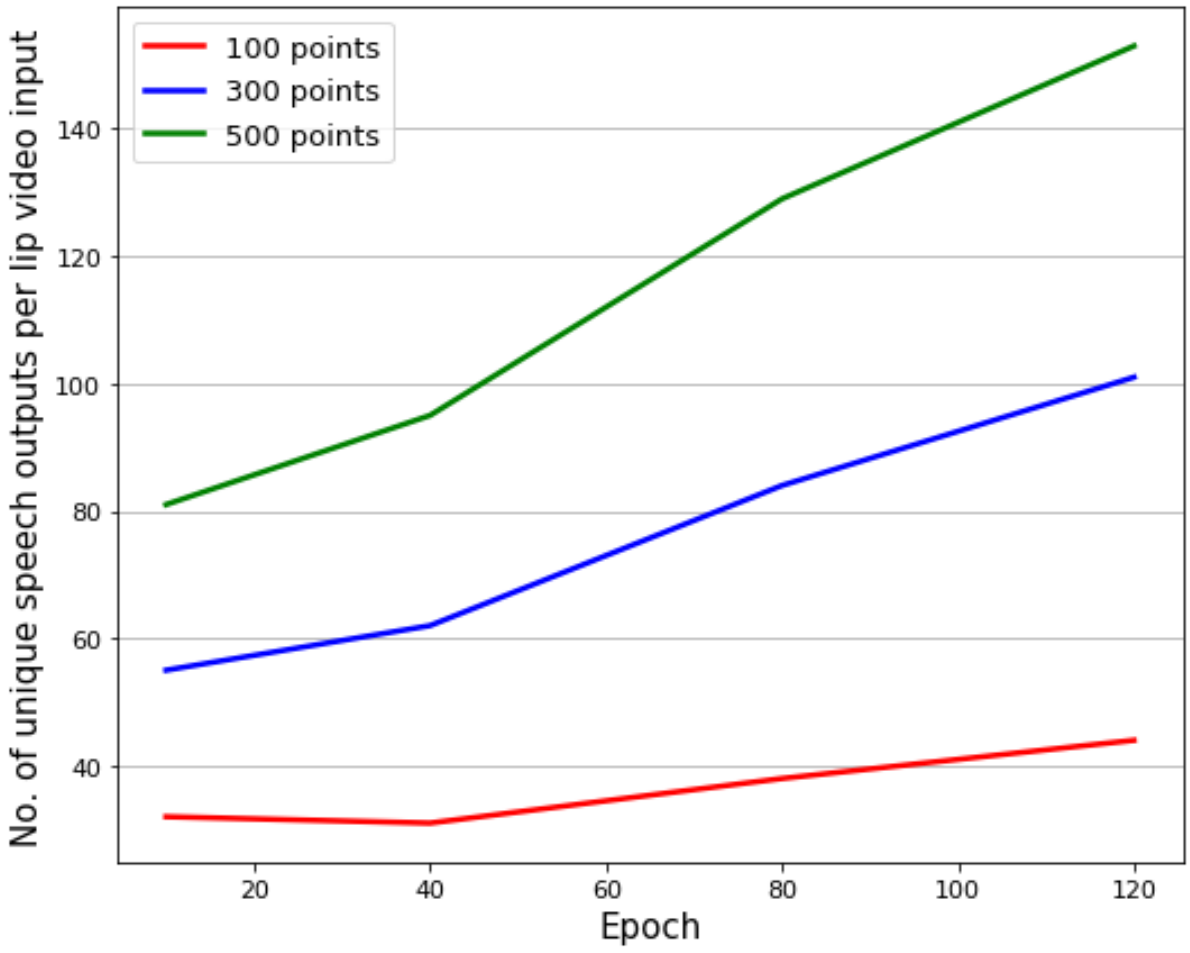

Figure 4.7 The average number of unique speech outputs generated by the model for each input lip video is plotted. This ”generative strength” [152] is shown at different stages of training. It can be observed that as the training progresses, the model captures more variations in the latent space, indicating an increase in the diversity of generated outputs over time.

Figure 4.8 Activation maps of the visual encoder shows that the model strongly attends to the lip region while generating speech, despite variations in head pose and lip location.

4.6.10 Where does the model attend in the visual front-end?

Figure 4.8 contains the plotted activation maps from the visual encoder to highlight that the model predominantly attends to the lip region.

4.6.11 Plotting the distributions

Multiple instances of different words from both the content and lip distributions were plotted. Embeddings from videos and audio for particular instances of different words were taken and then used to create a 2-dimensional TSNE plot. The plot is shown in Figure 4.9. Homophones were also plotted in the same graph. It was observed that the content and lip distributions for particular words are closer to each other. Additionally, homophones like ”Million” and ”Billion” were found to lie close to each other, which aligns with the intended approach.

4.7 Summary

In this chapter of the thesis, the problem of unconstrained lip-to-speech synthesis is explored for the first time. The complexities inherent to this task are examined, and the reasons why existing methods struggle in such unconstrained environments are discussed. A VAE-GAN model is proposed as a solution specifically designed to handle the stochastic nature of lip-to-speech synthesis. The proposed approach maps both lip movements and speech content into distributions during training. A KL-divergence loss between these distributions is employed to tie them together. During inference, only the lip movements are available, resulting in only the lip distribution being used. This variational step in the pipeline allows for the handling of randomness inherent in the task. Through extensive ablation studies, the components of the architecture are validated, demonstrating that the model produces speech outputs that are more intelligible and realistic compared to existing models. However, significant limitations are noted, particularly in the quality of the synthesized speech. Despite advancements, the quality often falls short of expectations. A critical aspect of evaluating the model’s effectiveness involves comparing its output with the ground truth (GT) audio. This comparative requirement highlights the need for further improvements in speech synthesis quality. It is observed that while the model was able to generate speech that sounded like the ground truth, it was still considerably different from it. An

Automatic Speech Recognition (ASR) test on the generated speech resulted in around 90% Word Error Rate (WER), rendering it virtually unusable in practical applications. Recognizing these limitations, alternative approaches are considered, taking into account recent advancements in related fields. It is noted that lip-to-text technologies have progressed significantly in parallel, and text-to-speech (TTS) systems continue to improve. The most straightforward solution to the lip-to-speech problem might appear to be a combination of lip-to-text followed by text-to-speech. However, as discussed in the previous chapter (refer to Chapter 3), this approach presents a critical issue: the generated speech would not be synchronized with the video. A novel approach is proposed in the next chapter to address this synchronization challenge while leveraging the advancements in both lip-to-text and TTS technologies. This approach introduces a visually conditioned TTS (VTTS) system that is specifically designed to condition on lip movements. By implementing a two-step process of lip-to-text followed by VTTS, it is shown that both the quality of speech generation and its synchronization with the visual input can be significantly improved. This innovative method aims to combine the accuracy of lip-to-text conversion with the flexibility and naturalness of advanced TTS systems, all while maintaining crucial temporal alignment with the original lip movements.

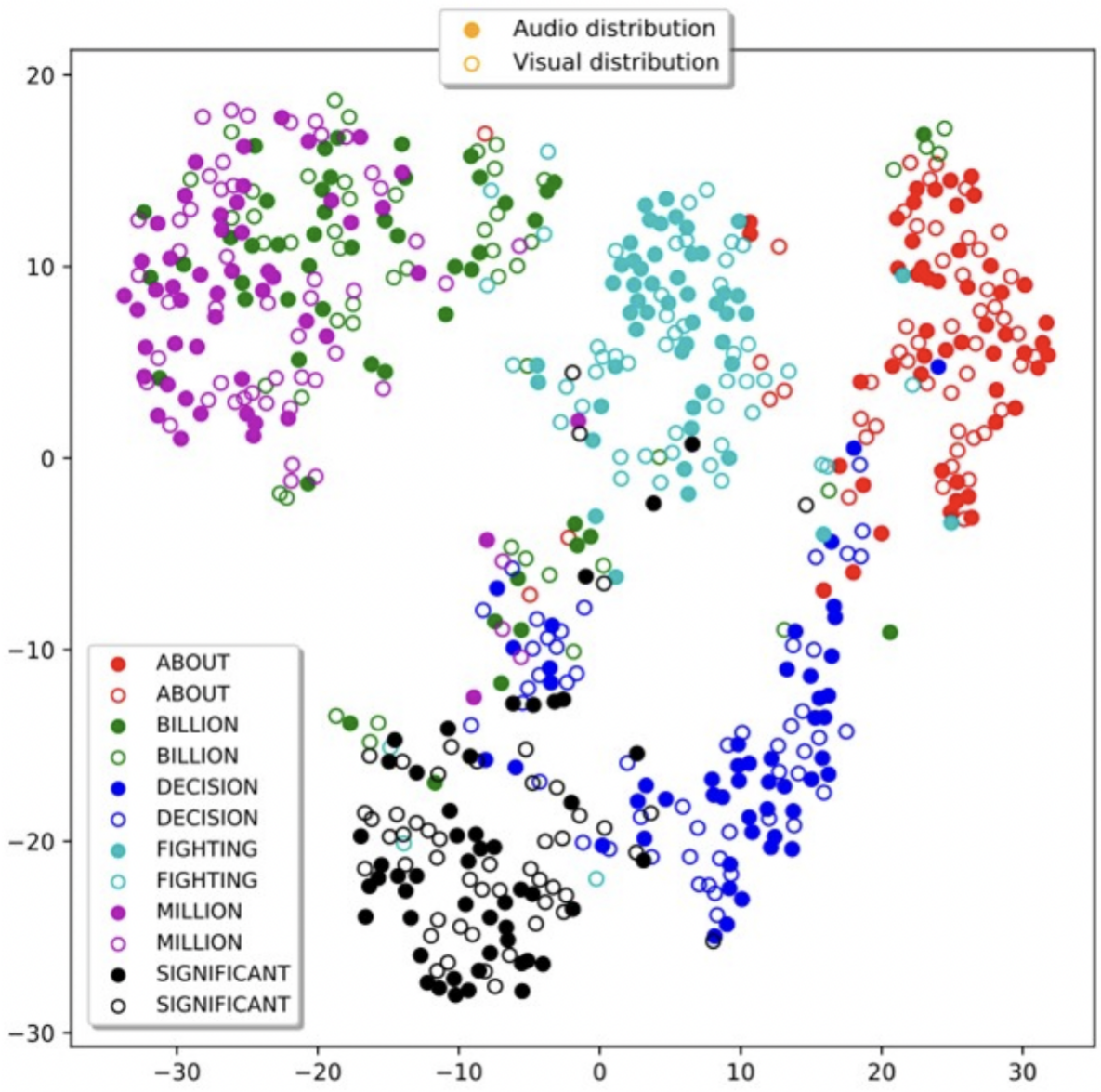

Figure 4.9 TSNE plot of multiple instances of different words from both content and lip distributions, including homophones. The plot shows that content and lip distributions for particular words are close to each other. Additionally, homophones like ”Million” and ”Billion” lie close to each other, demonstrating the approach’s effectiveness.

Chapter 5

Accurate Lip-to-Speech Synthesis for arbitrary identities in the wild

5.1 Introduction

The development of multi-speaker lip-to-speech synthesis is an important step towards practical audio-visual communication systems. Such models are relevant in several realistic settings, for example as a component in assistive communication devices, as a tool for recovering missing or corrupted audio in constrained forensic scenarios, or as a fallback mechanism in telecommunication and video conferencing when the audio stream is degraded but the video is still available. In all of these cases, it is desirable to handle multiple speakers without requiring a large amount of speaker-specific training data, and to produce speech that is intelligible and reasonably consistent with the original speaker's voice and articulation, rather than aiming to "solve" any of these application domains outright. In the previous chapter, we took a first step towards this goal by extending lip-to-speech synthesis from a single speaker to the multi-speaker setting. There, we used a VAE–GAN style formulation to jointly model a latent distribution over lip movements and the corresponding speech, with the aim of capturing a shared representation that can generalise across speakers. While this approach reduced some of the ambiguity in the mapping from lips to speech, the supervision signal still came solely from the raw speech waveform. As a result, many factors such as voice characteristics, accents, and prosody remained entangled with linguistic content, and the overall quality and stability of the generated speech were still limited.

In this chapter, we build on that multi-speaker framework and introduce an additional level of structure by explicitly separating content and style. Noting the significant progress made by recent lip-reading models, which benefit from large-scale datasets and strong language modelling components, we propose to use lip-to-text generation as an intermediate step. Concretely, we first employ a pre-trained lip-to-text model to obtain noisy text transcriptions from the input lip video. A visual text-to-speech (TTS) system is then conditioned on three elements: the generated text to specify linguistic content, the lip movements to guide timing and some aspects of prosody, and a speaker embedding to control voice characteristics. This design aims to reduce the burden on the model to implicitly learn a language model from raw audio alone, and instead leverages explicit textual supervision while still retaining visual and speaker cues. By combining the strengths of advanced lip-reading models with a carefully conditioned multi-speaker TTS system, this chapter presents a more structured approach to multi-speaker lip-to-

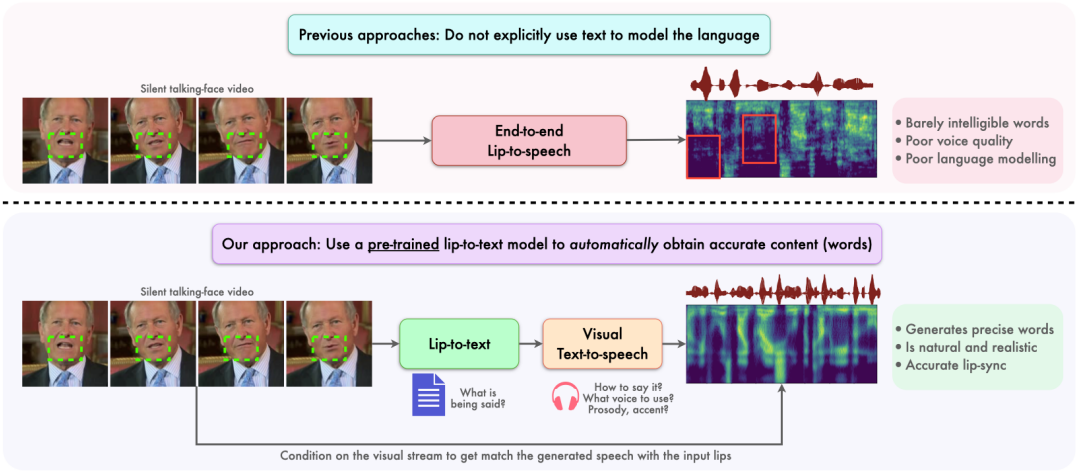

Figure 5.1 Overview of the proposed multi-speaker lip-to-speech system. Instead of learning a language model directly from raw speech, which provides only weak supervision due to acoustic variability (voice, accents, prosody), we leverage recent lip-to-text models to obtain noisy text transcriptions and condition a visual TTS network on both the text and the lip video.

speech synthesis. The focus is on empirically improving robustness and intelligibility over the previous VAE–GAN-based method, rather than on claiming a complete solution to all potential application areas.

5.1.1 Contributions

In this chapter, a novel approach is introduced that integrates noisy text supervision into the multi-speaker lip-to-speech synthesis process. A state-of-the-art lip-to-text network is used to generate text from lip movements, and a visual text-to-speech network then uses this text to produce speech that is more accurate in terms of content while remaining aligned with the lip video. This intermediate textual representation improves content distillation and makes it easier to incorporate linguistic structure, while the visual and speaker conditioning help to maintain reasonable prosody and speaker identity. The contributions of this chapter are threefold:

1. We analyse the limitations of relying solely on speech supervision in multi-speaker lip-to-speech generation, as in the previous VAE–GAN-based chapter, and motivate the need for an explicit textual intermediate representation.

2. We detail a method that uses noisy text transcriptions from a pre-trained lip-to-text model to guide speech synthesis, conditioning a visual TTS system jointly on text, lip movements, and a speaker embedding to better disentangle content from speaker- and style-related factors.

3. We demonstrate, through experiments and ablation studies, that the proposed approach improves over the previous multi-speaker baseline in terms of intelligibility and alignment, and discuss its applicability in realistic assistive and communication-oriented scenarios.

In the following sections, each of these contributions is delved into. This exploration highlights the potential of multimodal learning in deep learning and its significant impact on enhancing human lives.

5.2 Issues and challenges in existing works

The challenges present in standard lip-to-speech systems, similar to those discussed in Chapters 3 and 4, are first discussed.

5.2.1 Learning language from speech

As previously discussed, speech is directly generated from lips in all current lip-to-speech works. It is known that learning a language model is crucial for accurately reading lips. However, current multi-speaker lip-to-speech models are sub-par because a language model is attempted to be learned in the speech modality, which contains a large diversity of speaker identities, styles, accents, and prosody. Thus, it is argued that some other way of incorporating language knowledge is needed.

5.2.2 The missing block of lip-to-speech: lip-to-text

Two tasks are involved in Lip-to-Speech synthesis models: (i) inferring the content from lips and (ii) inferring the style in which that content is spoken. If the content being spoken is known, then the task is reduced to just generating speech that matches the silent lip video. This is the premise of this chapter. But how can this text information be obtained, especially when only a silent lip video is available as input? It is shown that this text information can be obtained from pre-trained lip-to-text models: a class of models closely related to the task at hand but largely ignored in previous works on lip-to-speech synthesis. An approach is designed that can build upon current works [71, 86] in lip reading, i.e., pre-trained lip-to-text models, to generate far more accurate speech outputs.

5.2.3 Achieving accurate lip-sync

Now that the text is available, the next step is to generate the ”right” kind of speech output. That is, the generated speech must match the input lip sequence. It is noted that a sentence can be uttered in many ways, but only one will match the input lip video. Thus, it is worth noting that a trivial text-to-speech will not serve this task. Instead, a text-to-speech model that is also conditioned on video input is needed.

5.3 Essential Background

Before delving into the proposed approach, it is necessary to provide a concise overview of the key concepts and technologies that underpin this work. This section aims to familiarize the reader with the essential background knowledge required to fully comprehend the subsequent discussions.

5.3.1 Pre-trained Lip-to-Text

As detailed in Chapter 2, an extensive background on lip-to-text technology has been provided. In this context, the focus is placed on the state-of-the-art subword level lip reading model [71], a prominent advancement in the transcription of silent talking face videos. Significant progress in transcribing spoken content from visual cues with remarkable accuracy is exemplified by this model. Two primary attributes by which this model is distinguished are: its data efficiency, essential for training on publicly available datasets, and its robust visual backbone, adept at accurately extracting lip features. The model's effectiveness is not confined to lip reading; utility in related tasks such as visual keyword spotting [153] and identifying mouth movements in sign language [154] has also been demonstrated. The adoption of sub-word units for text representation is particularly advantageous, as a more refined approach to managing the inherent ambiguities of lip reading than traditional character-based models is offered. A language prior is provided by these semantically meaningful sub-word tokens, enhancing overall performance. Moreover, the model's visual representations, called the Visual Transformer Pooling (VTP) features, skillfully track and aggregate spatio-temporal features of lip movements, largely due to its sophisticated attention-based pooling mechanism. Therefore, this model has been selected for generating text and visual features from silent lip videos in this research.

5.3.2 A transformer-based TTS model: The FastSpeech2

To contextualize the proposed approach, we briefly summarize FastSpeech2 [59], which we use as the backbone for our modified text-to-speech (TTS) module. The model is built around a feed-forward Transformer block that combines multi-head self-attention with 1D convolutions over a sequence of phoneme embeddings, transforming them into a hidden representation suitable for speech generation. A key component is the variance adaptor, which augments this hidden sequence with explicit duration, pitch, and energy information. The duration predictor, trained with an MSE loss, estimates how long each phoneme should be voiced; phoneme-level durations are obtained from forced alignment using the Montreal Forced Aligner [155], reducing the mismatch between input and output timing. For pitch, instead of predicting the raw contour directly, the model applies a continuous wavelet transform (CWT) [156] to obtain a smoother pitch spectrogram as the prediction target. Energy is computed at the frame level from the melspectrogram, quantized, and then injected into the hidden sequence. Both the pitch and energy predictors share the same structure: a two-layer 1D convolutional network with ReLU activations, followed by layer normalization, dropout, and a final linear layer. By explicitly modelling duration, pitch, and energy, the variance adaptor helps reduce the one-to-many mapping problem in TTS and offers better control over prosody. This setup forms a practical and well-tested starting point, which is built on in this chapter for our visual (lip-conditioned) speech synthesis experiments.

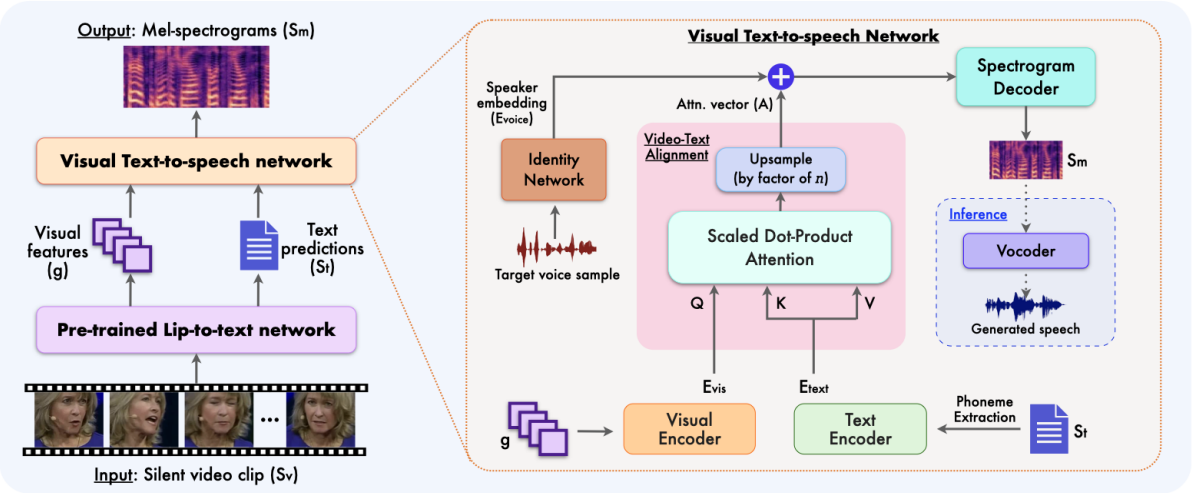

Figure 5.2 An overview of the proposed approach is presented. Visual features and text predictions are first extracted from a pre-trained lip-to-text network. Speech outputs that synchronize with the silent video input are then generated using a visual text-to-speech (TTS) model. The visual and textual (in the form of phonemes) inputs are encoded and aligned in time by the visual TTS using scaled dot-product attention. For each query video time-step, the phoneme to be uttered at that time is retrieved using this attention mechanism. After the addition of the speaker identity embedding, these are upsampled and decoded into melspectrograms. Finally, the melspectrograms are converted into natural waveforms using a pre-trained vocoder.

5.4 The proposed approach

As defined in the previous two chapters, the goal is to take a sequence of lip movements $I = (I_1, I_2, ..., I_N)$ to generate a speech segment $S = (S_1, S_2, ..., S_{T'})$ corresponding to the lip movements I. A speaker identity vector V is also taken to generate S in the voice of V. The discussion is started by examining the issues in previous methods, and appropriate changes are proposed to enable learning in a significantly more unconstrained multi-speaker setting. As mentioned previously in Chapter 3, mel-spectrograms are used to represent speech. The melspectrogram is represented as $Y = (Y_1, Y_2, ..., Y_T)$. A state-of-the-art lip-to-text model [71] is used to obtain the lip features and noisy text transcriptions for the given silent lip video $W = W_1, W_2, ..., W_{N'}$. A visual text-to-speech model is designed, conditioned on (i) the noisy text and (ii) the lip features, to produce high-quality speech outputs Y that are in sync with the input silent lip video. This solves the task of lip-to-speech synthesis. An overview of the proposed two-stage framework is depicted in Figure 5.2.

5.4.1 Adopting VTP for the lip-to-speech task

The lip-reading model employed processes a sequence of lip movements denoted as $I = (I_1, I_2, ..., I_N)$, where each individual frame is initially passed through a spatio-temporal residual CNN block. Foundational visual features from each frame are extracted by this block. Following this initial feature extraction, these frame-wise features are further refined in a visual transformer pooling block. The fea-

tures are enhanced through a self-attention mechanism by this block, comprised of several transformer layers. The output of this processing is a self-attended feature map, represented as z_N. A spatially weighted average of the map itself is resulted when this feature map z_N is integrated with a learnable query vector Q_{att}. A series of compact, per-frame visual representations is the outcome of this integration. These representations are then methodically arranged along the temporal dimension to form a temporal embedding sequence g, expressed as $g \in \mathbb{R}^{N \times fd}$, where fd represents the dimensionality of the transformer feature space.

The visual representations encapsulated in g are then fed into a transformer encoder-decoder network. The textual output in the form of sub-word tokens is predicted by this network, adhering to an auto-regressive model. To enhance the accuracy and coherence of the final output, advanced decoding strategies are employed by the model, including beam search [157] and language model rescoring [158], culminating in the final sentence outputs $W = W_1, W_2, ..., W_{N'}$, with N' denoting the number of words in the sentence.

The architectural specifics of the VTP network have not been delved into deeply in this discussion, as it is not one of the primary contributions. Instead, this network is employed off-the-shelf, focusing on how its outputs integrate with and enhance the proposed model. This approach allows concentration on the innovative aspects of the work, particularly the synthesis of speech that is in synchrony with lip movements, leveraging the strengths of the VTP network to bolster the system's performance.

This module operates on feature maps extracted from the video frames by a spatio-temporal residual 3D CNN encoder. The process begins with sub-clips of the video, each consisting of 5 frames, being transformed into spatial feature maps by the CNN. These feature maps are then individually processed by the VTP block for every frame. The overall structure of the 3D CNN encoder is given in Table 5.1. It should be noted that each Conv3D block consists of a Conv3D layer followed by a batch normalization layer. A residual connection is added inside the block optionally. The final output of the block is passed through a ReLU activation function.

In the VTP module, each feature map undergoes a flattening and projection step to align it with a predefined Transformer feature dimension. This step is crucial as the feature map is prepared for further processing. Once the feature map is in the correct format, spatial positional encodings are added. These encodings are essential for the Transformer to understand the spatial relationships within the data. The core of the VTP module is an encoder composed of several Transformer layers. The feature map is enhanced by this encoder by applying self-attention mechanisms, a process that allows the model to focus on the most relevant parts of the feature map for each frame. After encoding, a key step involves the use of a learnable query vector. A visual attention mask is produced by the interaction of this vector with the encoded feature map. This mask essentially highlights the areas of each frame that are most important. The final step in the VTP module is the aggregation of these attention-weighted features. This is achieved through a weighted average calculation, which condenses the information from each frame into a compact representation. By stacking these representations over time, a comprehensive spatio-temporal embedding for each frame in the video is created by the VTP module. This embedding

captures the essential visual characteristics of the video, making it a powerful tool for further analysis or processing in the pipeline. For those interested in a more detailed exploration of the model's architecture and its specific training strategies, a comprehensive description is available in [71].

In the work presented in this chapter, the pre-trained VTP (Visual Text Prediction) network is utilized, primarily for two purposes: (i) to generate text predictions, denoted as the final decoded output W of the model, and (ii) to acquire per-frame visual representations, represented by g. These text predictions are then directly fed into the speech generation module. Crucially, the visual representations act as a conditional element for the speech generation process, ensuring that the synthesized speech aligns accurately with the movements in the silent input video.

Table 5.1 This table outlines the architecture of the Conv3D feature extractor, detailing each layer's configuration, including the type, kernel size, stride, number of filters, presence of residual connections, and the input and output sizes.

Layer Type	Kernel Size	Stride	# Filters	Residual	Input Size	Output Size
Conv3D	$5 \times 5 \times 5$	$1 \times 2 \times 2$	64	$\times$	$96 \times 96 \times 3$	$48 \times 48 \times 64$
Conv3D	$1 \times 3 \times 3$	$1 \times 2 \times 2$	128	$\times$	$48 \times 48 \times 64$	$24 \times 24 \times 128$
Conv3D	$1 \times 3 \times 3$	$1 \times 1 \times 1$	128	✓	$24 \times 24 \times 128$	$24 \times 24 \times 128$
Conv3D	$1 \times 3 \times 3$	$1 \times 2 \times 2$	256	$\times$	$24 \times 24 \times 128$	$12 \times 12 \times 256$
Conv3D	$1 \times 3 \times 3$	$1 \times 1 \times 1$	256	✓	$12 \times 12 \times 256$	$12 \times 12 \times 256$
Conv3D	$1 \times 3 \times 3$	$1 \times 1 \times 1$	256	✓	$12 \times 12 \times 256$	$12 \times 12 \times 256$
Conv3D	$1 \times 3 \times 3$	$1 \times 2 \times 2$	512	$\times$	$12 \times 12 \times 256$	$6 \times 6 \times 512$
Conv3D	$1 \times 3 \times 3$	$1 \times 1 \times 1$	512	✓	$6 \times 6 \times 512$	$6 \times 6 \times 512$
Conv3D	$1 \times 3 \times 3$	$1 \times 1 \times 1$	512	✓	$6 \times 6 \times 512$	$6 \times 6 \times 512$

The architecture of the VTP module combines transformer blocks and patch projectors within its architecture. Patch projectors first convert spatial feature maps into patch representations, either through rearrangement of data dimensions or by applying a dense layer. These patches are then sequentially fed into transformer blocks, each tailored to handle specific data resolutions and dimensions.

5.4.2 Visual Text-to-Speech

Once accurate text predictions have been obtained, the next step is to generate the corresponding speech melspectrogram sequence Y, which is in sync with the input video clip I. As explained in Section 5.2.3, out-of-sync speech that does not match the input video would be generated if state-of-the-art TTS models were directly used to synthesize speech from text inputs. A TTS network is thus designed by conditioning the model on the input video features. The visual TTS network majorly comprises five components: (i) Text Encoder, (ii) Visual Encoder, (iii) Visual-Text Attention, (iv) Speaker Embedding, and (v) Spectrogram Decoder. Each of these components is delved into below.

5.4.2.1 Text Encoder

As followed in most of the TTS networks [59], phoneme representations are extracted from text input, which is then given as input to the Transformer encoder layers. The text encoder block is similar to the one used in FastSpeech2 [59], which consists of a positional encoding layer and Feed-Forward Transformer (FFT) layers. The phonemes are transformed to encode the semantic representation and the text embedding vectors, E_{text} of dimension: $N \times d$, are output, where d is the transformer feature dimension.

5.4.2.2 Visual Encoder

The input to the visual encoder is the visual feature sequence g obtained from the lip-to-text network. It is highlighted that these visual features, which capture the lip shape and motion, play a crucial role in generating speech that syncs with the input video. Since these representations were also learned using text supervision, they are likely to reflect accurate content information. This starkly contrasts with previous works where visual representations were directly learned from speech supervision only, which can lead to sub-par visual representations that might contain other unnecessary information, such as the input face identity. The superiority of these representations is one of the critical reasons for the overall network's performance. The extracted $N \times T \times fd$ dimensional representations are given as input to the Transformer encoder layers as shown in Figure 5.2. Similar to the text encoder network, the visual encoder consists of a positional encoding followed by a series of FFT blocks. The learned visual embeddings, E_{vis} of dimension: $N \times T \times d$, are output by the encoder network.

5.4.2.3 Visual-Text Attention

Once the text and the visual embeddings have been obtained, the next and most important step is to find the alignment between these embeddings in time: which phoneme must be uttered when? The generated speech must take the content from the text embeddings, and simultaneously, it should also be temporally aligned (synced) with the video frames. In order to achieve this, a scaled-dot product attention [106] mechanism is employed to learn the correspondence between text and video frames. Specifically, the visual embeddings E_{vis} act as query, and the text embeddings E_{text} act as keys and values.

$$Attention(Q, K, V) = ScaledDot\ ProductAttention(E_{vis}, E_{text}, E_{text}) \in \mathbb{R}^{T \times d} \tag{5.1}$$

Through this attention module, the video-text temporal alignment is learned by the network, which synchronizes the generated speech with the input video frames. Now between the video sequence and melspectrograms, it is known that a natural temporal alignment exists. The length of the melspectrograms is a constant n times the length of the video. Thus, the attention output A is up-sampled n times to directly obtain the melspectrogram duration. This eliminates the need to train a separate duration

predictor as done in the original FastSpeech2 [59]. In other words, the duration of each phoneme in the speech output is already determined by the text-to-video alignment network.

5.4.2.4 Speaker Embedding

Similar to Chapter 4, the voice input of the target speaker is also needed by the model to generate the speech in his/her voice. A random one-second audio segment of the target speaker is considered and the speaker embedding vector V is extracted using the pre-trained identity network[1].

5.4.2.5 Spectrogram Decoder

The speaker embedding vector V is added to the upsampled attention output A to obtain voice-aware content representation. This representation is ingested by the spectrogram decoder, consisting of transformer decoder layers, and the melspectrogram sequence Y' is generated. A L_1 loss is calculated between Y' and Y to train the network. To further improve speech quality, as done in most of the TTS networks, a pre-trained neural vocoder model BigVGAN [103] is adopted to synthesize the speech from the melspectrogram output. It should be noted that this step is only used during inference to obtain high-quality speech outputs.

5.4.3 Datasets and Training Settings

5.4.3.1 Datasets

The model's performance is evaluated on both constrained and unconstrained datasets. The first corpus experimented with is the TCD-TIMIT [61] lip speaker dataset, which comprises lab-recorded videos of 3 speakers. Next, the word-level LRW [30] dataset is considered, consisting of around 150 hours of single-word utterances from hundreds of speakers. More challenging large-scale datasets are then moved on to: LRS2 [28] and LRS3 [29]. The LRS2 data comprises thousands of speakers from BBC programs with a vocabulary of 59000 and around 230 hours of video clips (both "train" and "pre-train" sets together). The LRS3 dataset, on the other hand, is also a large-scale dataset with a total of approximately 430 hours ("train" and "pre-train" sets) of video data with 150000 utterances. It consists of thousands of spoken sentences from TED and TEDx talks in English. The performance of the proposed network is trained and tested using the official splits of LRW, LRS2 and LRS3 datasets, and the train-test split proposed in Lip2Wav for TIMIT dataset [61] is used.

5.4.3.2 Data pre-processing

The video frames are sampled at 25 FPS and the pre-processing procedure of VTP [71] is followed to obtain the face crops. For the speech segments, STFT is computed and then melspectrograms of 80 mel-bands, with a hop length of $10ms$ and a window length of 25ms, sampled at 16kHz, are generated.

[1] `github.com/CorentinJ/Real-Time-Voice-Cloning`

An open-source grapheme-to-phoneme tool is used for text processing to obtain the phoneme inputs for the Visual TTS model.

5.4.3.3 Model configuration and training

The Visual TTS model is comprised of 4 FFT blocks in the text and visual encoders and 6 FFT blocks in the spectrogram decoder network. 512-dimensional embeddings for each frame are obtained as visual embeddings from VTP. In the video-text attention sub-network, the upsample factor, n, is set to 4. A 256 dimensional vector for the speaker embedding is output by the identity network for each speech sample. For the lip-to-text network, the publicly released pre-trained model[2] (trained on LRS2 [28] and LRS3 [29]) is used. The visual TTS model is trained on a single NVIDIA 2080 Ti GPU. The Adam optimizer [130] is used with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-9}$ and the same learning rate schedule as done in [71] is followed. The batch size is set to 16 for all the datasets and the model is trained using L_1 reconstruction loss for approximately $900k$ steps (until convergence). The BigVGAN vocoder [103] is also trained and used during inference to generate the speech from the output melspectrograms.

5.5 Experiments

The quantitative results of the proposed approach and comparisons with existing methods, including those discussed in the previous chapters, are presented. As automatic speech metrics are imperfect, MOS scores using human evaluation are also shown. Finally, a real-world application of lip-to-speech is demonstrated for the first time by voicing the silent lip movements of an ALS patient.

5.5.1 Quantitative Evaluations

5.5.1.1 Metrics

The quality of the generated speech is measured using the standard speech metrics: Perceptual Evaluation of Speech Quality (PESQ), Short-Time Objective Intelligibility measure (STOI), and its extended version (ESTOI). PESQ measures the clarity and overall perceptual quality of speech, while STOI and ESTOI measure the intelligibility of speech. Further, as discussed previously, it is crucial to generate speech that is in sync with the input video. The lip-sync metrics, Lip-Sync Error - Confidence (LSE-C) and Lip-Sync Error - Distance (LSE-D) [15], are used to evaluate whether the output speech matches the input lip movements. The public implementations of all the above metrics are used for a fair comparison. For more information regarding the metrics, Section 3.4.4 can be referred to.

5.5.1.2 Speech Synthesis in Constrained Settings

5.5.1.2.1 Comparisons To evaluate lip-to-speech methods on the constrained single-speaker TCD-TIMIT dataset, four existing approaches are compared: (i) GAN-based [47], (ii) the proposed Lip2Wav

[2] `https://github.com/prajwalkr/vtp`

presented in Chapter 3, (iii) the proposed VAE-GAN in Chapter 4, and (iv) VCA-GAN [56]. The same settings as Lip2Wav are adopted and the scores and VCA-GAN [56] are reported. Finally, a parallel state-of-the-art Lip-to-Speech system called LipVoicer [119] is also compared on LRS3 [29] dataset. The Wav2Lip [27] repository is used to compute the LSE-C and LSE-D scores for each method. The metrics for a particular model that were not mentioned in the original papers or for which no publicly available pre-trained checkpoint exists have been excluded.

5.5.1.2.2 Results Table 5.2 contains the results on the TCD-TIMIT dataset. It is observed that the proposed approach achieves comparable results to previous methods in constrained settings with minimal data of only 3 speakers. However, the significant benefits of the proposed approach can be seen in unconstrained settings, which are described below.

Table 5.2 The state-of-the-art methods are compared on several standard multi-speaker benchmarks using standard metrics. The generated outputs from the model are found to be the most natural (PESQ), the most accurate (STOI, ESTOI), and in perfect sync with the video input (LSE-C, LSE-D) in the in-the-wild videos of LRW [30], LRS2 [28], and LRS3 [29].

Dataset	Method	PESQ↑	STOI↑	ESTOI↑	LSE-C↑	LSE-D↓
TCD-TIMIT [61]	GAN-based [47]	1.22	0.51	0.32	-	-
	Lip2Wav	1.35	0.56	0.36	6.610	7.815
	VCA-GAN [56]	**1.43**	0.58	0.40	-	-
	VAE-GAN	1.35	0.55	0.35	-	-
	Ours	1.34	**0.61**	**0.42**	**6.623**	**6.901**
LRW [30]	GAN-based [47]	0.72	0.10	0.02	1.983	9.426
	Lip2Wav	1.19	0.54	0.34	2.526	8.286
	VAE-GAN	0.78	0.15	0.03	2.538	8.173
	VCA-GAN [56]	1.33	0.56	0.36	-	-
	SVTS [57]	1.49	0.64	0.48	-	-
	Multi-task L2S [118]	1.56	0.64	0.47	4.876	8.102
	Lip-to-Text + TTS baseline	0.69	0.10	0.01	1.993	12.872
	Ours	**1.61**	**0.71**	**0.56**	**6.812**	**6.974**
LRS2 [28]	Lip2Wav	0.58	0.28	0.11	1.874	11.48
	VAE-GAN [47]	0.60	0.34	0.17	2.507	8.155
	VCA-GAN [56]	1.24	0.40	0.13	4.016	7.914
	SVTS [57]	1.34	0.49	0.29	-	-
	Multi-task L2S [118]	1.36	0.52	0.34	4.001	8.192
	Lip-to-Text + TTS baseline	0.53	0.19	0.02	2.013	15.891
	Ours	**1.47**	**0.65**	**0.47**	**8.083**	**6.586**
LRS3 [29]	VAE-GAN [47]	0.51	0.30	0.15	2.063	8.256
	VCA-GAN [56]	1.23	0.47	0.20	3.905	8.392
	SVTS [57]	1.25	0.50	0.27	-	-
	Multi-task L2S [118]	1.31	0.48	0.26	3.876	8.677
	LipVoicer [119]	1.08	0.36	0.19	6.239	8.266
	Lip-to-Text + TTS baseline	0.42	0.16	0.01	1.771	17.882
	Ours	**1.39**	**0.58**	**0.37**	**7.886**	**6.850**

5.5.1.3 Speech Synthesis in Unconstrained Settings

5.5.1.3.1 Comparisons In order to assess the performance of lip-to-speech methods in unconstrained scenarios, three datasets are employed: word-level LRW [30], sentence-level LRS2 [28], and LRS3 [29]. While the GAN-based [47] and Lip2Wav models have been re-trained by the authors of VAE-GAN in a multi-speaker context, the scores from their original study are presented for comparison. For VCA-

GAN [56], SVTS [57], and Multi-task Lip-to-Speech synthesis [118], the speech metric (PESQ, STOI and ESTOI) scores are adopted from [118]. Further, publicly accessible pre-trained checkpoints for VCA-GAN[3] and Multitask-L2S[4] are used to generate speech on LRS2 and LRS3 test sets for the former, and LRS2, LRS3, and LRW test sets for the latter. These generations are utilized to compute the LSE metrics for both techniques, wherever applicable. Lastly, results for a baseline approach that leverages lip-to-text conversion followed by multi-speaker TTS without a visual stream in the TTS model are included. The evaluation does not include all metrics that were not originally reported in the papers or for which no pre-trained model is publicly available.

5.5.1.3.2 Results The results on the challenging LRW, LRS2, and LRS3 datasets are presented in Table 5.2. The proposed model consistently outperforms the existing methods by a significant margin on all these datasets. Since the GAN-based [47] model was proposed to work for constrained laboratory recorded datasets, it can be observed that extending this model in unconstrained settings does not yield satisfactory results. Lip2Wav performs decently on the word-level LRW dataset; however, it fails to learn the audio-visual alignment on the LRS2 dataset, thus leading to very poor performance. This model is discarded for further comparison on the LRS3 dataset. VAE-GAN, VCA-GAN [56], SVTS [57] and Multitask-L2S [118] generate speech that is in-sync with the input video; however, they fail to synthesize accurate content. The quality of the generated speech is often non-intelligible and leads to lower scores in speech quality metrics. The Lip-to-Text + TTS baseline model is on the opposite spectrum, where the content is generated well by the model but lip-sync is failed to be captured, mainly because the speed, prosody, and accents of speakers cannot be inferred by the model just from the text input. The model proposed in this chapter, on the other hand, is capable of generating both the actual spoken content as well as maintaining precise lip synchronization. As can be seen from the table, all the speech quality metrics are outperformed by the proposed method, indicating the robustness and superiority of the proposed approach. In Figure 5.3, how the model temporally aligns video and text sequences in the process of generating speech is depicted. The reader is encouraged to view the provided demo video comprising multiple qualitative samples and comparisons.

5.5.1.3.3 Qualitative comparisons Qualitative results from the model proposed in this chapter are presented in the form of a linked video in Figure 5.4, providing a dynamic and illustrative demonstration of the research findings.

5.5.2 Human Evaluations

To evaluate the applicability of the method in real-world scenarios, subjective human evaluations were conducted. 25 volunteers were asked to assess the quality of speech generations. The participant group had an almost equal male-female ratio, spanning an age group of $20-45$ years. 10 long sentences

[3] https://github.com/ms-dot-k/Visual-Context-Attentional-GAN
[4] https://github.com/ms-dot-k/Lip-to-Speech-Synthesis-in-the-Wild

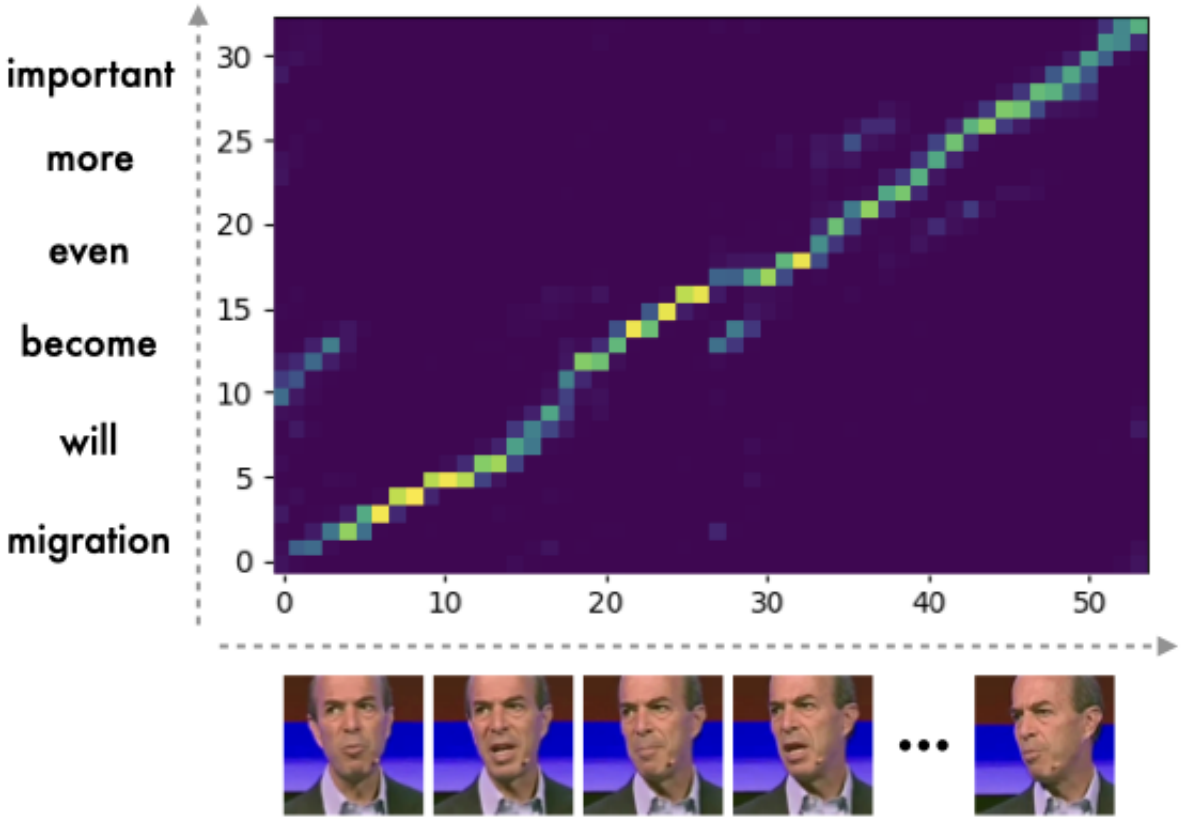

Figure 5.3 The video-text alignment from the scaled dot product attention step of the model is visualized. It is observed that the model learns a strong monotonic near-diagonal attention, as expected.

(10 seconds or longer) were randomly selected from the test set of LRS3 [29] and the results from different methods were presented to the participants. They were asked to rate the samples on a scale of $1-5$ based on the following criteria: (A) Intelligibility (is the speech meaningful?), (B) Content clarity (are the words clear?), (C) Sync Accuracy, and (D) Overall perceptual quality of the talking head video $+$ audio. The mean opinion scores are reported in Table 5.3. In line with the quantitative evaluations, the method was highly rated over other approaches in all the criteria listed above. As expected, the speech intelligibility was rated slightly higher for the Lip-to-Text $+$ TTS baseline. However, the overall perceptual quality for this baseline sharply fell due to the lack of sync between the spoken content and the lip movements. Overall, Table 5.3 clearly signifies that the network is able to generate speech with more clarity, which sounds more natural and is of considerably higher quality.

5.6 Applications in Assistive Technology

Lip-to-Speech synthesis has a host of applications in an increasingly digital world. Simple applications such as performing video calls in quiet environments, filling in audio interruptions due to technical

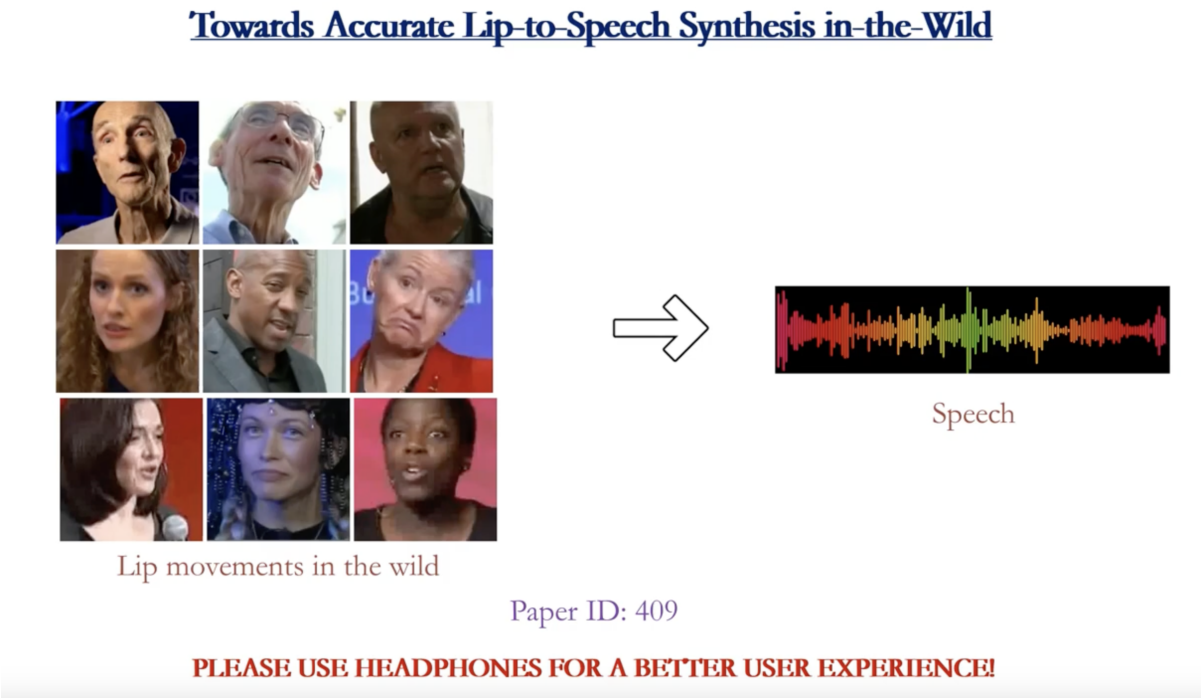

Figure 5.4 A video is presented in this link: `https://youtu.be/6WNSazF9vyQ`. This video contains qualitative results and comparisons from the model proposed in this chapter. This image is presented as the thumbnail for this video.

Table 5.3 (A) Intelligibility, (B) Content clarity, (C) Sync Accuracy, (D) Overall perceptual quality. The model produces natural and realistic speech outputs that is largely preferred by the users in comparison to other approaches.

Method	(A)	(B)	(C)	(D)
GAN-based [47]	2.05	1.87	1.99	2.12
Lip2Wav	1.01	1.03	1.34	1.01
VAE-GAN	1.07	1.33	2.18	2.57
VCA-GAN [56]	2.18	1.88	2.97	2.54
Multi-task L2S [118]	2.19	1.85	3.01	2.64
Lip-to-Text + TTS baseline	**3.61**	2.87	1.01	2.96
Ours	3.49	**3.52**	**3.82**	**3.31**

issues, and eliminating unwanted background chatter can be made possible with accurate lip-to-speech. It is believed that the most significant application of lip-to-speech can be found in assistive technologies. The current assistive systems used to improve the communication ability of people suffering from various disorders affecting their speech can be revolutionized. Words can be mouthed by patients suffering from vocal cord disabilities to communicate naturally with the world around them. The synthesized speech can be personalized and also be in sync with the speaker's lip movements.

5.6.1 Generating Speech for a Patient suffering from ALS

A recent work [159] proposed a lip reading technique for a patient suffering from ALS. The patient has feeble vocal cord movements but can mouth words silently. Limited amounts of data were collected from the patient by the authors of the paper, and a model was trained to recognize words and sentences from his lip movements. As a result of significant improvement in this task, the benefit of using lip-to-speech as a future assistive technology can now be demonstrated. Through the help of the authors of [159], the lip-to-speech system was evaluated on the patient's data. Please note that the data was anonymized and used only for research purposes. It was found that the lip-to-text module generates fairly accurate text (WER of $\approx 37\%$), and the visual TTS model generates clear speech in sync with the patient's lip movements. This is the first demonstration of automatic lip-to-speech synthesis for an unseen speaker in an entirely out-of-domain real-world application. Furthermore, the model was also tested on other deaf speakers studied in [159] and accurate performance was observed.

5.6.2 Ethical Considerations

It is acknowledged that the work has the potential to generate synthetic speech for videos, given that only a $1-$second voice sample from any target speaker is required. However, since context is provided by the video and the output speech is constrained, it is likely that the generated speech will closely follow the original content. The importance of ethical considerations regarding the use of such models is recognized, and it is ensured that the models will only be shared with users who consent to limit their usage to research-oriented and ethically valid tasks.

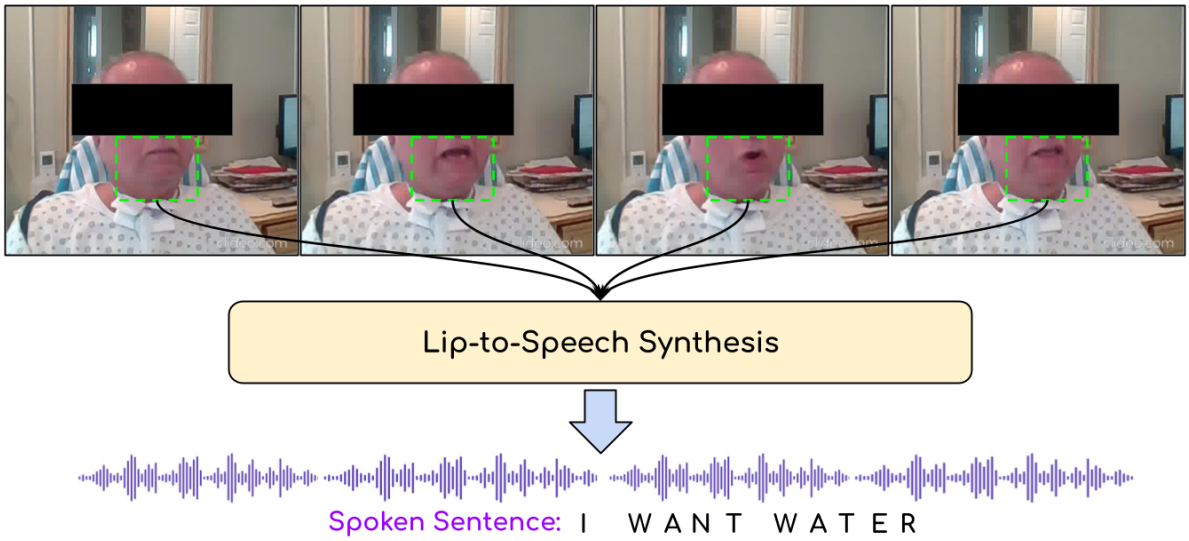

Figure 5.5 The model is demonstrated on an ALS patient who cannot voice words but can mouth them. The speech corresponding to the silent lip movements can be generated. Lip-to-Speech can thus be a cheap and non-invasive method to assist someone who has lost their voice.

5.7 Ablation Studies

In this section, several ablation studies are performed to understand the effect of different components of the proposed model. All the ablation experiments are conducted on the LRS2 [28] test set until and unless stated otherwise.

5.7.1 Effect of different pre-trained lip-to-text models

Additional experiments were conducted using other pre-trained lip-to-text models, specifically DeepLR [86] and AV-HuBERT [88]. The former had a WER of 51.3 on the LRS2 test set, while the latter had a WER of 46.6. The input text to the pre-trained Visual TTS module was taken from different Lip-to-Text models. Additionally, the ground-truth text from the LRS2 test set was directly provided. As shown in Table 5.4, speech was recovered that was somewhat accurate, despite the presence of noisy text transcripts from both models. This can be attributed to two factors: (i) the correction of errors made by the lip-to-text network by the pipeline to some extent (demonstrated in the demo video); and (ii) the reduced difference in scores between homonyms such as ”ship and sheep” or ”berth and birth” in the audio domain.

5.7.2 Effect of different visual representations

The proposed Visual Text-to-Speech module was trained with RGB face crops instead of the VTP embeddings to generate speech conditioned on text and lip movements. Based on observations from Ta-

Table 5.4 A comparison of using generated text from different lip-to-text networks in the pipeline is presented. The WER of the lip reading model (L2T-WER) on the LRS2 test set is also reported as a reference.

Method	L2T-WER	PESQ↑	STOI↑	ESTOI↑	LSE-C↑	LSE-D↓
Deep Lip Reading [87]	51.3	1.17	0.40	0.22	7.847	6.904
AV-HuBERT [88]	46.1	1.27	0.53	0.40	7.960	7.003
VTP (Ours)	22.6	1.47	0.65	0.47	8.083	6.586
GT text	-	1.51	0.69	0.50	8.781	6.106

ble 5.5, VTP embeddings are found to be the most suitable for this task because they excel in localizing and representing the shape of the speaker's lips.

Table 5.5 The effect of using different visual representations for training the Visual TTS module is presented in this table.

Method	PESQ↑	STOI↑	ESTOI↑	LSE-C↑	LSE-D↓
Face crops	1.17	0.40	0.22	7.847	6.904
VTP (Ours)	1.47	0.65	0.47	8.083	6.586

5.7.3 Effect of using the speaker-embedding

The generation of prosody, pitch, and other speech nuances is a complex challenge when relying solely on facial information. A speaker identity network that is *NOT* speaker-specific is used in the proposed approach. With just a 1-second speech sample of the target speaker, the network extracts the speaker embeddings, capturing the desired voice. As a result, speech in the voice and style of any in-the-wild speaker can be generated by the proposed model. An ablation study was performed without using the identity network, and the results are reported in Table 5.6. The results show that without the identity network, an average (robotic-kind of) voice, lacking style and voice quality, is generated by the model.

Table 5.6 Identity network ablation on LRS2 test set.

Method	PESQ↑	STOI↑	ESTOI↑	LSE-C↑	LSE-D↓
W/o speaker embedding	1.06	0.43	0.30	5.109	11.204
With speaker embedding	**1.47**	**0.65**	**0.47**	**8.083**	**6.586**

5.7.4 Using different vocoders for generating speech

Different vocoders were trained to improve the quality of the final generated speech from the mel-spectrograms. BigVGAN [103], HiFiGAN [102], MelGAN [101], and the non-learnable Griffin-Lim [107] were used to convert melspectrograms into raw waveforms, and the performance is reported in Table 5.7. BigVGAN was selected as the vocoder of choice for all the datasets based on its slightly superior performance in this experiment.

Table 5.7 Using a vocoder network during inference to generate speech produces better quality outputs.

Speech generation	PESQ↑	STOI↑	ESTOI↑	LSE-C↑	LSE-D↓
MelGAN [101]	1.27	0.61	0.43	8.072	6.586
Griffin-Lim [107]	1.26	0.57	0.39	8.079	6.581
HiFiGAN [102]	1.28	0.58	0.46	8.082	6.602
BigVGAN [103]	1.47	0.65	0.47	8.083	6.586

5.7.5 Word-error-rate comparison

In Table 5.8, the Word Error Rate (WER) of the state-of-the-art lip-to-text system, Auto AVSR [89] and the generated speech from the proposed model, transcribed using WhisperX [160] is compared. The same comparison is also conducted on the generations of LipVoicer [119]. The LRS3 [29] test set is used for this comparison. The text input to the proposed lip-to-speech system is also generated from Auto AVSR for this experiment for a fair comparison.

Table 5.8 Comparison of WER (After ASR) for the proposed lip-to-text + Visual TTS model and LipVoicer models using the LRS3 test set.

Model	WER ↓
Ours	0.26
LipVoicer [119]	0.36
Auto AVSR [89]	0.19

The table shows that the proposed model outperforms LipVoicer in terms of WER after passing through WhisperX, with a lower WER of 0.26 compared to 0.36. However, Auto AVSR achieves the best performance with a WER of 0.19. This suggests that while the proposed model is effective and produces describable speech, there is still room for improvement to reach the performance level of Auto AVSR.

5.7.5.1 Comparison of phoneme-error-rates

Word error rates (WER) are not always the most suitable metric for lip reading, as they can be overly sensitive to small differences. For instance, words like "work" and "want" will be classified completely differently, even if only a single phoneme between them differs. Therefore, phoneme error rates (PER) are calculated instead of WER. By transcribing the sentences into a sequence of phonemes and then applying the standard WER calculation method to these phoneme sequences, a more nuanced and accurate evaluation of the lip-to-speech model's performance is achieved.

It is observed that the proposed model performs closer to Auto AVSR than the WER suggests. This indicates that many phonemes are predicted correctly, where only a part of the word is wrong. At a phoneme level, the proposed model does much better than the competitive LipVoicer and also Auto AVSR. While Auto AVSR's error rate goes from 0.19 to 0.16, the proposed model's error rate goes from 0.26 to 0.19. This clearly shows that the proposed model predicts many phonemes correctly, even if

Table 5.9 Comparison of PER for the proposed lip-to-text + Visual TTS model and LipVoicer models using the LRS3 test set.

Model	**PER** ↓
Ours	0.19
LipVoicer [119]	0.29
Auto AVSR [89]	0.16

the entire word might be wrong. Please note that the ASR system itself may inject some noise into this analysis.

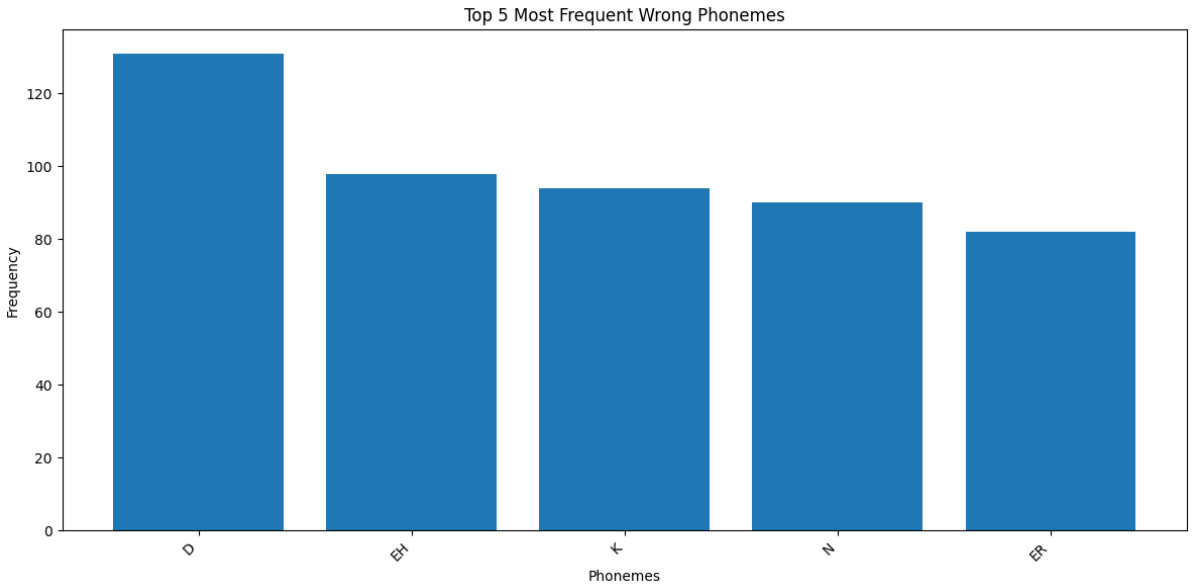

Figure 5.6 The top-5 wrongly predicted phonemes by the proposed lip-to-speech network are plotted in this Figure.

To further analyze this, the phonemes that are most frequently missed were examined. The frequencies of the top 5 phonemes that are incorrectly predicted are plotted in Figure 5.6. It was observed that the phonemes that are mostly wrong actually have little to no lip movements.

The phonemes D, EH, K, N, and ER primarily involve articulatory movements that do not significantly engage the lips. For the phoneme /d/, which is a voiced alveolar stop, the tongue tip touches the alveolar ridge behind the upper front teeth, with no primary involvement of the lips. The /e/ phoneme, a mid-front unrounded vowel, involves positioning the tongue halfway between high and low in the mouth and towards the front, with the lips slightly open but not significantly involved in shaping the sound. The /k/ sound, a voiceless velar stop, is produced by the back of the tongue contacting the soft palate (velum), with no lip involvement. Similarly, the /n/ phoneme, a voiced alveolar nasal, is produced with the tongue tip touching the alveolar ridge and the velum lowered, again without lip involvement. Finally, the /er/ phoneme, a mid-central r-colored vowel, involves a mid-central tongue position with

a slight curl back of the tongue tip and only a slight rounding of the lips. Given that the lips are not the primary articulators for these sounds, their production relies more on the tongue's interaction with various parts of the mouth. This lack of distinct lip movement likely contributes to the challenges faced by lip-reading systems in accurately interpreting these phonemes. A very similar outcome is also noted for the lip-to-text Auto AVSR network.

Figure 5.7 The wrongly predicted phonemes are produced without moving the lips, and thus, both the lip-to-text and lip-to-speech models struggle to predict these successfully.

5.7.6 Comparison based on gender

To check for gender bias, a test was conducted on the LRS3 test set using the DeepFace repository [161]. In this test, 10 frames from each video were passed through the gender detection algorithm. A voting mechanism was employed, where if more than 5 frames were classified as a particular gender, the video was predicted to be of that gender. The results of this test are shown in Table 5.10.

Table 5.10 Gender bias test results on the LRS3 test set

Gender	PESQ↑	STOI↑	ESTOI↑	LSE-C↑	LSE-D↓	WER ↓
Male	1.42	0.60	0.36	7.88	6.82	0.23
Female	1.36	0.56	0.38	7.89	6.88	0.28

It is observed that the outputs of the proposed model are not affected by gender.

5.7.7 Comparison based on age

To check for age bias, a test was conducted on the LRS3 test set using the DeepFace repository [161]. In this test, 20 frames from each video were passed through the age detection algorithm. A voting mechanism was employed, where the age range with the most votes determined the predicted age range of the video. If no age range was a clear winner, the video was skipped (250 videos were skipped). The age ranges used were: $0-18$, $19-30$, $31-45$, $45-60$, and $60+$. The results of this test are shown in Table 5.11.

The above table shows that the model remains consistent in terms of quality for all age groups with no major variations.

Table 5.11 Age bias test results on the LRS3 test set

Age	PESQ↑	STOI↑	ESTOI↑	WER ↓
0-18	1.17	0.48	0.32	0.19
19-30	1.19	0.47	0.30	0.27
31-45	1.20	0.49	0.32	0.26
46-60	1.17	0.46	0.30	0.25
60+	1.18	0.48	0.32	0.27

5.7.8 Comparison based on race

To check for racial bias, a test was conducted on the LRS3 test set using the DeepFace repository [161]. In this test, 20 frames from each video were passed through the race detection algorithm. A voting mechanism was employed, where the race with the most votes determined the predicted race of the video. If no race was a clear winner, the video was skipped. Approximately 300 videos were removed where the classifier was not confident. The racial categories used were: Asian, Black, Latino/Hispanic, Middle Eastern, and White. The results of this test are shown in Table 5.12.

Table 5.12 Race bias test results on the LRS3 test set.

Race	PESQ ↑	STOI ↑	ESTOI ↑	WER ↓
Asian	1.19	0.47	0.30	0.22
Black	1.20	0.50	0.32	0.26
Indian	1.18	0.48	0.30	0.53
Latino/Hispanic	1.50	0.54	0.35	0.19
Middle Eastern	1.19	0.48	0.31	0.35
White	1.18	0.47	0.31	0.26

From the results, it is observed that while the speech quality metrics (PESQ, STOI, and ESTOI) remain relatively consistent across different racial categories, the Word Error Rates (WER) for Indian and Middle Eastern categories are notably higher. Specifically, the WER for Indian speakers is 0.53, and for Middle Eastern speakers, it is 0.35, compared to lower WERs for other racial categories. This discrepancy indicates that while the model maintains a consistent speech quality, it struggles with accurate word prediction for certain racial groups, highlighting an area for further improvement in the model's robustness and fairness.

5.7.9 Comparison for Different Emotions

To evaluate the performance of the proposed model across different emotions, a speaker from the MEAD dataset [62] was used as shown in Figure 5.8. The dataset provides labeled videos with various emotional expressions. For this comparison, the frontal video of a single identity speaking English was selected. The emotions were chosen at their maximum intensity, and the categories include neutral, angry, disgust, contempt, fear, happy, sad, and surprise. The results of this evaluation are presented

in Table 5.13, showing how the model performs in terms of PESQ, STOI, ESTOI, and WER for each emotion.

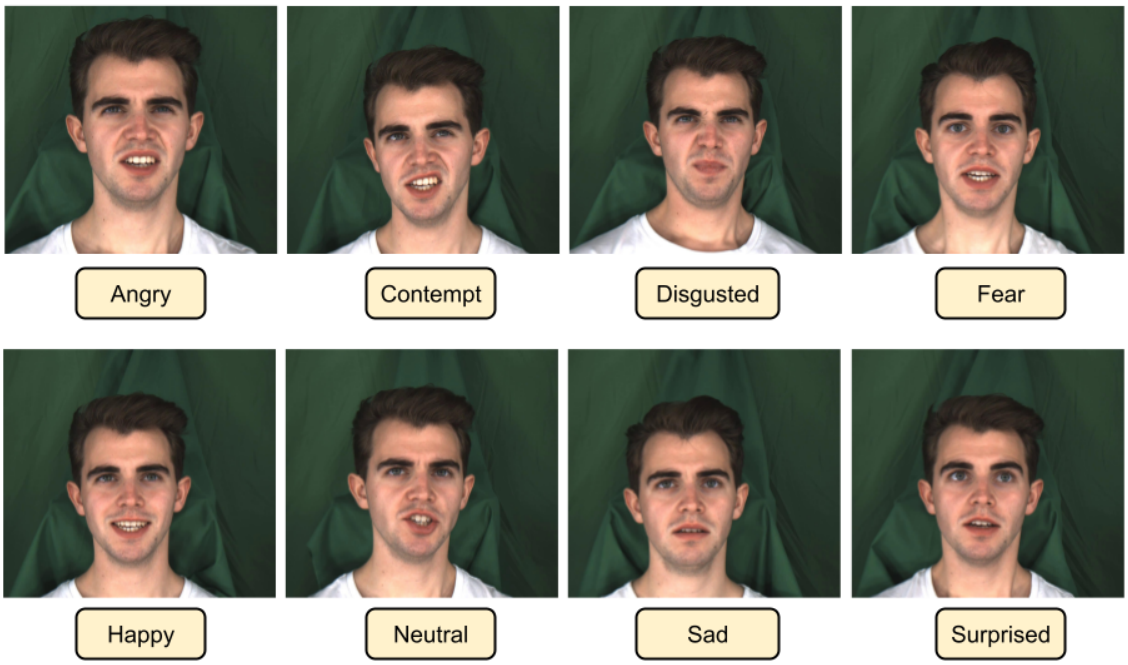

Figure 5.8 Multiple emotions are selected from the MEAD dataset for a particular speaker to evaluate lip-to-speech synthesis using the proposed model.

Table 5.13 Emotion comparison results using the MEAD dataset

Emotion	**PESQ** ↑	**STOI** ↑	**ESTOI** ↑	**WER** ↓
Angry	1.08	0.12	0.07	0.38
Disgust	1.13	0.15	0.09	0.34
Contempt	1.10	0.13	0.08	0.36
Neutral	1.15	0.26	0.19	0.27
Fear	1.09	0.14	0.09	0.35
Happy	1.14	0.16	0.10	0.32
Sad	1.12	0.15	0.09	0.30
Surprise	1.11	0.14	0.09	0.33

Among the various emotions evaluated using the MEAD dataset, the sad and neutral emotion exhibits the best performance. This is primarily because the sad emotion in the MEAD dataset involves less exaggerated lip movements, making it easier for the model to interpret and generate accurate speech. In contrast, other emotions such as angry and surprise have more exaggerated lip movements, which complicates the lip-reading process. The angry emotion, in particular, introduces a lot of teeth visibility, which further challenges the model. Similarly, the surprise emotion, with its pronounced lip movements, also proves difficult for accurate interpretation.

5.7.10 Comparison for Different Head Poses

To assess the impact of different head poses on the performance of the proposed model, the same speaker of the MEAD dataset [62] was used again. The different head poses or camera positions are shown Figure 5.9. For this comparison, neutral videos of the same identity were selected, covering a range of head poses: left-60, left-30, down-30, front, up-30, right-30, and right-60. The results of this evaluation are presented in Table 5.14, illustrating the performance of the model in terms of PESQ, STOI, ESTOI, and WER for each head pose.

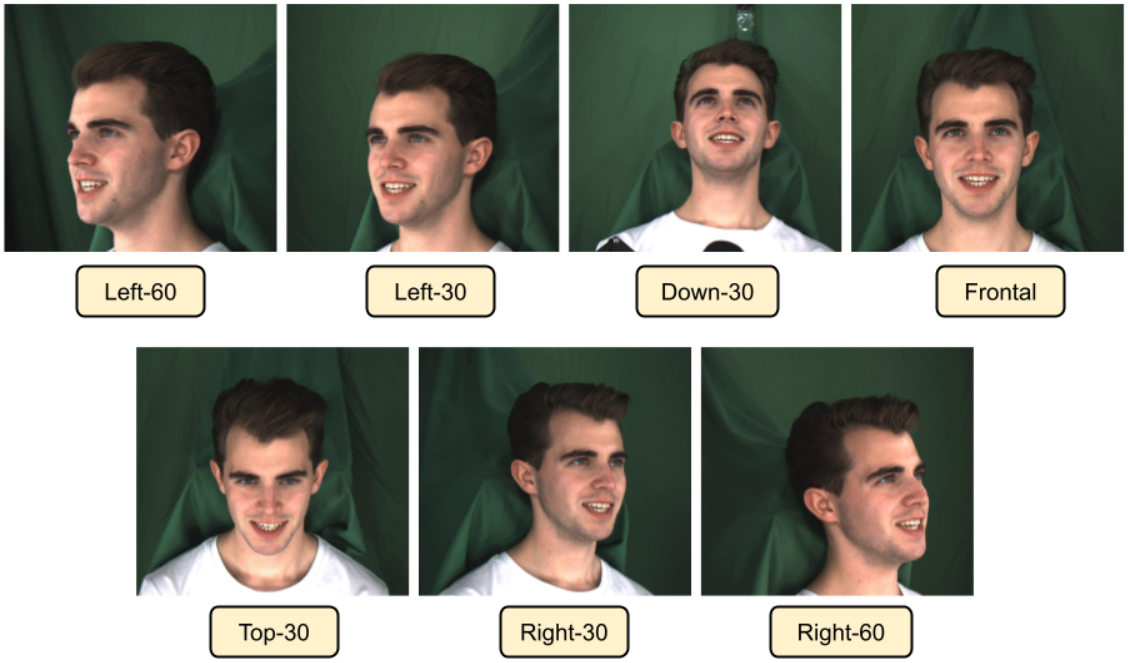

Figure 5.9 Multiple camera positions are selected from the MEAD dataset for a particular speaker to evaluate lip-to-speech synthesis using the proposed model. The camera positions mimic different head poses in real world situations.

Table 5.14 Head pose comparison results using the MEAD dataset

Head Pose	PESQ ↑	STOI ↑	ESTOI ↑	WER ↓
Left-60	1.09	0.13	0.08	0.36
Left-30	1.12	0.15	0.09	0.34
Down-30	1.11	0.14	0.09	0.33
Front	1.18	0.17	0.11	0.28
Up-30	1.14	0.16	0.10	0.30
Right-30	1.13	0.15	0.09	0.31
Right-60	1.10	0.14	0.09	0.35

The evaluation of different head poses using the MEAD dataset indicates that the frontal head pose performs the best across all metrics. This can be attributed to the fact that a frontal view provides the

clearest and most direct visual information of lip movements, which is crucial for accurate lip-to-speech synthesis. As the head moves away from the frontal position, the visibility of lip movements decreases, resulting in a decline in model performance.

5.7.11 Limitations

In this work, the problem of lip-to-speech networks not being able to learn a language model directly from speech supervision is addressed by using a pre-trained lip-to-text network. While the proposed model does not require ground-truth text annotations, the lip-to-text model upon which it is built has been trained with text supervision. However, recent efforts in self-supervised pre-training have led to a sharp decrease in the number of text annotations required for training accurate lip-to-text models [88], making it easier to extend such models to lip-to-speech using the proposed approach. Currently, the model has only been tested in English, and validation in other languages remains to be done.

5.8 Summary

In conclusion of this chapter, the proposed research presents an innovative approach to unconstrained multi-speaker lip-to-speech synthesis that outperforms previous methods by incorporating language and visual information from a highly accurate lip-to-text model. Significant improvements in lip-to-speech synthesis are demonstrated, generating high-quality outputs that seamlessly synchronize with silent lip video. This study can potentially open up exciting avenues for future research. The success of this approach in assistive technology is particularly encouraging, where it has been shown that the method can generate accurate speech from silent lip movements of individuals with speech impairments. Overall, optimism exists about the possibilities of this approach to improve communication and enhance the quality of life for people with speech impairments, and it is anticipated that this work will drive further progress in this field.

Chapter 6

Adding Degraded Speech to Lip-to-Speech Synthesis setup

In this thesis, the primary focus has been on lip-to-speech synthesis, which involves generating speech solely from lip movements. However, a new avenue of research is explored in this chapter by considering a slightly modified problem: the integration of noisy speech along with lip movements. It is hypothesized that even with highly degraded audio input, the addition of visual information could significantly improve speech reconstruction. This approach leads to the exploration of Audio-Visual Speech Enhancement (AVSE), a field where visual and auditory information processing intersect. Traditionally, speech enhancement techniques have been predominantly audio-centric, often struggling in environments with compromised audio signals. The AVSE problem, as addressed in this chapter, is conceptualized as an extension of lip-to-speech synthesis, incorporating an additional layer of noisy or low-quality audio inputs.

Two distinct yet interconnected problems within AVSE are explored:

1. Audio-visual speech super-resolution: This involves enhancing the resolution and clarity of speech signals. The task is particularly challenging when audio data is not only of low quality but also sparse in detail. Visual cues from lip movements are integrated to reconstruct and augment the audio data, aiming to achieve higher-resolution speech output that surpasses the limitations of traditional audio-only enhancement methods.

2. Audio-visual speech denoising: This focuses on removing noise from speech signals in environments where the audio is heavily contaminated with background noise or interference. Visual information from lip movements is crucial in distinguishing speech from noise, allowing for more effective and precise noise reduction.

In summary, this chapter explores how integrating lip movement data with traditional speech processing techniques can significantly improve speech signal quality and intelligibility, even in scenarios where conventional audio-based approaches fall short.

6.1 Audio-Visual Speech Super-resolution

Speech super-resolution is recognized as an important task in various scenarios, including the enhancement of historical recordings, improvement of telephonic communications, and compression of

speech data for bandwidth reduction. Here, "low-resolution" speech refers to speech sampled at a lower rate (e.g., 16 kHz or 24 kHz), while "high-resolution" speech denotes a higher sampling rate (e.g., 48 kHz) that preserves more high-frequency detail; this terminology is borrowed directly from image super-resolution, where one aims to reconstruct a higher-resolution signal from a lower-resolution input. Traditionally, audio-only methods have been limited to scale factors of $2\times$ and, at most, $4\times$. Building upon the success of lip-to-speech synthesis, it was hypothesized that incorporating visual information from lip movements could enable super-resolution at significantly higher scale factors. This approach can be conceptualized as lip-to-speech synthesis with the addition of low-sampling-rate audio input.

To test this hypothesis, an innovative audio-visual network was developed. This network aims to super-resolve speech from extremely low sampling rates (as low as 1 kHz) to standard audio quality (16 kHz), achieving scale factors of $8\times$ and $16\times$. These scale factors represent a significant advancement over previous audio-only methods. The proposed approach not only outperforms state-of-the-art audio-only techniques in speech quality and intelligibility metrics but also demonstrates applicability in real-world, unconstrained settings. Additionally, a "pseudo-visual model" was developed to synthesize lip movements from low-resolution speech inputs, extending the method's applicability to scenarios where visual data is unavailable.

6.1.1 Background on audio-only speech super-resolution

The speech community has studied the problem of upsampling the frequency of speech signals for a long time. This problem was popularly known as "bandwidth extension" in the pre-deep learning era. Initially, classical signal processing approaches were utilized [162, 163] to solve the task of bandwidth extension. This was followed by methods relying on Gaussian mixture models to predict high-frequency speech based solely on low-frequency input [164]. With the advent of deep learning, renewed interest in this problem was sparked under a new alias, "audio super-resolution". Inspired by the success of image super-resolution techniques using deep learning, a simple neural network was first used by [165] to learn mappings between high-resolution and low-resolution audio signals. This was further improved by various techniques such as residual-based bottleneck network [166], "Temporal-Film (TFiLM)" [167], and the diffusion probabilistic model "NU-Wav" [168], which significantly enhanced SR performance. While current state-of-the-art methods work directly on low-resolution speech, they are proposed for $2\times$ and $4\times$ SR, in stark contrast to the $16\times$ SR attempted in this work. To improve these networks' robustness and real-world applicability, the use of additional assistance in terms of visual modality, particularly lip movements, is considered. Super-resolving low-resolution speech is explored by using lip movements as additional cues. The assistance from the visual stream allows for handling large-scale factors like $16\times$ compared to $4\times$ as done in previous works. Furthermore, to enable the model to be applied in practical situations, the pseudo-visual approach proposed in [169] is extended for speech SR. Thus, along with the audio-visual approach, an audio-only system that incorporates the advantages of the visual stream without requiring a real visual stream is also developed.

6.1.2 Formally formulating the problem

Our goal is to take a sequence of lip movements $I = (I_1, I_2, \ldots, I_N)$ and a corresponding low-resolution speech segment $S^{lr} = (S_1^{lr}, S_2^{lr}, \ldots, S_{T'}^{lr})$ to generate a high-resolution speech segment $S = (S_1, S_2, \ldots, S_{T'})$. The generated speech S should correspond to the lip movements I and be an enhanced version of S^{lr}. It should be noted that in this work, unlike previous chapters where mel-spectrograms were used, linear spectrograms are utilized to represent speech. The high-resolution linear spectrogram is denoted as $Y = (Y_1, Y_2, \ldots, Y_T)$. The linear spectrogram from the low-resolution speech is used as another input to the network other than the lip movements. It is denoted by $Y^{lr} = (Y_1^{lr}, Y_2^{lr}, ..., Y_T^{lr})$.

6.1.3 The Architecture

When the speech resolution is low, the loss of information is so paramount that the semantic details of speech are almost completely lost. In such cases, it is shown that the visual stream can aid in recovering the content, thereby improving the quality and coherence of super-resolved speech. The proposed audio-visual model comprises three modules: (i) Speech Encoder, (ii) Visual Encoder, and (iii) Speech Decoder. Each of these modules is elaborated upon below.

6.1.3.1 Speech Encoder

In the proposed approach, a 1-second segment of low-resolution (LR) speech, denoted as S_{lr}, is initially considered. Linear interpolation is applied to this segment to upscale it to the desired target resolution, resulting in an upsampled version, S. This upsampling step is crucial for maintaining a consistent architecture regardless of the input resolution. The raw waveforms are transformed into a more analyzable format using the Short-Time Fourier Transform (STFT). For the STFT computation, a window length of 25ms and a hop length of 10ms are employed, with the sampling rate set at 16kHz. A complex STFT with dimensions $(T, 257)$ is produced as the outcome of this process. This complex STFT is then decomposed into its magnitude and phase components, each of which is normalized within the $[0, 1]$ range to ensure data representation consistency. The magnitude and phase components are subsequently concatenated along the frequency axis, resulting in a representation with dimensions $(T, 514)$, which serves as the input for the speech encoder. The speech encoder, comprising a series of residual 1D convolution layers, processes these time-frequency representations, ultimately generating speech embeddings with dimensions $(T, 600)$.

In the methodology, the raw speech is converted into linear spectrograms using STFT. A representation with $T \times 514$ dimensions is obtained by concatenating the magnitude and phase components of the STFT. This representation is then fed into the speech encoder, which is processed using a stack of 1D convolution layers, as described in Table 6.1, to produce speech embeddings with a dimensionality of

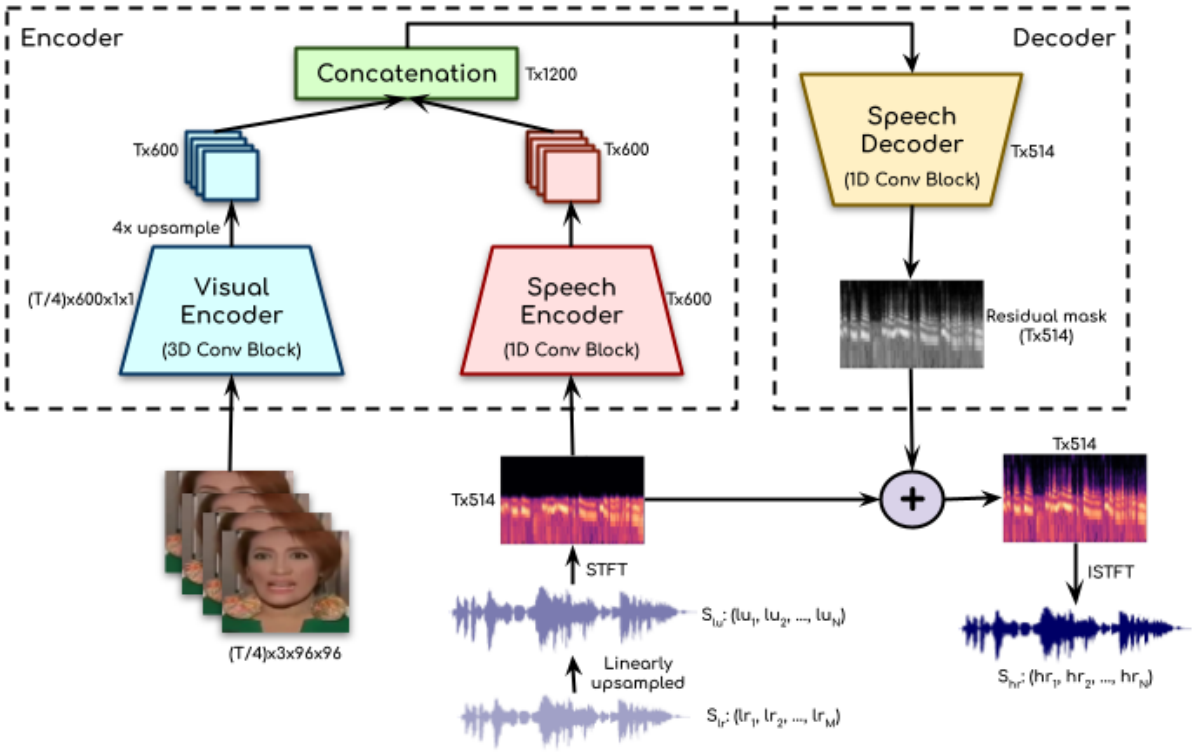

Figure 6.1 The proposed audio-visual network for speech super-resolution at large scale factors ($8\times$ and $16\times$) is illustrated. Three major components are comprised in the SR model: (i) visual encoder, (ii) speech encoder, and (iii) speech decoder. A sequence of frames is ingested by the visual encoder and processed, and visual embeddings are generated. The speech encoder takes the spectrogram representation from the linearly upsampled speech signal to create speech embeddings. These learned visual and speech embeddings are then fused and subsequently processed by the speech decoder. A residual mask is output by the network, which is added to the input spectrogram to generate realistic, high-quality (16kHz) speech signals.

$T \times 600$. Through this process, a robust and efficient translation of raw speech waveforms into a format that can be effectively utilized for further speech-processing tasks is ensured.

6.1.3.2 Visual Encoder

The visual features from the visual stream input I are extracted using the visual encoder. The visual encoder is designed to process the input frames of dimension $(\frac{T}{4}, 3, 96, 96)$ (Here $N = \frac{T}{4}$) by gradually reducing the spatial dimension to $(\frac{T}{4}, 600, 1, 1)$ using a stack of 3D convolution layers with residual connections. The visual encoder is similar to the visual stream of the "Perfect Match" model [18]. The short-range motion information is captured using a temporal receptive field of 5 frames in the first convolution layer. The output of the visual encoder is upsampled 4-times along the temporal axis using nearest neighbor interpolation to match the spectrogram temporal dimension T. Thus, the visual embeddings of dimension $(T, 600)$ are finally obtained.

Table 6.1 Details of the speech encoder.

Layer	# Filters	Kernel	Stride	Residual	Output
input	-	-	-	-	$T \times 257$
conv1	600	3	1	$\times$	$T \times 600$
conv2	600	3	1	✓	$T \times 600$
conv3	600	3	1	✓	$T \times 600$
conv4	600	3	1	✓	$T \times 600$
conv5	600	3	1	✓	$T \times 600$
conv6	600	3	1	✓	$T \times 600$
conv7	600	3	1	$\times$	$T \times 600$

The visual encoder ingests the input frame sequences of dimension $T/4 \times 3 \times 96 \times 96$. It generates visual embeddings of dimension $T \times 600$ using 3D convolution layers as described in Table 6.2.

Table 6.2 Details of the visual encoder.

Layer	# Filters	Kernel	Stride	Residual	Output
input	3	-	-	-	$^T/_4 \times 3 \times 96 \times 96$
transpose1	-	-	-	-	$3 \times {}^T/_4 \times 96 \times 96$
conv1	32	$5 \times 5 \times 5$	$1 \times 2 \times 2$	$\times$	$32 \times {}^T/_4 \times 48 \times 48$
conv2	32	$5 \times 5 \times 5$	$1 \times 1 \times 1$	✓	$32 \times {}^T/_4 \times 48 \times 48$
conv3	64	$1 \times 3 \times 3$	$1 \times 2 \times 2$	$\times$	$64 \times {}^T/_4 \times 24 \times 24$
conv4	64	$1 \times 3 \times 3$	$1 \times 1 \times 1$	✓	$64 \times {}^T/_4 \times 24 \times 24$
conv5	128	$1 \times 3 \times 3$	$1 \times 2 \times 2$	$\times$	$128 \times {}^T/_4 \times 12 \times 12$
conv6	128	$1 \times 3 \times 3$	$1 \times 1 \times 1$	✓	$128 \times {}^T/_4 \times 12 \times 12$
conv7	256	$1 \times 3 \times 3$	$1 \times 2 \times 2$	$\times$	$256 \times {}^T/_4 \times 6 \times 6$
conv8	256	$1 \times 3 \times 3$	$1 \times 1 \times 1$	✓	$256 \times {}^T/_4 \times 6 \times 6$
conv9	512	$1 \times 3 \times 3$	$1 \times 2 \times 2$	$\times$	$512 \times {}^T/_4 \times 3 \times 3$
conv10	512	$1 \times 3 \times 3$	$1 \times 1 \times 1$	✓	$512 \times {}^T/_4 \times 3 \times 3$
conv11	600	$1 \times 3 \times 3$	$1 \times 3 \times 3$	$\times$	$600 \times {}^T/_4 \times 1 \times 1$
conv12	600	$1 \times 1 \times 1$	$1 \times 1 \times 1$	$\times$	$600 \times {}^T/_4 \times 1 \times 1$
transpose2	-	-	-	-	$^T/_4 \times 600 \times 1 \times 1$
squeeze	-	-	-	-	$^T/_4 \times 600$
upsample	-	-	-	-	$T \times 600$

6.1.3.3 Speech Decoder

The speech decoder's aim is to generate a residual mask, which is added to the input spectrogram S_{lr} to obtain the output spectrogram. Initially, the learned speech and visual embeddings are fused in the latent space to form $(T, 1200)$-dimension features. These features are ingested by the decoder, which consists of 1D convolution layers, and a residual mask of dimension $(T, 514)$ is output. In the experiments, the addition mask is used to get the spectrogram, as it was found that the output quality is significantly better than using multiplicative masks (see Table 6.10 for comparison). The

mean absolute error ($L1$) between the generated and the ground-truth HR spectrograms is used as the loss to train the network. Finally, the inverse-STFT (ISTFT) is used to obtain the speech from the generated spectrograms.

Table 6.3 Details of the speech decoder.

Layer	# Filters	Kernel	Stride	Residual	Output
input	-	-	-	-	$T \times 1200$
conv1	1024	3	1	$\times$	$T \times 1024$
conv2	1024	3	1	✓	$T \times 1024$
...					
conv13	1024	3	1	✓	$T \times 1024$
conv14	1024	3	1	✓	$T \times 1024$
conv15	514	3	1	$\times$	$T \times 514$

6.1.4 Speech Super-Resolution using Pseudo-Visual Stream

During inference, the formulation can be modified to work with a synthetically generated talking head video. In cases where the original visual inputs $I = (I_1, I_2, ..., I_N)$ are unavailable, synthetic visual inputs are used instead. These synthetic inputs are denoted as $I = (I'_1, I'_2, ..., I'_N)$. This modification allows the model to generate speech using synthetically generated lip movements, even in scenarios where real visual data is absent. But standard pre-trained synthetic talking head generation models expect clean speech as input. Therefore, a new model to generate synthetic lip movements from low-resolution speech is required.

6.1.4.1 Synthetic generation of frames from degraded speech

In the proposed audio-visual model, frontal or near-frontal talking-face videos of the speaker are required as input. However, it is observed that situations can arise, especially in real-world applications, where the visual stream may be corrupted, unreliable, or entirely absent. Videos where lip movements are occluded, out-of-focus, or out-of-sync with the speech cannot be considered suitable visual stream input. For such cases, a synthetic generation of the visual stream using the proposed pseudo-visual model is proposed. It should be noted that the speech SR network is trained only using the real visual stream (as it is available during training) but can ingest the pseudo-visual stream during testing. This adaptation to synthetic data during inference clearly demonstrates the proposed SR network's capability and robustness.

6.1.4.2 Can the current lip synthesis models be readily used for noisy speech inputs?

It was found that the current state-of-the-art unconstrained speech-to-lip models [25, 27, 134], which work for arbitrary speakers, voices, and languages, are highly inaccurate on noisy speech segments. This inaccuracy is understandable, as these models were never designed to tackle such cases. Moreover,

these methods are speaker-independent and are designed to work on unconstrained videos by training on thousands of speakers with substantial variations in pose, expressions, backgrounds, etc. It was determined that naively fine-tuning the pre-trained lip synthesis model on noisy speech is not ideal. However, for the task at hand, a lip synthesis model trained on thousands of speakers is unnecessary; a single speaker is sufficient.

6.1.4.3 A single identity is all you need.

A sequence of accurate lip movements is needed, preferably on a static image of a single identity where only the lips move according to the speech. This task is relatively easier and well-aligned with the requirements. If the only visual changes are in the lip shapes, the model will naturally focus on learning more accurate, fine-grained speech-lip correspondences. It should be noted that this identity-specific model must work for any speech in any voice and language. The challenge is to train a lip synthesis model that can lip-sync for a single identity image while handling any speech.

6.1.4.4 Distilling lip motion knowledge for a single identity

The fact that the current state-of-the-art model, Wav2Lip [27], can generate accurate lip motion for arbitrary static face images conditioned on any *clean* speech is exploited. The core idea is to achieve this accuracy using noisy speech (the harder part) as input but on just a single identity (the easier part). To accomplish this, a student network is trained to map the noisy speech inputs to lip motion on a static face image. Wav2Lip is employed as the teacher network, and its predictions on the same identity image but with clean speech inputs are used.

6.1.4.5 A Visual Noise Filter

It is hypothesized that since the only visual differences are in the lip shapes, a strong correspondence between the underlying speech and the lip motion is forced to be learned by the student network. Furthermore, noise cannot be meaningfully represented in the generated images by the student network, which is thereby forced to represent only the speech components that the teacher network accurately indicates. Thus, the images generated by the student network act as a ”visual noise filter,” manifesting only the speech component for the downstream speech enhancement network.

6.1.4.6 Training the Student Model

As described above, a student model is trained by learning from a pre-trained lip-synthesis network as a teacher. The student model, a simple encoder-decoder model, inputs a low-resolution speech segment and outputs a lip-synced mouth region of a pre-determined person. This model is adapted from the Wav2Lip architecture [27] by discarding the face identity branch because lip movements need to be generated only for a single identity image.

6.1.4.7 Learning from a Lip Synthesis Teacher

To train the student model, an accurate lip-synced ground truth of a single target is needed. This is obtained from the teacher, a pre-trained lip-synced network, by feeding the clean speech S as the audio input. The audios present in the LRS3 [29] dataset are used as the clean speech data. For the lip synthesis teacher, Wav2Lip [27], a publicly available[1] state-of-the-art speech-to-lip synthesis model, is used.

As it is a speaker-independent model, an identity image also needs to be fed. A near-frontal face image of Taylor Swift is chosen, on which Wav2Lip morphs the lips to match the clean speech inputs. The lip-synced output from Wav2Lip is accurate, as the audio is clean. Furthermore, the output is always the same face image, with only the lip and jaw regions changing while the rest of the face regions remain static. The lower half of the generated face output containing Wav2Lip's prediction is used as the ground truth for the student model. Thus, the new student network is trained to generate correct lip movements (matching the clean speech) given a noise-corrupted input of the same speech.

The student network is trained to minimize the $L1$ loss between its predicted images and the lip-synced ground truth from Wav2Lip. This network is trained for $150K$ iterations with a batch size of 64 on a single NVIDIA RTX 2080Ti GPU. Other hyper-parameters are the same as those of Wav2Lip [27]. For the speech enhancement model in the next section, this trained student network is used to generate lip motion when given a noisy speech segment. The overall method is depicted in Figure 6.2.

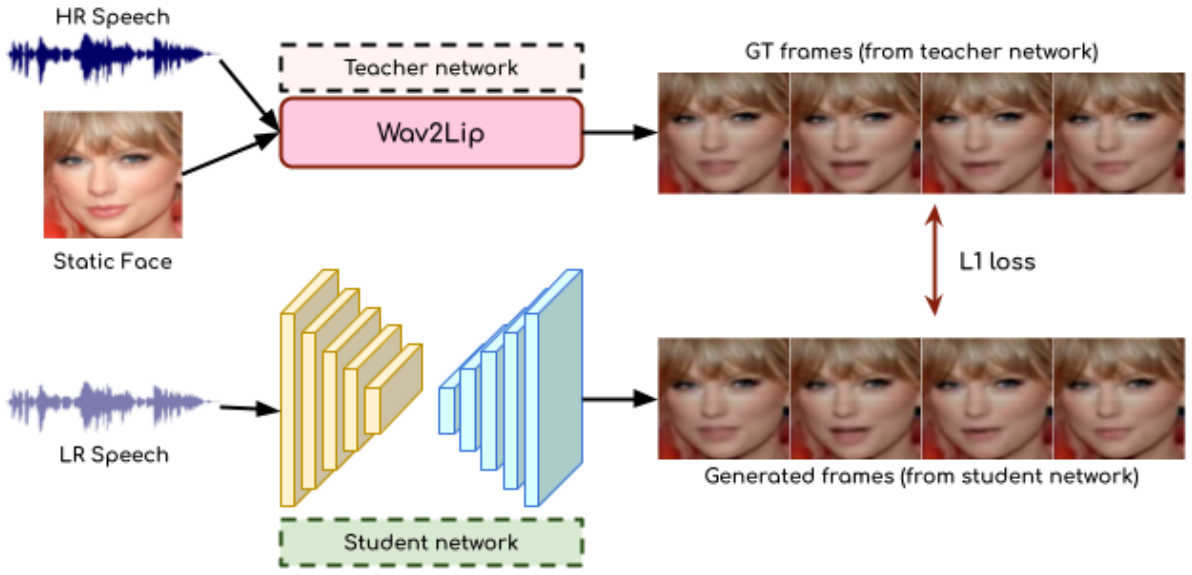

Figure 6.2 The applicability of the proposed SR network is demonstrated by synthesizing the lip movements in cases where the visual stream is absent. A student-teacher network is set up to generate the visual stream from the LR speech input synthetically. The student model is trained to imitate the outputs from the pre-trained teacher model (Wav2Lip [27]), which ingests the HR speech and a static identity to produce accurate lip movements.

[1] https://github.com/Rudrabha/Wav2Lip

6.1.5 Experiments and Results

6.1.5.1 Dataset and Training Settings

6.1.5.1.1 Dataset The publicly available VoxCeleb2 [31] dataset, which consists of over 1 million utterances for ~ 6000 identities, is used. This dataset is highly challenging and popular due to the wide variations in identities, languages, and extensive vocabulary. For testing, the official test split from the VoxCeleb2 dataset is used. Note that there are no overlaps between the identities used in the training and the testing set; thus, evaluating it demonstrates the generalization ability of the model on completely unseen identities.

6.1.5.1.2 Training Setup The network is trained by randomly sampling a 1-second audio segment at 16kHz and its corresponding video frames from the VoxCeleb2 train set. The linear spectrograms extracted from the audio waveform correspond to $T = 100$ timesteps for a 1-second segment. The corresponding frames are considered at 25 FPS and are resized to 96×96 before being fed to the visual encoder. All experiments are performed at scale factors of $4\times$, $8\times$, and $16\times$ with the fixed output resolution of 16kHz. The network is trained using the Adam Optimizer [130] with a learning rate of 10^{-3} and a batch size of 32, stopping the training when the validation loss plateaus. In the experiments, the model was trained for 50 epochs.

6.1.5.2 Results

The results of the audio-visual speech super-resolution for scale factors of $4\times$, $8\times$, and $16\times$ are now presented. The various existing approaches used for comparison are first discussed. This is then followed by the quantitative evaluation (Section 6.1.5.3) along with details of the different metrics used. Finally, a human evaluation (Section 6.1.5.4) is conducted to highlight the real-world applicability of the approach.

6.1.5.2.1 Comparison Comparisons are started with standard "linearly interpolated" outputs. Next, since the existing works in the speech SR literature are limited to lower scale factors, they are trained on the same dataset as the model at all the scale factors for a fair comparison. Additionally, the network's audio-only (AO) baseline is trained by discarding the visual stream input. Thus, comparisons are made against the following models: (i) Linear interpolation, (ii) DNN [165], (iii) U-Net [166], (iv) TFiLM [167], (v) NU-Wav [168], and (vi) AO baseline.

6.1.5.3 Quantitative Evaluation

Qualitative results from the Audio-Visual Speech Super-resolution model proposed in this chapter are presented as a video linked in Figure 6.3, providing a dynamic and illustrative demonstration of the research findings. However, it should be noted that opening the document in Adobe Acrobat Reader

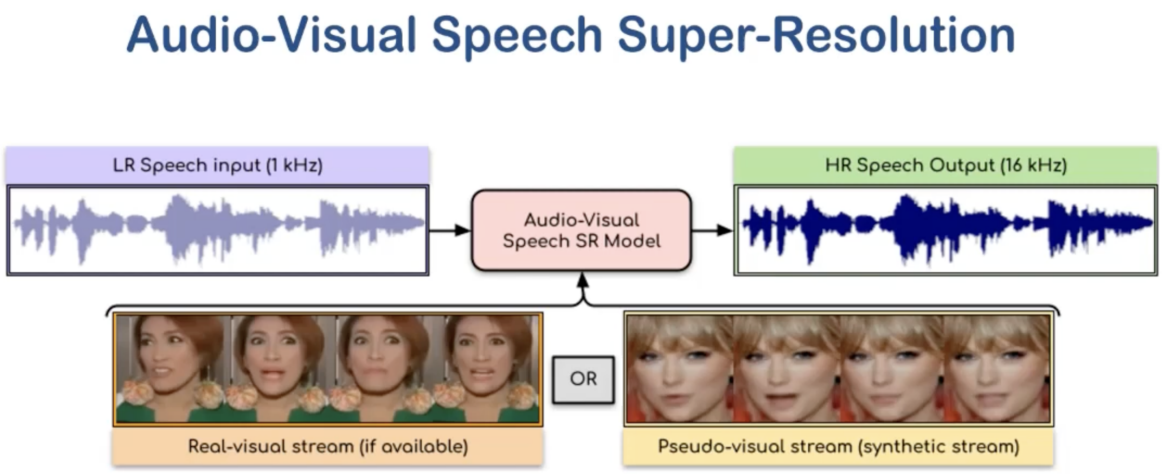

Figure 6.3 A video is presented in this link: `https://youtu.be/bc0ZsTmhLM0`. This video contains qualitative results and comparisons from the Audio-Visual Speech Super-resolution model proposed in this chapter. This image is presented as the thumbnail for this video.

is essential to access this feature, as other PDF viewers may not support the interactive multimedia functionality.

6.1.5.3.1 Evaluation Metrics Popular speech metrics are used to measure the quality of speech generation, as done in the previous chapters. The Perceptual Evaluation of Speech Quality (PESQ) [133], which estimates the perceptual quality of the generated speech, is reported. To evaluate the intelligibility of speech, the Short-Time Objective Intelligibility (STOI) [131] and Extended Short-Time Objective Intelligibility (ESTOI) [132] are computed. Finally, as done in most of the speech SR works [166–168], the Log-spectral Distance (LSD) [170] metric is also reported. It should be noted that these metrics have also been used in the previous three chapters.

Table 6.4 shows the results at scale-factors of $4\times$, $8\times$ and $16\times$. It can be observed that at smaller scale factors like $4\times$, the performance of all the approaches is very similar; the boost obtained using the visual stream is insignificant. However, as the scale factor increases, all the audio-only methods struggle to recover plausible speech outputs. At higher scale factors of $8\times$ and $16\times$, the proposed method outperforms the other methods by a large margin, especially in perceptual quality. It is interesting to note that the proposed pseudo-visual model not only surpasses all the current techniques but is also very close to the proposed approach that uses the real-visual stream. This validates the precise lip shape generations of the proposed pseudo-visual network. Sample spectrograms shown in Figure 6.4 (a) depict that the proposed model successfully reconstructs the high-frequency elements even from very low-resolution inputs.

Table 6.4 Quantitative comparison of different approaches at scale-factors of $4\times$, $8\times$ and $16\times$. The proposed method outperforms the existing audio-only approaches by a large margin, illustrating the benefits from the visual stream.

Scale factor	Method	Linear	DNN [165]	U-Net [166]	TFiLM [167]	NU-Wav [168]	AO baseline	**Ours (pseudo)**	**Ours**
$4\times$	PESQ↑	3.289	3.304	3.318	3.342	3.397	3.363	3.416	**3.429**
	STOI↑	0.871	0.888	0.904	0.889	0.892	0.912	0.916	**0.917**
	ESTOI↑	0.739	0.819	0.825	0.837	0.855	0.843	0.861	**0.869**
	LSD↓	6.112	6.012	6.004	5.803	5.801	5.799	**5.686**	5.694
$8\times$	PESQ↑	2.330	2.243	2.268	2.275	2.219	2.399	2.401	**2.814**
	STOI↑	0.756	0.749	0.765	0.771	0.774	0.804	0.818	**0.832**
	ESTOI↑	0.590	0.638	0.667	0.681	0.663	0.705	0.721	**0.755**
	LSD↓	10.79	7.681	7.325	6.830	9.541	6.220	6.014	**5.069**
$16\times$	PESQ↑	1.842	1.639	1.651	1.654	1.526	1.925	2.188	**2.237**
	STOI↑	0.550	0.653	0.671	0.684	0.598	0.702	0.726	**0.762**
	ESTOI↑	0.327	0.432	0.480	0.551	0.482	0.593	0.614	**0.651**
	LSD↓	11.405	9.306	8.993	8.082	9.780	7.841	6.601	**5.500**

In addition to the qualitative evaluation of the VoxCeleb2 dataset [31], the model is further assessed on the official LRS2 dataset [28] test set. Note that the model is not fine-tuned on the LRS2 [28] dataset; thus, evaluating it demonstrates the generalization ability of the model on new datasets (with significantly different pre-processing and image resolutions used during data collection).

Table 6.5 Quantitative comparison of different approaches at scale factors of $8\times$ and $16\times$ on LRS2 [28] dataset.

Scale factor	$8\times$				$16\times$			
Method	PESQ↑	STOI↑	ESTOI↑	LSD↓	PESQ↑	STOI↑	ESTOI↑	LSD↓
Linear	2.201	0.753	0.558	10.771	2.017	0.570	0.318	11.612
TFiLM [167]	2.191	0.768	0.675	7.103	2.042	0.679	0.491	8.623
NU-Wav [168]	2.250	0.761	0.651	7.946	2.005	0.628	0.524	8.711
AO baseline	1.914	0.802	0.692	6.242	1.706	0.701	0.525	8.007
Ours (pseudo)	2.584	0.808	0.702	6.005	2.616	0.739	0.622	6.991
Ours	**2.805**	**0.815**	**0.725**	**5.197**	**2.637**	**0.766**	**0.649**	**5.838**

In line with the results on the VoxCeleb2 dataset [31], the model performs remarkably well compared to the existing audio-only approaches, as shown in Table 6.5 on the LRS2 dataset [28]. The model performs consistently better at both scale factors, thereby significantly improving the generated speech quality and intelligibility. Additionally, the pseudo-visual model achieves a substantial boost compared to all the audio-only approaches, although neither the pseudo-visual nor the speech SR models have been fine-tuned on LRS2 data [28]. This demonstrates the robustness of the method and its ability to generalize to different identities and datasets.

6.1.5.3.2 How does the performance vary when the scale factor increases? In Figure 6.4 (b), the performance of different models at various scale factors ($4\times$, $8\times$, $16\times$, $24\times$, and $32\times$) is compared.

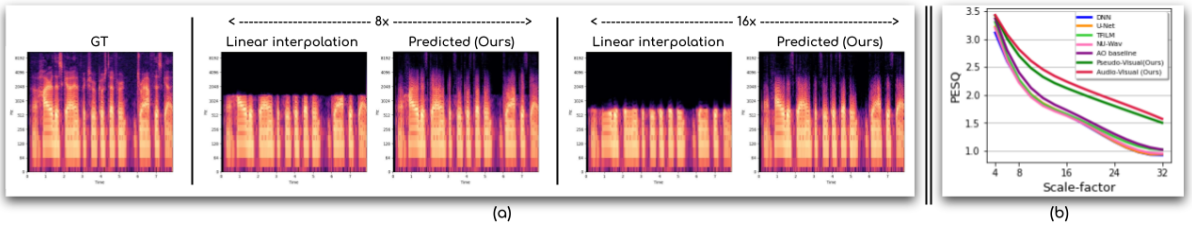

Figure 6.4 (a) Spectrograms of the ground-truth (GT), linearly upsampled speech, and the proposed predicted speech. It can be observed that the proposed network can reconstruct the LR speech, which is close to the GT speech, even at large-scale factors. (b) Performance comparison (metric: PESQ) at different scale factors. At higher scale factors, the gap in the performance of "audio-only" and "audio-visual" methods emphasizes the importance of the visual stream at larger scales.

The gap between the existing audio-only approaches and the proposed model is clearly noticeable. This difference in performance increases with the increase in scale factor. Although there is room for improvement at scale factors of $24\times$ and $32\times$, the impact and usefulness of the visual stream at these scales are impressive. It allows the model to recover the lost information at larger factors that would otherwise be much harder to retrieve by solely using the audio modality.

6.1.5.3.3 Computation comparison In Table 6.6, the number of parameters and the inference time for all the models are compared. Except for NU-Wav [168], the parameters of the "audio-visual" model are similar to those of other "audio-only" models. It should be noted that although NU-Wav has fewer parameters than the proposed model, in terms of performance, it surpasses NU-Wav by a large margin, especially at higher scale factors. To compare the inference time, a 1-second audio segment is processed on a single NVIDIA Geforce RTX 2080Ti GPU. As shown in the table, the audio-visual model is faster (2^{nd} best) compared to most of the existing audio-only models. This is mainly because all the other approaches (except the AO baseline) operate at the waveform level. In contrast, the spectrogram approach is taken, which is considerably faster and also better in terms of performance.

Table 6.6 Comparison of the model size (in million parameters) and the inference time (in seconds). The proposed "audio-visual" model has parameters similar to most of the "audio-only" approaches, with a very low inference time.

	DNN [165]	U-Net [166]	TFiLM [167]	NU-Wav [168]	AO baseline	**Ours (pseudo)**	**Ours**
# params (M)↓	69.9	70.9	68.2	3.0	8.1	90.0	69.3
inf. time (sec)↓	1.113	1.268	0.971	2.921	0.638	0.929	0.873

6.1.5.4 Human Evaluation

A human study was conducted to assess the perceptual quality of the speech generations. 15 samples were randomly selected from the test set of the VoxCeleb2 dataset [31], and the super-resolved signals

Figure 6.5 Activation maps of the visual encoder for different identities. Although the proposed model is highly attentive to the lip region, the contributions from other facial areas, such as eyes and cheeks, are also noteworthy.

were generated at a scale factor of $16\times$. The outputs from the proposed approach and all the comparison methods were played in random order. 30 participants were asked to rate each of these speech samples on a scale of 1-5 based on (a) Quality and (b) Intelligibility. The participant group consisted of people in the age group of 20-50 and had a nearly equal male-female ratio. The mean opinion scores (MOS) are reported in Table 6.7. In line with the quantitative evaluations, the method generated speech that was largely preferred over the other methods.

Table 6.7 Mean opinion scores of different methods based on: (i) Quality and (ii) Intelligibility. The proposed method generates plausible speech outputs with higher perceptual satisfaction.

Measure	Linear	TFiLM [167]	NU-Wav [168]	AO baseline	**Ours (pseudo)**	**Ours**
Quality	2.057	2.571	2.343	2.643	3.152	**3.415**
Intelligibility	1.928	2.685	2.369	2.599	3.064	**3.282**

6.1.6 Ablation Studies

Several ablation experiments were performed to analyze various aspects of the model. All the experiments were conducted for $16\times$ SR on the VoxCeleb2 test set [31].

6.1.6.1 What kind of Visual Input is the Best?

Different forms of the visual input were analyzed: (i) the lower half of the face containing the lip and jaw region and (ii) the full face. As observed in Table 6.8, providing the full face performed better. This is also reflected by the activation map in Figure 6.5, which shows that the facial regions like the eyes, cheeks, and forehead also play a crucial role along with the significant attention on the lip and jaw regions.

Table 6.8 Feeding full face to the visual encoder achieves better performance.

Method	PESQ↑	STOI↑	ESTOI↑	LSD↓
Lower half	2.425	0.743	0.638	5.633
Full face	2.237	0.762	0.651	5.500

6.1.6.2 Robustness to Noise

The robustness of the network in handling noisy inputs is demonstrated here. Gaussian noise was added at three SNR levels: 5dB, 10dB, and 15dB to the LR speech input. As observed from Table 6.9, the model can generate plausible speech outputs even for severely degraded speech inputs.

Table 6.9 The propsoed model is robust to noisy inputs and generates plausible speech outputs.

Noise level	PESQ↑	STOI↑	ESTOI↑	LSD↓
5dB	2.035	0.702	0.586	6.253
10dB	2.062	0.711	0.602	6.190
15dB	2.075	0.714	0.619	6.033

6.1.6.3 Additive Mask v/s Multiplicative Mask

The performance of different types of masks used in prior works was investigated: (i) addition mask (used in this work), (ii) multiplication mask (used in [120]), and (iii) complex ratio mask (cRM) (used in [63]). The results are reported in Table 6.10. It was observed that although several works benefit from the popular cRM, in this case, a simple addition mask performs better than the other kinds of masks.

Table 6.10 Addition mask achieves better performance compared to multiplication masks.

Masks	PESQ↑	STOI↑	ESTOI↑	LSD↓
Addition (Ours)	2.237	0.762	0.651	5.500
Multiplication	2.171	0.702	0.601	6.042
cRM	2.217	0.698	0.612	5.848

6.1.6.4 Importance of the Student Network

The need for a student model to synthetically generate the visual stream during inference (if the real visual stream is absent or unreliable) was assessed. Table 6.11 compares directly using the teacher Wav2Lip model [27], Wav2Lip trained on the LR inputs, and the proposed student network. The teacher Wav2Lip was fine-tuned on the VoxCeleb2 [31] dataset for a fair comparison, and the linearly upsampled speech signal was given as input (Wav2Lip takes speech inputs at 16kHz). As observed in Table 6.11, directly using the teacher model fails to generate plausible speech; this is evident as this network was not intended to work on LR inputs. The teacher model trained (from scratch) on LR inputs also gives poor performance. The best results are obtained using the proposed student model, which learns to imitate the accurate teacher model's output, thus validating the claim of the student-teacher setup.

6.1.6.5 Model's Variation to Identity Attributes

The behavior of the speech SR model on identity attributes such as gender and age (from the test set of VoxCeleb2 data [31]) was analyzed, as shown in Table 6.12. For gender classification, a gen-

Table 6.11 The student network yields the best performance compared to other alternatives.

Pseudo-visual models	PESQ↑	STOI↑	ESTOI↑	LSD↓
Teacher Wav2Lip [27]	1.012	0.637	0.553	8.628
Wav2Lip trained on LR	1.684	0.690	0.581	7.647
Student network (Ours)	2.237	0.762	0.651	5.500

der detection tool [151] was used, which automatically categorizes the identities into male and female categories. To identify the age of the speakers, the public implementation from[2] was used. As seen in Table 6.12, the speech SR network is consistent across the different age groups, but scores vary slightly across the gender of the identities.

Table 6.12 Effect of the identity attributes such as gender and age on model's performance.

Attribute	**Class**	PESQ↑	STOI↑	ESTOI↑	LSD↓
Gender	Female	2.562	0.740	0.687	5.515
	Male	2.520	0.784	0.636	5.504
Age	< 25	2.213	0.776	0.665	5.652
	$25 - 50$	2.287	0.751	0.650	5.727
	> 50	2.271	0.773	0.638	5.452

6.1.6.6 Comparison of Pseudo-Lip Identities

The performance of the pseudo-visual model was compared when different static identities were used for the generation of lip movements. Table 6.13 shows the results on the test set of VoxCeleb2 data [31]. As seen in the table, the model's performance does not vary across the pseudo-lip identities.

Table 6.13 The proposed pseudo-visual model is invariant to pseudo-lip identities.

Identities	PESQ↑	STOI↑	ESTOI↑	LSD↓
Taylor Swift	2.237	0.762	0.651	5.500
Paul McCartney	2.218	0.763	0.643	5.618
Barack Obama	2.239	0.758	0.645	5.498

6.1.7 Summary of the audio-visual super-resolution

This work presents the first audio-visual network for super-resolving speech signals, demonstrating that lip-to-speech synthesis aided by degraded speech is effective. In fact, it is proven that the method can now tolerate significant degradation. While previous works were restricted to $4\times$ SR, the proposed method effectively super-resolves at higher factors of $8\times$ and $16\times$ by incorporating it into the lip-to-speech pipeline. The importance of the visual stream is emphasized, particularly in handling very

[2] `https://github.com/yu4u/age-estimation-pytorch`

low-resolution inputs and significantly improving the generated speech quality. The real-world applicability of the method is shown in handling "in-the-wild" speech signals without an associated visual stream. The designed pseudo-visual model accurately synthesizes lip movements solely from the LR speech input. The method achieves a considerable boost over state-of-the-art audio-only approaches in quantitative metrics and user studies. This work takes a significant step forward in the audio-visual space.

6.2 Audio-Visual Speech Denoising

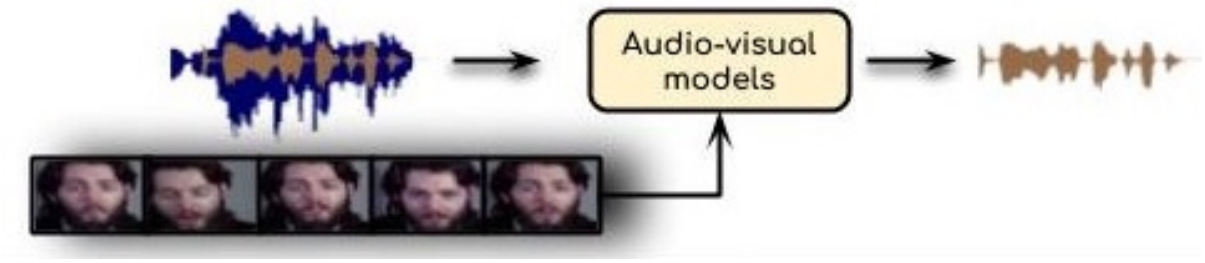

Figure 6.6 Schematic representation of lip-to-speech Synthesis with integration of noisy speech Input, depicting the process of converting visual lip movement data into synthesized speech while incorporating an additional noisy speech signal for enhanced realism and robustness.

Speech super-resolution (SR) is practically important for transmission and other related applications. However, in day-to-day life, speech denoising is likely more crucial, as illustrated by the scenario of trying to communicate with a friend over the phone while on a crowded bus. The noise of the bus, the wind, and the nearby moving vehicles make it nearly impossible for clear communication to occur, underscoring a ubiquitous daily challenge: speech corruption by ambient noise. Speech enhancement serves as an essential tool, particularly as work-related meetings increasingly take place through phone calls from home environments. The applications of this technology extend significantly beyond voice calls, influencing numerous aspects of contemporary life. In the entertainment industry, the capability to isolate human speech from background music proves invaluable for generating automatic subtitles or lyrics for films and music content. This functionality not only enhances the viewing experience for general audiences but also significantly improves accessibility for individuals with hearing impairments, thereby broadening the reach and inclusivity of multimedia content. In the world of vlogging and content creation, which has seen a significant surge in recent years[3], speech enhancement plays a pivotal role. Independent content creators often grapple with filtering out outdoor noises prevalent in their vlogs and short movies. Effective speech enhancement can dramatically improve the audio quality of these videos, making them more appealing to viewers. Moreover, speech enhancement finds critical applications in large public gatherings, such as conferences or sports events, where ensuring clear speech amidst the cacophony of background sounds is crucial. Improved sound systems, equipped with

[3] https://www.omnicoreagency.com/youtube-statistics/

advanced speech enhancement technologies, can significantly enhance the auditory experience for attendees, ensuring that speeches and announcements are heard clearly. Preserving historically important speeches is another vital application. By enhancing the audio quality of these recordings, cultural and historical heritage can be safeguarded for future generations, making these valuable resources more accessible and comprehensible. Lastly, and perhaps most importantly, speech enhancement is essential for everyday voice calls. Enhancing noisy speech over voice calls can vastly improve communication quality, reducing misunderstandings and frustration in an increasingly connected world. Each of these applications demonstrates the wide-ranging impact and necessity of speech enhancement in daily life, from personal communication to preserving historical records.

Similar to the speech SR discussed in the previous section, the concept of lip-to-speech synthesis can be extended to address the additional challenge of denoising speech corrupted by background noise. In this section, the focus is on audio-visual speech denoising, which is also studied in famous works like [63, 120]. Unlike traditional methods that only use audio for speech enhancement, these approaches combine both the visual aspect of lip movements and the auditory signal. This combination is particularly effective in scenarios where external noise heavily distorts the audio. By using visual cues from lip movements, speech can be more accurately separated from the noise, leading to clearer and more understandable audio output. This method is a step forward in speech denoising, especially in real-world situations where noise is a common problem.

6.2.1 Audio-only Speech Enhancement

The works using only the audio stream, with no additional information to denoise speech, are first reviewed. Classical signal denoising techniques like Wiener filtering [171] became the first popular approach for speech enhancement. However, it was often found to be ineffective in denoising speech in real-world situations, as the Wiener filter requires an estimate of the noise a priori. Like many other problems, deep learning models have recently become increasingly popular for speech enhancement. Initial works [172] used standard denoising auto-encoders and LSTM-based approaches [173, 174] for cleaning noisy speech. Feature-based loss functions were also proposed in [175], while the most popular advancement came from models like [176–178] using generative adversarial networks (GANs), which produce relatively higher quality speech from noisy audio segments.

Even though significant progress has been made in the last few years, speech enhancement models are still confined to being trained on datasets [174, 179–181] that have been collected in constrained environments recorded by a selected set of speakers. The types of noises [182, 183] that are synthetically added to the clean speech while training such models are also limited to a few types. Thus, these models often fail to perform satisfactorily in natural, unconstrained settings. They fail to cope with hundreds of speakers of different dialects and languages, and the level and type of noise vary abruptly in a speech segment. In this work, a method is proposed to generate high-quality clean speech from a given noisy audio and lip movements in such unconstrained conditions.

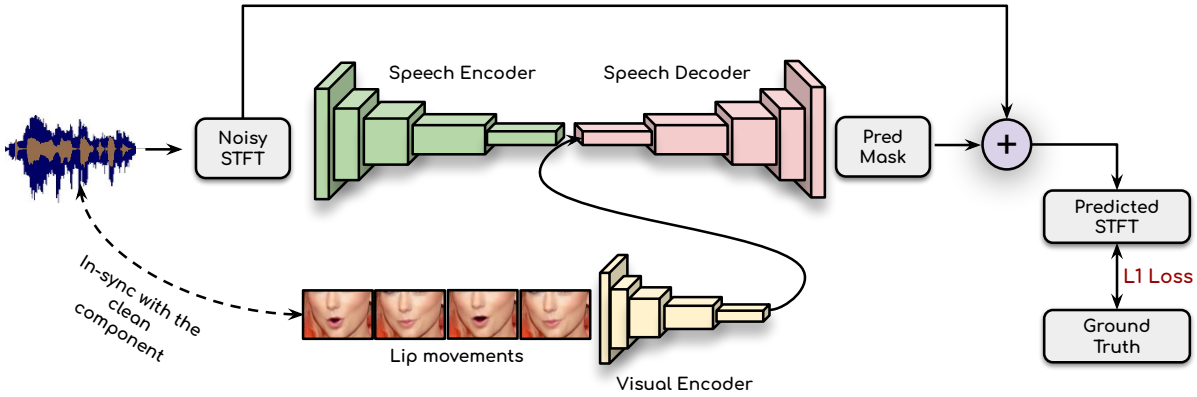

Figure 6.7 The enhancement model ingests the noisy spectrogram along with the lip movements and outputs a mask for clean speech.

6.2.2 Formally formulating the problem

The setup for speech super-resolution is modified to handle noisy speech. A sequence of lip movements $I = (I_1, I_2, \ldots, I_N)$ and a corresponding noisy speech segment $S^{noisy} = (S_1^{noisy}, S_2^{noisy}, \ldots, S_{T'}^{noisy})$ are taken to generate a high-resolution speech segment $S = (S_1, S_2, \ldots, S_{T'})$. The generated speech S should correspond to the lip movements I and be an enhanced version of S^{noisy}. It should be noted that in this work, unlike previous chapters where melspectrograms were used, linear spectrograms are utilized to represent speech. The high-resolution linear spectrogram is denoted as $Y = (Y_1, Y_2, \ldots, Y_T)$. The linear spectrogram from the noisy speech is used as another input to the network in addition to the lip movements. It is denoted by $Y^{noisy} = (Y_1^{noisy}, Y_2^{noisy}, \ldots, Y_T^{noisy})$.

6.2.3 Architecture

The overview of the proposed audio-visual speech denoising network is depicted in Figure 6.7.

6.2.3.1 Audio representation

0.2 seconds of noisy speech, S^{noisy}, is considered as the input. However, instead of representing the audio at the mel-scale spectrograms, the linear spectrogram representation is adopted by using a short-time Fourier transform (STFT). Generating linear spectrograms allows for direct inversion back to a waveform without needing vocoders. To compute the STFT, a window length of 25ms with a hop length of 10ms samples at 16kHz is considered. The computed STFT from the raw audio waveforms is a complex array of time-frequency representation with a dimension of $T \times F$. Here, T is the number of STFT time steps, which corresponds to 20 in the experiments (0.2 second audio segment), and F is the frequency which corresponds to 257. The complex STFT array is further decomposed into the magnitude and the phase components, which act as input to the magnitude and phase sub-networks, respectively. All STFT representations are normalized between $[0, 1]$.

6.2.3.2 Visual representation

In this approach, continuous real-face crops from videos are directly used to represent visual information. A corresponding sequence of face crops is extracted for every 0.2 second of the audio window, aligning with the audio segments. This results in a sequence of 5 face crops that correspond to the lip movements for the given STFT input segment. Each face crop is then passed through a visual encoder, which comprises 12 layers of residual 2D-convolution blocks, to obtain a visual embedding for each frame. Since all input videos are recorded at 25 FPS, 5 frames are effectively aligned with each 0.2 second audio window, providing synchronized visual information for the speech denoising task.

6.2.3.3 Network architecture

The speech enhancement model consists of the magnitude and the phase sub-networks as illustrated in Figure 6.7. Inspired by [120], a phase sub-network is incorporated to avoid the "robotic voice" in the generated speech. These sub-networks are described below.

6.2.3.3.1 Magnitude sub-network The input noisy magnitude spectrogram is processed by the magnitude sub-network, which is a stack of 7 1D-convolution blocks with residual connections. Convolutions are performed along the temporal dimension by considering the frequency component of the input spectrograms as channels. The output of the visual encoder module is $4\times$ up-sampled using nearest-neighbor interpolation to match the spectrogram temporal dimension. The audio and the visual streams are then combined by concatenating the learned features of each stream along the channel dimension. This fused representation is then given to the magnitude decoder, a stack of 14 1D convolution layers with residual connections. The decoder outputs a mask that is added to the input noisy magnitude, followed by a sigmoid activation to generate the enhanced magnitude output between $[0, 1]$. The $L1$ distance between the predicted magnitude spectrogram and the ground truth is minimized.

6.2.3.3.2 Phase sub-network The phase sub-network is conditioned on the enhanced magnitude, the embedded lip movements, as well as the input noisy phase spectrogram. The phase sub-network is similar to the decoder of the magnitude sub-network, and the resultant mask is added to the noisy input phase, followed by a sigmoid activation to produce the clean phase.

Thus, the output from the magnitude sub-network, the learned lip features from the visual encoder, and the input noisy phase representations are concatenated. This fused representation is processed by the phase decoder, which consists of a stack of 14 1D convolution layers. Similar to the magnitude sub-network, a mask is predicted to be added to the input noisy phase, followed by a sigmoid activation to produce the clean phase.

The cosine similarity between the predicted phase and the ground truth is maximized. The final loss function is similar to the one used in [120]. Finally, the predicted magnitude and phase spectrograms are combined, and the enhanced clean waveform is obtained using inverse-STFT (ISTFT). A later ablation

study shows that using cosine similarity instead of $L1$ reconstruction loss for phase produces significant improvements in metrics and quality.

6.2.4 Experiments and Results

6.2.5 Dataset

The publicly available LRS3 dataset [29] is used, which consists of thousands of spoken sentences from TED videos. For training, the "pre-train" and "train-val" sets from the dataset are used, which has around 430 hours of video data with $150K$ utterances. This is a challenging dataset that covers a large number of speakers ($9K$), thus encouraging the trained model to be speaker-independent.

6.2.5.1 Experimental setup

For evaluating in unconstrained settings, the following three synthetic test sets are created, each having three noise levels of 0db, 5db and 10db: (i) Test split of LRS3 [29] + Noise from VGGSound [184], (ii) Test split of LRS3 [29] + Noise from QUT-NOISE-TIMIT [179] (unseen noise), (iii) Test split of LRS2 [28] + Noise from VGGSound [184] (unseen speakers).

6.2.5.2 Evaluation

The following standard speech enhancement evaluation metrics (higher is better) are used to evaluate the proposed method. Perceptual Evaluation of Speech Quality (PESQ) [133] (-0.5 to 4.5) is computed to measure the overall perceptual quality, and the short-time objective intelligibility measure (STOI) [131] (0 to 1) is computed to correlate with the intelligibility of speech. Objective measures such as the mean opinion score (MOS) prediction of the signal distortion (CSIG) (1 to 5), the MOS prediction of background noise (CBAK) (1 to 5), and the overall MOS prediction score (COVL) (1 to 5) are also used.

6.2.5.3 Results

Exhaustive qualitative and quantitative evaluations of the proposed approach under different noise conditions are performed, and the results are compared with existing methods. The speech enhancement network is evaluated for constrained noise conditions on the test split of the TIMIT [181] dataset along with noises from the QUT-NOISE-TIMIT [179] corpus. For evaluation in unconstrained real-world conditions, two synthetic test sets are created by mixing the samples from VGGSound [184] data with the clean speech in the test splits of LRS3 [29] and LRS2 [28] datasets. Standard speech evaluation metrics are used, namely, the perceptual evaluation of speech quality (PESQ), the mean opinion score (MOS) prediction of the signal distortion (CSIG), the MOS prediction of background noise (CBAK), the overall MOS prediction score (COVL), and the short-time objective intelligibility measure (STOI).

Finally, human evaluations are also performed on a newly curated real-world test bed, and the MOS is reported to analyze the effectiveness of the proposed method.

Table 6.14 Quantitative comparison of different approaches. The first section contains clean speech from LRS3 [29] test set mixed with VGGSound [184] noises at different SNR levels. In the second section, the performance on "unseen noises" is specifically evaluated by mixing the LRS3 [29] test set audios with the QUT [184] city-street noises at different noise levels. Finally, in the third section, evaluation is specifically conducted on "unseen speakers" by mixing the speeches of the unseen LRS2 [28] test set speakers with VGGSound [184] noises. The proposed method outperforms the audio-only approaches in all three sections and is comparable ($< 3\%$ difference) to the real visual-stream method.

SNR	0db					5db					10db				
Method	Noisy	[176]	[175]	AO	AV	Noisy	[176]	[175]	AO	AV	Noisy	[176]	[175]	AO	AV
PESQ	1.93	1.84	2.17	2.62	2.80	2.29	2.24	2.52	2.93	3.05	2.66	2.65	2.95	3.12	3.25
CSIG	2.31	2.10	2.82	3.02	3.25	2.79	2.67	3.22	3.26	3.39	3.15	3.17	3.36	3.45	3.56
CBAK	1.83	1.87	2.30	2.36	2.51	2.23	2.30	2.46	2.54	2.71	2.40	2.55	2.63	2.70	2.84
COVL	1.68	1.53	2.02	2.14	2.29	2.04	2.00	2.21	2.29	2.41	2.16	2.26	2.31	2.46	2.58
STOI	0.76	0.75	0.83	0.87	0.90	0.85	0.86	0.88	0.90	0.94	0.88	0.88	0.90	0.92	0.95
PESQ	1.86	1.77	2.14	2.54	2.73	2.26	2.14	2.54	2.83	3.01	2.67	2.61	2.90	3.05	3.21
CSIG	2.46	2.37	2.78	2.93	3.10	2.90	2.86	3.18	3.15	3.33	3.30	3.30	3.38	3.35	3.48
CBAK	1.55	1.89	2.10	2.31	2.46	1.94	2.23	2.30	2.49	2.63	2.42	2.53	2.59	2.62	2.77
COVL	1.69	1.68	1.95	2.04	2.20	2.02	1.99	2.06	2.20	2.34	2.14	2.20	2.30	2.35	2.45
STOI	0.75	0.77	0.80	0.84	0.89	0.85	0.86	0.87	0.89	0.93	0.87	0.90	0.91	0.91	0.95
PESQ	1.94	1.82	2.10	2.58	2.79	2.32	1.87	2.55	2.87	3.04	2.69	2.41	2.79	3.10	3.22
CSIG	2.55	2.29	2.80	3.07	3.23	3.01	2.85	3.16	3.21	3.38	3.23	3.13	3.33	3.36	3.49
CBAK	1.86	1.82	2.22	2.31	2.47	2.28	2.12	2.42	2.48	2.63	2.58	2.59	2.57	2.63	2.73
COVL	1.81	1.59	1.93	2.07	2.20	2.16	2.03	2.14	2.17	2.31	2.20	2.11	2.23	2.27	2.39
STOI	0.75	0.73	0.80	0.84	0.89	0.84	0.83	0.85	0.88	0.90	0.85	0.86	0.89	0.89	0.93

This section evaluates the performance of audio-visual speech enhancement methods, focusing on the effectiveness of using real visual information alongside audio for speech denoising. Synthetic test sets, as described in Section 6.2.5.1, are used for evaluation, with comparisons made to state-of-the-art audio-visual methods.

6.2.5.3.1 Robustness to Various Noise Levels The results of testing on the LRS3 test set mixed with VGGSound data at different noise levels are summarized in the first section of Table 6.14. These results demonstrate the effectiveness of audio-visual methods in speech enhancement, particularly in challenging noise conditions. The audio-visual method consistently performs well across various noise levels, indicating the value of incorporating visual information in speech-denoising processes.

6.2.5.3.2 Robustness to Unseen Noises To assess the generalization capabilities of the audio-visual approach, results with unseen noise types from the QUT (city-street noise) dataset [179] are presented in the second section of Table 6.14. The performance of the audio-visual method remains robust even in the presence of unfamiliar noise types, highlighting its adaptability to diverse real-world scenarios.

6.2.5.3.3 Robustness to Unseen Speakers

An additional comparative study is conducted on unseen speakers from the LRS2 test set, as shown in the third section of Table 6.14. The audio-visual method exhibits remarkable consistency in performance, indicating its robustness to variations in speaker characteristics.

6.2.6 Summary of Audio-Visual Speech Denoising

In summary, the effectiveness of lip-to-speech synthesis, aided by noisy audio, was assessed through a series of rigorous tests. Synthetic datasets were used to compare these methods with leading audio-visual solutions, highlighting their proficiency in handling diverse, challenging noise environments. A key finding from the tests conducted on the LRS3 dataset, combined with VGGSound noises [184], is the remarkable performance of lip-to-speech methods across a spectrum of noise levels, including 0dB, which is typically considered a high noise level. This robustness is further emphasized in scenarios involving unfamiliar noise types, such as city-street noises from the QUT dataset. The lip-to-speech approach maintains its effectiveness, illustrating its versatility and adaptability to various real-world settings.

Moreover, the study extends to evaluate the performance of speakers not represented in the training data, using the LRS2 test set for this purpose. The results consistently indicate a high level of performance, reinforcing the lip-to-speech method's ability to handle variations in speaker attributes effectively. This evaluation highlights lip-to-speech enhancement methods as highly effective tools, showcasing their superior performance and adaptability across different noise conditions and speaker variations. This robustness positions them as valuable assets for practical speech enhancement applications in real-world scenarios, demonstrating that high noise levels can be managed effectively.

Chapter 7

Conclusion

The primary objective of this thesis is to advance the field of lip-to-speech synthesis by generating speech from silent lip movements. This task addressed the significant challenge of the one-to-many relationship between lip movements and spoken language, a complex problem in speech processing research. Initial studies in this domain were largely constrained, focusing on speaker-specific models that limited broader applicability. This thesis aimed to overcome these limitations, with a focus on developing more versatile and generalizable methods for lip-to-speech synthesis. The goal was to extend the utility of these systems beyond controlled environments and specific speakers, thereby enhancing their practicality for real-world applications. The research presented herein has systematically addressed the challenges of constrained settings and speaker dependency, contributing to the development of more adaptable lip-to-speech synthesis technologies. In this thesis, a comprehensive survey of the field was conducted to provide context for the research on lip-to-speech synthesis. The survey, presented in the introduction, aimed to give an overview of progress in related areas. Following this, background information was provided on lip-to-text, text-to-speech, and lip-to-speech technologies. Relevant works in each of these fields were discussed to offer a clear understanding of current advancements. This survey helped in identifying research gaps and potential areas for contribution. The goal was to present a thorough yet accessible overview of the field, acknowledging the work of other researchers and highlighting the interconnected nature of these research areas. This examination was instrumental in shaping the research questions and approaches adopted in this thesis.

The first contribution of this thesis was the extension of lip-to-speech synthesis from the constrained single-speaker approaches prevalent during $2017 - 18$ to an unconstrained single-speaker framework. To facilitate this advancement, a dedicated dataset was meticulously collected and tailored specifically for this purpose. This dataset, characterized by its diversity and comprehensiveness, was instrumental in developing and refining our novel approach, named Lip2Wav. Lip2Wav, a sequence-to-sequence model, represents a significant improvement in the field of lip-to-speech synthesis. This approach was rigorously tested and demonstrated state-of-the-art results in both constrained and unconstrained single-speaker lip-to-speech synthesis scenarios. The Lip2Wav model demonstrated that it is possible to learn the correlation between lip movements and speech given sufficient speaker-specific data. This approach showed improvements in lip-to-speech synthesis, addressing some of the task's inherent complexities.

The model's performance on natural, unscripted speech provided insights into handling the nuances and variations present in real-world scenarios. These findings contributed to the ongoing research in the field of lip-to-speech synthesis.

The second contribution aimed to extend lip-to-speech synthesis into the realm of unconstrained multi-speaker settings introduced a set of formidable challenges, distinctively more complex than those encountered in single-speaker scenarios. Multi-speaker text-to-speech (TTS) synthesis itself stands as a significantly challenging domain, where a single model is tasked with generating speech in the voices of multiple speakers. This complexity is magnified manifold when the task shifts to generating speech solely from lip movements across different speakers. In this multi-speaker context, the model must not only accurately interpret the lip movements corresponding to speech but also capture the unique vocal characteristics of each speaker. This includes varying pitch, tone, accent, and speaking style, which differ substantially from one individual to another. Additionally, in an unconstrained multi-speaker lip-to-speech setting, the model encounters a diverse range of speaking conditions, facial expressions, and backgrounds, all of which add layers of complexity to the speech synthesis process. The challenge here is twofold: first, to develop an algorithm capable of robustly deciphering speech from the silent lip movements of multiple speakers, and second, to ensure that this speech retains the naturalness and individuality of each speaker's voice. Addressing these challenges requires a deep understanding of speech processing and computer vision and innovative approaches in machine learning to handle the variability and complexity inherent in multi-speaker lip-to-speech synthesis.

In addressing the challenge of homophones in multi-speaker lip-to-speech synthesis, a significant focus of this thesis was placed. Homophones are recognized as presenting a unique challenge in this field, as different words with the same lip movements can lead to ambiguous interpretations. An innovative approach based on Variational Autoencoder (VAE) and Generative Adversarial Network (GAN) was introduced to tackle this issue. In this framework, lip movements are first mapped to a distribution of potential homophones. This mapping allows for the creation of a diverse set of potential speech outputs corresponding to a given set of lip movements. The key to this approach is the use of KL divergence loss to effectively tie the lip movement distribution to a corresponding speech distribution. During the training phase, sampling is predominantly done from the speech distribution, a strategy akin to teacher forcing, to enhance the model's ability to generate accurate speech. In contrast, at testing time, the model relies solely on the lip movement distribution for speech generation. Through this method, not only is the homophone challenge addressed, but it is also ensured that the synthesized speech maintains the natural characteristics of the speaker's voice. The VAE-GAN approach is thus considered to represent an advancement in multi-speaker lip-to-speech synthesis, offering a solution to one of the intricate challenges in the field.

Despite the progress made with the VAE-GAN approach in addressing the challenge of homophones, our research revealed that this alone was insufficient for optimal content representation in lip-to-speech synthesis. The intricacies involved in accurately capturing and translating the content from lip movements were further compounded by the vast variations in speech attributes and lip dynamics across

different speakers. These variations posed a significant challenge, making ensuring that the synthesized speech accurately reflected the intended content increasingly difficult. Recognizing this, it became apparent that a more effective method was needed to distill content information from lip movements. The goal was to develop a solution that could more robustly handle the diverse range of lip articulations and speech characteristics inherent in multi-speaker scenarios.

Alongside the contributions in lip-to-speech synthesis, significant advancements were being made in the development of lip-to-text networks by the wider research community. These networks have demonstrated increased accuracy, making them practical for various applications. Recognizing the potential of these developments, the decision was made to integrate lip-to-text capabilities into this research to enhance content distillation in lip-to-speech synthesis. This integration led to the formulation of a visual text-to-speech (VTTS) model. The model begins with a lip-to-text (L2T) network that translates lip movements into textual content. The synthesized text, alongside the original lip movements, then serves as dual input for the VTTS module. This approach ensures that the final speech output is conditioned on the textual interpretation of the lip movements and the visual cues. This dual-input system, leveraging both phonemic context from text and articulation guidance from lip movements, aims to produce a more accurate and naturally sounding speech output, especially in multi-speaker scenarios. It is crucial to note that during inference, the only requirement is the lip movement input. The integrated L2T network first converts these movements into text, which is then utilized by the VTTS module along with the lip movements to generate the final speech output. This process ensures that the system can effectively function in real-world scenarios where only visual cues are available. While this system represents an important advancement in multi-speaker lip-to-speech synthesis, it is part of a broader, dynamic field. It highlights the potential of multimodal integration in speech synthesis and reflects the collaborative nature of scientific progress. This work is viewed as a step in the ongoing journey of speech synthesis research, with the hope that it will pave the way for further innovations in the field.

A significant extension of our research in lip-to-speech synthesis involved incorporating an additional dimension: the integration of noisy or low-quality speech as input alongside lip movements. This advancement aligns with the broader community's efforts in audio-visual speech enhancement but is uniquely positioned in our research framework. By integrating auditory cues, even if noisy or of low quality, with the visual modality of lip movements, we aimed to create a more robust and versatile speech synthesis system. This approach is grounded in the understanding that lip movements and speech, though closely related, can provide complementary information. In scenarios where the audio quality is highly compromised, be it due to background noise, poor recording equipment, or other distortive factors—the visual component of lip movements becomes crucial. It can provide the missing or obscured linguistic details that noisy audio fails to convey. Conversely, even low-quality audio can offer valuable context, such as voice tone, pitch, and rhythm, which are not discernible from visual information alone. In essence, this extension of lip-to-speech synthesis to include noisy audio input is a strategic enhancement of the system's capabilities. It allows the model to leverage the strengths of both audio and visual inputs, resulting in a more accurate and natural speech output under a wider range

of conditions. This method does not merely treat the visual and auditory inputs as separate entities but rather synergizes them, enhancing the overall efficacy of the speech synthesis process. Within the ambit of audio-visual speech enhancement, this research focused on two areas: speech super-resolution and speech denoising. In both cases, highly degraded audio was provided as an additional condition in a lip-to-speech setup. Speech super-resolution involved increasing the sampling rate of low-resolution audio, enhancing clarity and detail while maintaining synchronization with visual lip movements. Speech denoising concentrated on extracting clear speech from noisy audio by integrating visual information from lip movements with the degraded audio input, effectively filtering out background noise. A simple yet robust encoder-decoder setup with dual encoders for lip movements and low-quality audio inputs was utilized for both tasks. The fused encoded representations were processed by a decoder to generate high-quality speech output. This approach demonstrated significant improvements over existing audio-only methods in terms of quality and clarity. These models highlight the potential of combining audio-visual technologies to improve speech synthesis and enhancement, advancing our understanding and application of multimodal approaches in challenging speech processing scenarios.

7.1 Limitations and future research directions

In this section, the limitations of the current research are presented, and potential avenues for future research are proposed. These insights are presented with the understanding that there is significant room for improvement and exploration in the field of lip-to-speech synthesis.

7.1.1 Frontal faces dependency

The current models have shown proficiency with frontal face orientations; however, their performance is observed to decline with non-frontal or angled views, which limits their real-world applicability. To enhance versatility, algorithms with improved robustness to face orientation could be developed. One of the first works in lip-to-speech synthesis utilized multiple views of the face to predict speech [46]. This approach could be further explored and expanded upon. Today, there are numerous novel view generator networks using techniques such as Neural Radiance Fields (NeRFs) [185] and advanced 3D modeling [186], which could be leveraged to synthesize a variety of novel views. While this approach was initially limited by the difficulty in collecting multi-view data, modern synthetic data generation techniques could potentially overcome this challenge. The integration of these sophisticated computer vision algorithms for interpreting facial movements from various angles might be a promising direction for future research, potentially leading to more versatile and robust lip-to-speech systems.

7.1.2 Language specificity

The methods presented in this research have primarily been validated in English, which raises questions about their performance with other languages. It is acknowledged that multi-lingual lip reading

itself has not been extensively attempted to date. While multi-lingual lip-to-speech might be more achievable due to the reduced need for annotations, challenges remain. Datasets such as AVSpeech [63] and VoxCeleb2 [31], which include multiple languages, could potentially be utilized for training. Additionally, a version of LRS3 [187] with 13 languages exists, though it has not been made public. Training on these diverse datasets could help expand the linguistic scope of these models. However, it is recognized that a language classifier might be necessary to select the appropriate language during inference. A similar classifier has already been proposed by Afouras et al. in [188]. Future research could focus on creating diverse linguistic datasets and developing models that are attuned to the nuances of different languages and dialects.

7.1.3 Model Size and Efficiency

The substantial size of our multi-speaker lip-to-speech model proposed in Chapter 5, which is one of the most practically usable models in terms of quality, is acknowledged as a significant challenge for deployment in resource-constrained environments. This model has over 100 million parameters, making it impractically large for deployment on edge devices where many potential use cases of lip-to-speech could flourish. Therefore, the need for lightweight models is evident. Advanced model compression techniques could be researched, and more efficient neural network architectures could be explored to develop lightweight models suitable for edge computing scenarios.

Easy extensions to address this issue include replacing the transformer layers with more efficient alternatives, such as Performer layers [189], which are much more efficient transformers. Additionally, advanced model compression techniques could be researched, and more efficient neural network architectures could be explored to develop lightweight models suitable for edge computing scenarios.

7.1.4 Inherent Challenges in Lip Reading

Despite the advancements made, it is recognized that certain limitations in lip reading persist, particularly in lip-to-text conversion for diverse speakers and unconstrained environments. An area that remains largely unexplored is single-speaker lip-to-text systems, which could potentially offer significant improvements in accuracy and personalization. While it is acknowledged that collecting extensive data (such as 20 hours per speaker) is not practically scalable, even a modest amount of speaker-specific data could provide key insights into individual speaking styles. These insights could inherently improve the performance of lip-reading systems for that particular speaker. This approach of personalization could have a significant impact, especially in medical use cases where data is often scarce, but the need for accurate, patient-specific models is high. For instance, a personalized lip-to-text system could be invaluable in scenarios involving patients with speech impairments or those undergoing speech therapy. Such a system, trained on a limited but focused dataset of the patient's lip movements, could potentially offer more accurate and reliable results compared to generic models. Future research in this direction might explore efficient ways to adapt general lip-reading models to individual speakers with minimal additional data. Techniques such as transfer learning, few-shot learning, or continual learning could be

investigated to enable quick and effective personalization of lip-reading systems. By addressing these aspects, future advancements in lip-reading technology could not only improve overall accuracy but also open up new possibilities for highly specialized and personalized applications, particularly in fields where such customization can make a substantial difference in an individual's quality of life or treatment outcomes.

7.1.5 Voice Attributes Optimization

It is acknowledged that there remains potential for enhancing the accuracy of capturing diverse voice attributes. Future research efforts could be directed toward a deeper understanding of human voice qualities through advanced acoustic modeling and speech prosody analysis. The integration of emotional expression into speech synthesis to produce more natural and emotionally resonant outputs might also be explored. Recent advancements in text-to-speech (TTS) technology have led to the emergence of several noteworthy services and models. Coqui TTS [190], an open-source TTS system, has gained attention for its high-quality voice synthesis and customization options. ElevenLabs [191] has made significant strides in producing natural-sounding voices with emotional nuances. Other services like Resemble AI [192] have also shown promise in creating highly realistic and emotionally expressive synthetic voices. Furthermore, the application of diffusion models in speech synthesis, as demonstrated by recent works such as Grad-TTS [193], has opened new avenues for generating high-fidelity speech. These models have shown potential in capturing fine-grained details of voice characteristics and producing more natural-sounding speech. The rapid progress in these areas suggests that future lip-to-speech systems could potentially leverage these advanced TTS technologies to improve the quality, naturalness, and emotional expressiveness of the generated speech. Integration of these state-of-the-art TTS models with lip-reading systems could lead to more sophisticated and versatile lip-to-speech synthesis capabilities.

7.1.6 Ethical concerns

The ethical implications of lip-to-speech and related technologies are recognized as demanding careful consideration. Primary concerns include the risks of unauthorized surveillance and the creation of deceptive deepfake content, which are acknowledged as significant ethical challenges. It is understood that these technologies have the potential to infringe on individual privacy rights and, at a broader level, could undermine public trust in digital media and communications. Addressing these ethical issues is deemed crucial to ensure that the advancements in this field are used responsibly and do not contribute to the erosion of privacy or the spread of misinformation. Future research efforts might focus on developing robust safeguards and ethical guidelines for the use and deployment of these technologies.

In conclusion, the field of lip-to-speech synthesis has been extensively explored in this thesis, revealing both its complexities and potential. The comprehensive survey and in-depth analysis provided aim to serve as a valuable resource for future research and development in this area. As the field progresses,

it is recognized that foundational techniques that learn fundamental patterns in data are becoming increasingly crucial. These techniques are sought to identify patterns that are common across diverse populations, transcending differences in language, ethnicity, and accents. The contributions made in this thesis are intended to align with this approach, striving to uncover these universal patterns in lip movements and speech. The rapid advancement of the field is acknowledged, with the potential for other facial aspects, such as emotions, to be inherently learned and incorporated into future models. It is anticipated that as more data becomes available and processing capabilities improve, these models will become more sophisticated and versatile. In the broader context, lip-to-speech synthesis is viewed as part of a movement towards more natural human-computer interaction. While its applications are far-reaching, from enhancing accessibility to creating novel solutions in various sectors, the ethical implications of these advancements must be carefully considered. The ongoing challenges in this field are seen as opportunities for further interdisciplinary research and innovation, with the expectation that such efforts will lead to more robust, inclusive, and ethically sound technologies that contribute significantly to breaking down communication barriers. The journey in lip-to-speech synthesis has been both comprehensive and detailed. The goal was to connect complex theoretical foundations with practical applications, aiming for accurate visual speech interpretation and natural audio output. This work contributes to both academic research and practical implementation, enhancing our understanding and supporting future advancements in improving human communication through technology.

Bibliography

[1] A. M. Bell, *Popular Manual of Vocal Physiology and Visible Speech*. N.D.C. Hodges, 1867. [Online]. Available: https://books.google.co.in/books/about/Popular_Manual_of_Vocal_Physiology_and_V.html?id=0ZoOAAAAYAAJ&redir_esc=y

[2] A. G. Bell, "Visible speech as a means of communicating articulation to deaf-mutes," *American Annals of the Deaf and Dumb*, vol. 17, no. 1, pp. 1–21, 1872. [Online]. Available: http://www.jstor.org/stable/44460666

[3] W. PENFIELD and E. BOLDREY, "SOMATIC MOTOR AND SENSORY REPRESENTATION IN THE CEREBRAL CORTEX OF MAN AS STUDIED BY ELECTRICAL STIMULATION1," *Brain*, vol. 60, no. 4, pp. 389–443, 12 1937. [Online]. Available: https://doi.org/10.1093/brain/60.4.389

[4] J. Lucy, "Sapir–whorf hypothesis," in *International Encyclopedia of the Social & Behavioral Sciences*, N. J. Smelser and P. B. Baltes, Eds. Oxford: Pergamon, 2001, pp. 13 486–13 490. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B0080430767030424

[5] V. A. Utzinger, "A study of lip movements in speech," *Quarterly Journal of Speech*, vol. 15, no. 4, pp. 480–484, 1929. [Online]. Available: https://doi.org/10.1080/00335632909360829

[6] H. H. Yeung and J. F. Werker, "Lip movements affect infants' audiovisual speech perception," *Psychological Science*, vol. 24, no. 5, pp. 603–612, 2013, pMID: 23538910. [Online]. Available: https://doi.org/10.1177/0956797612458802

[7] B. Dodd, "Lip reading in infants: Attention to speech presented in- and out-of-synchrony," *Cognitive Psychology*, vol. 11, no. 4, pp. 478–484, 1979. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0010028579900215

[8] L. Zhang and Y. Du, "Lip movements enhance speech representations and effective connectivity in auditory dorsal stream," *NeuroImage*, vol. 257, p. 119311, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S105381192200430X

[9] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[10] M. Masapollo, L. Polka, L. Ménard, L. Franklin, M. Tiede, and J. Morgan, "Asymmetries in unimodal visual vowel perception: The roles of oral-facial kinematics, orientation, and configuration," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 44, no. 7, pp. 1103–1118, 2018.

[11] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, and C. L. Zitnick, "Vqa: Visual question answering," *International Conference on Machine Learning*, pp. 2425–2434, 2015.

[12] J. S. Nam, E. Kim, and B. Kim, "Multimodal residual learning for visual qa," *International Conference on Computer Vision*, pp. 3174–3183, 2017.

[13] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," *International Conference on Machine Learning*, pp. 2048–2057, 2015.

[14] W. Chan, N. Xu, Q. V. Le, and T. N. Sainath, "Watch, listen, attend and spell: A deep multi-modal approach to speech recognition," *International Conference on Acoustics, Speech and Signal Processing*, pp. 4960–4964, 2016.

[15] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Workshop on Multi-view Lip-reading, Asian Conference on Compuer Vision (ACCV)*, 2016.

[16] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," *ArXiv*, vol. abs/2002.05709, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:211096730

[17] R. Arandjelovic and A. Zisserman, "Objects that sound," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 435–451.

[18] S.-W. Chung, J. S. Chung, and H.-G. Kang, "Perfect match: Improved cross-modal embeddings for audio-visual synchronisation," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3965–3969.

[19] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 95, 2017.

[20] R. Kumar, J. Sotelo, K. Kumar, A. D. Brébisson, and Y. Bengio, "Obamanet: Photo-realistic lip-sync from text," *ArXiv*, vol. abs/1801.01442, 2018.

[21] A. Jha, V. Voleti, V. Namboodiri, and C. V. Jawahar, "Cross-language speech dependent lip-synchronization," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7140–7144.

[22] O. Fried, A. Tewari, M. Zollhöfer, A. Finkelstein, E. Shechtman, D. B. Goldman, K. Genova, Z. Jin, C. Theobalt, and M. Agrawala, "Text-based editing of talking-head video," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–14, 2019.

[23] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner, "Neural voice puppetry: Audio-driven facial reenactment," *arXiv preprint arXiv:1912.05566*, 2019.

[24] A. Lahiri, V. Kwatra, C. Früh, J. Lewis, and C. Bregler, "Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization," *CoRR*, vol. abs/2106.04185, 2021. [Online]. Available: https://arxiv.org/abs/2106.04185

[25] J. S. Chung, A. Jamaludin, and A. Zisserman, "You said that?" in *British Machine Vision Conference*, 2017.

[26] A. Jamaludin, J. S. Chung, and A. Zisserman, "You said that?: Synthesising talking faces from audio," *International Journal of Computer Vision*, vol. 127, no. 11-12, pp. 1767–1779, 2019.

[27] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 484–492. [Online]. Available: https://doi.org/10.1145/3394171.3413532

[28] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 3444–3453.

[29] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[30] J. S. Chung and A. ZISSERMAN, "Lip reading in the wild," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 87–103.

[31] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[32] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230819302712

[33] A. Gupta, R. Mukhopadhyay, S. Balachandra, F. F. Khan, V. P. Namboodiri, and C. V. Jawahar, "Towards generating ultra-high resolution talking-face videos with lip synchronization," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 5198–5207.

[34] K. Cheng, X. Cun, Y. Zhang, M. Xia, F. Yin, M. Zhu, X. Wang, J. Wang, and N. Wang, "Videoretalking: Audio-based lip synchronization for talking head video editing in the wild," 2022.

[35] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, "Makeittalk: Speaker-aware talking-head animation," *ACM Transactions on Graphics*, vol. 39, no. 6, 2020.

[36] S. Wang, L. Li, Y. Ding, C. Fan, and X. Yu, "Audio2head: Audio-driven one-shot talking-head generation with natural head motion," *arXiv preprint arXiv:2107.09293*, 2021.

[37] Y. Ren, G. Li, Y. Chen, T. H. Li, and S. Liu, "Pirenderer: Controllable portrait image generation via semantic neural rendering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 759–13 768.

[38] 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*, 2021, pp. 3661–3670.

[39] R. Yi, Z. Ye, J. Zhang, H. Bao, and Y.-J. Liu, "Audio-driven talking face video generation with learning-based personalized head pose," *arXiv: Computer Vision and Pattern Recognition*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:212414741

[40] Y. Ma, S. Wang, Z. Hu, C. Fan, T. Lv, Y. Ding, Z. Deng, and X. Yu, "Styletalk: One-shot talking head generation with controllable speaking styles," in *AAAI Conference on Artificial Intelligence*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:255393710

[41] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. Liu, "Pose-controllable talking face generation by implicitly modularized audio-visual representation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[42] M. Agarwal, R. Mukhopadhyay, V. Namboodiri, and C. Jawahar, "Audio-visual face reenactment," *arXiv preprint arXiv:2210.02755*, 2022.

[43] A. Ephrat and S. Peleg, "Vid2speech: speech reconstruction from silent video," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[44] A. Ephrat, T. Halperin, and S. Peleg, "Improved speech reconstruction from silent video," *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 455–462, 2017.

[45] Y. Kumar, R. Jain, K. M. Salik, R. R. Shah, Y. Yin, and R. Zimmermann, "Lipper: Synthesizing thy speech using multi-view lipreading," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 2588–2595.

[46] H. Akbari, H. Arora, L. Cao, and N. Mesgarani, "Lip2audspec: Speech reconstruction from silent lip movements video," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2516–2520, 2017.

[47] K. Vougioukas, P. Ma, S. Petridis, and M. Pantic, "Video-driven speech reconstruction using generative adversarial networks," *arXiv preprint arXiv:1906.06301*, 2019.

[48] R. Yadav, A. Sardana, V. P. Namboodiri, and R. M. Hegde, "Speech prediction in silent videos using variational autoencoders," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7048–7052, 2020.

[49] M. Varshney, R. Yadav, V. P. Namboodiri, and R. M. Hegde, "Learning speaker-specific lip-to-speech generation," *arXiv preprint arXiv:2206.02050*, 2022.

[50] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "Learning individual speaking styles for accurate lip to speech synthesis," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13 793–13 802.

[51] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:206742911

[52] J.-H. Kim, J. Kim, and J. S. Chung, "Let there be sound: Reconstructing high quality speech from silent videos," *ArXiv*, vol. abs/2308.15256, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:261276980

[53] S. B. Hegde, K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "Lip-to-speech synthesis for arbitrary speakers in the wild," in *Proceedings of the 30th ACM International Conference on Multimedia*, ser. MM '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 6250–6258. [Online]. Available: https://doi.org/10.1145/3503161.3548081

[54] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*. PMLR, 2016, pp. 173–182.

[55] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 4485–4495.

[56] M. Kim, J. Hong, and Y. M. Ro, "Lip to speech synthesis with visual context attentional gan," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[57] R. Mira, A. Haliassos, S. Petridis, B. W. Schuller, and M. Pantic, "Svts: Scalable video-to-speech synthesis," *INTERSPEECH*, 2022.

[58] S. Hegde, R. Mukhopadhyay, C. Jawahar, and V. Namboodiri, "Towards accurate lip-to-speech synthesis in-the-wild," in *Proceedings of the 31st ACM International Conference on Multimedia*, ser. MM '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 5523–5531. [Online]. Available: https://doi.org/10.1145/3581783.3611787

[59] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[60] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[61] N. Harte and E. Gillen, "Tcd-timit: An audio-visual corpus of continuous speech," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, 2015.

[62] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy, "Mead: A large-scale audio-visual dataset for emotional talking-face generation," in *Computer Vision – European Conference on Computer Vision (ECCV) 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 700–717.

[63] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. Graph.*, vol. 37, no. 4, jul 2018. [Online]. Available: https://doi.org/10.1145/3197517.3201357

[64] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S3fd: Single shot scale-invariant face detector," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 192–201.

[65] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee *et al.*, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.

[66] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5203–5212.

[67] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1021–1030.

[68] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 2155–215 509.

[69] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.

[70] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[71] K. R. Prajwal, T. Afouras, and A. Zisserman, "Sub-word level lip reading with visual attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 5162–5172.

[72] C. Chen, J. W. on Pattern Recognition, M. Artificial Intelligence (1976, Hyannis, W. Recognition, and A. Intelligence, *Pattern Recognition and Artificial Intelligence: Proceedings of the Joint Workshop on Pattern Recognition and Artificial Intelligence, Held at Hyannis, Massachusetts, June 1-3, 1976*, ser. Academic Press Rapid Manuscript Reproduction. Academic Press, 1976. [Online]. Available: https://books.google.co.in/books?id=wW9QAAAAMAAJ

[73] S. S. Stevens, J. E. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *Journal of the Acoustical Society of America*, vol. 8, pp. 185–190, 1937.

[74] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[75] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *INTERSPEECH*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:119188316

[76] A. Baevski, H. Zhou, A. rahman Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *ArXiv*, vol. abs/2006.11477, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:219966759

[77] N. Dawalatabad, M. Ravanelli, F. Grondin, J. Thienpondt, B. Desplanques, and H. Na, "Ecapa-tdnn embeddings for speaker diarization," in *INTERSPEECH*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233024933

[78] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[79] R. J. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. A. J. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," *ArXiv*, vol. abs/1803.09047, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:4425995

[80] Y. Wang, D. Stanton, Y. Zhang, R. J. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:4349820

[81] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-vc: non-parallel many-to-many voice conversion using star generative adversarial networks," *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 266–273, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:46949741

[82] Y. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "Lipnet: End-to-end sentence-level lipreading," *arXiv: Learning*, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:7421075

[83] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[84] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[85] J. S. Chung and A. Zisserman, "Lip reading in profile," in *British Machine Vision Conference*, 2017.

[86] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 8717–8727, 2022.

[87] T. Afouras, J. S. Chung, and A. Zisserman, "Deep lip reading: a comparison of models and an online application," in *INTERSPEECH*, 2018.

[88] B. Shi, W.-N. Hsu, K. Lakhotia, and A. rahman Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," *ArXiv*, vol. abs/2201.02184, 2022.

[89] P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic, "Auto-avsr: Audio-visual speech recognition with automatic labels," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[90] Y. A. D. Djilali, S. Narayan, E. L. Bihan, H. Boussaid, E. Almazrouei, and M. Debbah, "Do vsr models generalize beyond lrs3?" *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6621–6630, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265445748

[91] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[92] D. H. Klatt, "Review of text-to-speech conversion for english," *Journal of the Acoustical Society of America*, vol. 82, pp. 737–793, 1987.

[93] F. Charpentier and M. Stella, "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," in *ICASSP '86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 11, 1986, pp. 2015–2018.

[94] D. O'Shaughnessy, L. Barbeau, D. Bernardi, and D. Archambault, "Diphone speech synthesis," *Speech Communication*, vol. 7, no. 1, pp. 55–65, 1988. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0167639388900210

[95] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, pp. 453–467, 1990.

[96] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. A. J. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *INTERSPEECH*, 2017.

[97] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," in *Advances in neural information processing systems*, 2017, pp. 2962–2970.

[98] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," *arXiv preprint arXiv:1710.07654*, 2017.

[99] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[100] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2410–2419.

[101] K. Kumar, R. Kumar, T. de Boissière, L. Gestin, W. Z. Teoh, J. M. R. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Neural Information Processing Systems*, 2019.

[102] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *ArXiv*, vol. abs/2010.05646, 2020.

[103] S. gil Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S.-H. Yoon, "Bigvgan: A universal neural vocoder with large-scale training," *ArXiv*, vol. abs/2206.04658, 2022.

[104] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*. PMLR, 2017, pp. 214–223.

[105] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *CoRR*, vol. abs/1312.6114, 2013.

[106] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[107] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[108] T. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein, and W. Matusik, "Speech2face: Learning the face behind a voice," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2019, pp. 7531–7540. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR.2019.00772

[109] M. Kamachi, H. Hill, K. Lander, and E. Vatikiotis-Bateson, "'putting the face to the voice': Matching identity across modality," *Current Biology*, vol. 13, no. 19, pp. 1709–1714, 2003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960982203006638

[110] Y. Wen, M. A. Ismail, W. Liu, B. Raj, and R. Singh, "Disjoint mapping network for cross-modal matching of voices and faces," *ArXiv*, vol. abs/1807.04836, 2018.

[111] A. Duarte, F. Roldan, M. Tubau, J. Escur, S. Pascual, A. Salvador, E. Mohedano, K. McGuinness, J. Torres, and X. G. i Nieto, "Wav2pix: Speech-conditioned face generation using generative adversarial networks," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[112] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[113] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *NIPS*, 2017.

[114] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," *ArXiv*, vol. abs/2005.08100, 2020.

[115] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015.

[116] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," in *International Conference on Learning Representations*, 2018.

[117] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6199–6203, 2019.

[118] M. Kim, J. Hong, and Y. M. Ro, "Lip-to-speech synthesis in the wild with multi-task learning," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[119] Y. Yemini, A. Shamsian, L. Bracha, S. Gannot, and E. Fetaya, "Lipvoicer: Generating speech from silent videos guided by lip reading," *ArXiv*, vol. abs/2306.03258, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259088588

[120] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," *ArXiv*, vol. abs/1804.04121, 2018.

[121] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," *arXiv preprint arXiv:1804.03641*, 2018.

[122] R. Mukhopadhyay, S. B. Hegde, V. Namboodiri, and C. Jawahar, "Audio-visual speech super-resolution," *British Machine Vision Conference (BMVC)*, 2021.

[123] D. J. Lewkowicz and A. M. Hansen-Tift, "Infants deploy selective attention to the mouth of a talking face when learning speech," *Proceedings of the National Academy of Sciences*, vol. 109, no. 5, pp. 1431–1436, 2012.

[124] L. I. Iezzoni, B. L. O'Day, M. Killeen, and H. Harker, "Communicating about health care: observations from persons who are deaf or hard of hearing," *Annals of Internal Medicine*, vol. 140, no. 5, pp. 356–362, 2004.

[125] H. Zhou, Z. Liu, X. Xu, P. Luo, and X. Wang, "Vision-infused deep audio inpainting," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 283–292.

[126] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[127] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on Pattern analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[128] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[129] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML'15, 2015, pp. 448–456.

[130] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[131] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 4214 – 4217, 04 2010.

[132] J. Jensen and C. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 1–1, 11 2016.

[133] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq): A new method for speech quality assessment of telephone networks and codecs," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2, pp. 749–752 vol.2, 02 2001.

[134] P. KR, R. Mukhopadhyay, J. Philip, A. Jha, V. Namboodiri, and C. Jawahar, "Towards automatic face-to-face translation," in *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 2019, pp. 1428–1436.

[135] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[136] P. Ma, Y. Wang, J. Shen, S. Petridis, and M. Pantic, "Lip-reading with densely connected temporal convolutional networks," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 2857–2866.

[137] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[138] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, 2016.

[139] J.-H. Choi, J.-H. Kim, M. Cheon, and J.-S. Lee, "Deep learning-based image super-resolution considering quantitative and perceptual quality," *Neurocomputing*, vol. 398, pp. 347–359, 2020.

[140] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 105–114.

[141] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.

[142] T. Afouras, A. Owens, J. S. Chung, and A. Zisserman, "Self-supervised learning of audio-visual objects from video," in *European Conference on Computer Vision*, 2020.

[143] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

[144] U. Demir and G. Ünal, "Patch-based image inpainting with generative adversarial networks," *ArXiv*, vol. abs/1803.07422, 2018.

[145] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *International conference on machine learning*. PMLR, 2016, pp. 1558–1566.

[146] M. Rosca, B. Lakshminarayanan, D. Warde-Farley, and S. Mohamed, "Variational approaches for auto-encoding generative adversarial networks," *arXiv preprint arXiv:1706.04987*, 2017.

[147] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," *Proceedings of Machine Learning Research*, vol. 70, pp. 214–223, 06–11 Aug 2017.

[148] T. Tieleman and G. Hinton, "Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude," COURSERA: Neural Networks for Machine Learning, 2012.

[149] M. Bińkowski, J. Donahue, S. Dieleman, A. Clark, E. Elsen, N. Casagrande, L. C. Cobo, and K. Simonyan, "High fidelity speech synthesis with adversarial networks," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=r1gfQgSFDr

[150] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ArXiv*, vol. abs/2010.11929, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:225039882

[151] A. Ponnusamy, "cvlib - high level computer vision library for python," https://github.com/arunponnusamy/cvlib, 2018.

[152] U. Jain, Z. Zhang, and A. Schwing, "Creativity: Generating diverse questions using variational autoencoders," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5415–5424.

[153] P. K. R, L. Momeni, T. Afouras, and A. Zisserman, "Visual keyword spotting with attention," in *BMVC*, 2021.

[154] L. Momeni, H. Bull, K. R. Prajwal, S. Albanie, G. Varol, and A. Zisserman, "Automatic dense annotation of large-vocabulary sign language videos," in *Computer Vision – European Conference on Computer Vision (ECCV) 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 671–690.

[155] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *INTERSPEECH*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:12418404

[156] S. Yunhui and R. Qiuqi, "Continuous wavelet transforms," in *Proceedings 7th International Conference on Signal Processing, 2004. Proceedings. ICSP '04. 2004.*, vol. 1, 2004, pp. 207–210 vol.1.

[157] M. Freitag and Y. Al-Onaizan, "Beam search strategies for neural machine translation," in *NMT@ACL*, 2017.

[158] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964.

[159] B. Sen, A. Agarwal, R. Mukhopadhyay, V. Namboodiri, and C. Jawahar, "Personalized one-shot lipreading for an als patient," *arXiv preprint arXiv:2111.01740*, 2021.

[160] M. Bain, J. Huh, T. Han, and A. Zisserman, "Whisperx: Time-accurate speech transcription of long-form audio," *arXiv preprint arXiv:2303.00747*, 2023.

[161] S. Serengil and A. Ozpinar, "A benchmark of facial recognition pipelines and co-usability performances of modules," *Journal of Information Technologies*, vol. 17, no. 2, pp. 95–107, 2024. [Online]. Available: https://dergipark.org.tr/en/pub/gazibtd/issue/84331/1399077

[162] P. Ekstrand, "Bandwidth extension of audio signals by spectral band replication," in *Proceedings of the 1st IEEE Benelux Workshop on Model Based Processing and Coding of Audio (MPCA '02*, 2002.

[163] E. Larsen and R. M. Aarts, *High-Frequency Bandwidth Extension for Audio*. John Wiley & Sons, Ltd, 2004, ch. 5, pp. 145–170. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/0470858710.ch5

[164] Y. Cheng, D. O'Shaughnessy, and P. Mermelstein, "Statistical recovery of wideband speech from narrowband speech," *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 544–548, 1994.

[165] K. Li, Z. Huang, Y. Xu, and C.-H. Lee, "Dnn-based speech bandwidth expansion and its application to adding high-frequency missing features for automatic speech recognition of narrowband speech," in *INTERSPEECH*, 2015.

[166] V. Kuleshov, S. Enam, and S. Ermon, "Audio super resolution using neural networks," *International Conference on Learning Representations Workshops (ICLR Workshops)*, vol. abs/1708.00853, 2017.

[167] S. Birnbaum, V. Kuleshov, Z. Enam, P. W. W. Koh, and S. Ermon, "Temporal film: Capturing long-range sequence dependencies with feature-wise modulations." in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

[168] J. Lee and S. Han, "Nu-wave: A diffusion probabilistic model for neural audio upsampling," *INTERSPEECH*, pp. 1634–1638, 2021.

[169] S. B. Hegde, K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "Visual speech enhancement without a real visual stream," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2021, pp. 1926–1935.

[170] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 380–391, 1976.

[171] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," *ICASSP '96 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 629–632, 1996.

[172] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2013.

[173] Y. Tu, I. Tashev, S. Zarar, and C. Lee, "A hybrid approach to combining conventional and deep learning techniques for single-channel speech enhancement and recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2531–2535.

[174] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech," in *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, 2016, pp. 146–152.

[175] F. G. Germain, Q. Chen, and V. Koltun, "Speech denoising with deep feature losses," in *INTERSPEECH*, 2019.

[176] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," in *Proc. INTERSPEECH 2017*, 2017, pp. 3642–3646.

[177] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5024–5028, 2018.

[178] S. Abdulatif, K. Armanious, K. Guirguis, J. Thaiparambil Sajeev, and B. Yang, "Aegan: Time-frequency speech denoising via generative adversarial networks," *arXiv preprint arXiv:1910.12620*, 10 2019.

[179] D. Dean, S. Sridharan, R. Vogt, and M. Mason, "The qut-noise-timit corpus for the evaluation of voice activity detection algorithms," in *INTERSPEECH*, 2010.

[180] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, 2013, pp. 1–4.

[181] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 11 1992.

[182] C. Valentini-Botinhao, "Noisy speech database for training speech enhancement algorithms and tts models, 2016," University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR), 2017. [Online]. Available: https://doi.org/10.7488/ds/2117

[183] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd Annual ACM Conference on Multimedia*. ACM Press, 2015, pp. 1015–1018.

[184] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 721–725.

[185] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf," *Communications of the ACM*, vol. 65, pp. 99 – 106, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:213175590

[186] V. Blanz and T. Vetter, *A Morphable Model For The Synthesis Of 3D Faces*, 1st ed. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3596711.3596730

[187] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *ArXiv*, vol. abs/1809.00496, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:52155419

[188] A. Triantafyllos, J. S. Chung, and A. Zisserman, "Now you are speaking my language: Visual language identification," in *Interspeech*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221317408

[189] K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser *et al.*, "Rethinking attention with performers," *arXiv preprint arXiv:2009.14794*, 2020.

[190] Coqui.ai, "Tts: A toolbox for text-to-speech synthesis," 2023. [Online]. Available: https://github.com/coqui-ai/TTS

[191] ElevenLabs, "Elevenlabs: Voice ai platform," 2023. [Online]. Available: https://elevenlabs.io

[192] R. AI, "Resemble ai: Ai voice generator," 2023, accessed: 2024-07-25. [Online]. Available: https://www.resemble.ai

[193] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. A. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *International Conference on Machine Learning*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:234483016