Beyond Text: Expanding Speech Synthesis with Lip-to-Speech and Multi-Modal Fusion

Thesis submitted in partial fulfilment
of the requirements for the degree of

*(Master of Science in **Electronics and Communication Engineering** by Research)*

by

Neha Sahipjohn
2021702012
neha.s@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
June 2024

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled ***"Beyond Text: Expanding Speech Synthesis with Lip-to-Speech and Multi-Modal Fusion"*** by Neha Sahipjohn, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Advisor: Dr. Vineet Gandhi

To the Past, Present and Future

Acknowledgments

I am deeply grateful to my advisor, Dr. Vineet Gandhi, for his invaluable guidance and support and for the opportunity to work under his guidance. His mentorship was instrumental in shaping my research journey. He helped me identify impactful problems to pursue and fostered in me a mindset of solving real-world issues through research. I am grateful to him for giving me the freedom to explore different ideas and for helping me in the overall completion of my degree. I deeply appreciate his patience and belief in me. He instilled in me not only strong research skills but also the confidence to pursue them, and for that, I am eternally thankful.

I am thankful to my professors, seniors and peers for shaping my AI-ML journey. Dr. Anoop Namboodiri's insightful classes on ML laid the foundation of my knowledge in the field. The rigorous coursework and assignments at IIIT significantly honed my coding skills and confidence. Additionally, the teachings of Dr. Anil Kumar Vuppala, Dr. Chiranjeevi Yarra, Dr. Ravi Kiran, Dr. Pawan Kumar and Dr. Manish Shrivastava helped me develop a better and deeper understanding of the field.

I am deeply indebted to Saiteja Kosgi and Neilkumar Shah for their invaluable guidance, support and enlightening discussions, which have contributed significantly to the development of this thesis. Engaging in discussions with Vishal T, Ayan, and Ritam Basu has been both enriching and enjoyable, providing me with diverse perspectives and insights into problem-solving. I am grateful to my lab mates Astitva, Chandradeep and Shantika for their guidance and assistance whenever I reached out.

Special thanks to my friends Arpit Sahni, Ritam Basu, Habeeba Khan, Laksh N, Pranav Gupta, Deepti and Rishal for making my time at IIIT Hyderabad fun and memorable. I also fondly remember the celebrations and camaraderie with Anjani, Shiva, Ashish, Chetan, and other seniors in MS ECE, which added a much-needed human touch to the demanding academic journey. I extend my heartfelt appreciation to everyone for their companionship.

Lastly, I am deeply grateful to my Mom, Nana, and family for their unwavering belief, trust, and support throughout my professional and academic journey.

Abstract

Speech constitutes a fundamental aspect of human communication. Therefore, the ability of computers to synthesize speech is paramount for achieving more natural human-computer interactions and increased accessibility, particularly for individuals with reading limitations. Recent advancements in AI and machine learning technologies, alongside generative AI techniques, have significantly improved speech synthesis quality. Text input serves as a common modality for speech synthesis, and Text-to-Speech (TTS) systems have achieved notable milestones in terms of intelligibility and naturalness. In this thesis, we propose a system to synthesize speech directly from lip movements and explore the idea of a unified speech synthesis model that can synthesize speech from different modalities, like text-only, video-only or combined text and video inputs. This facilitates applications in dubbing and accessibility initiatives aimed at providing voice to individuals who are unable to vocalize. This innovation promises streamlined communication in noisy environments as well. We propose a novel system for lip-to-speech synthesis that achieves state-of-the-art performance by leveraging advancements in self-supervised learning and sequence-to-sequence networks. This enables the generation of highly intelligible and natural-sounding speech even with limited data.

Existing lip-to-speech systems primarily focus on directly synthesizing speech or mel-spectrograms from lip movements. This often leads to compromised intelligibility and naturalness due to the entanglement of speech content with ambient information and speaker characteristics. We propose a modularized approach that uses representations that disentangle speech content from speaker characteristics, leading to superior performance. Our work sheds light on the information-rich nature of embedding spaces compared to tokenized representations. The system maps lip movement representations to disentangled speech representations, which are then fed into a vocoder for speech generation. Recognizing the potential applications in dubbing and the importance of synthesizing accurate speech, we explore a multimodal input setting by incorporating text alongside lip movements.

Through extensive experimentation and evaluation across various datasets and metrics, we demonstrate the superior performance achieved by our proposed method. Our approach demonstrates high correctness and intelligibility, paving the way for practical deployment in real-world scenarios. Our work contributes significantly to advancing the field of lip-to-speech synthesis, offering a robust and versatile solution for generating natural-sounding speech from silent videos with broader implications for accessibility, human-computer interaction, and communication technology.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Speech synthesis is a technology that enables computers and other devices to generate human-like speech. It has revolutionized the way humans interact with computers and devices. Traditionally, text serves as the primary input, but the future holds immense potential for incorporating additional modalities. Humans can infer speech content from various modalities such as lip movements, videos, whispers, hand gestures, and body language. Expanding the scope of speech synthesis to incorporate these modalities opens up a plethora of applications that enhance human-computer interaction experiences, fostering a more intuitive and natural experience and improve accessibility for speech-disabled individuals. Imagine a system that can synthesize speech by analyzing lip movements in videos, capturing the essence of a whispered conversation, or even interpreting hand gestures and body language. In this thesis, we explore how we can use modalities like lip movements, text, or both to synthesise speech. One such scenario, where a person can look into the camera, which will capture their lip movements and convert them to speech, is depicted in Fig. 1.1. This can be used in a scenario where the person wants to communicate in a crowded loud place or if the person is someone with speaking disability.

Speech synthesis has already made significant strides in improving accessibility. Tools that convert text to speech have empowered people with speech disabilities to communicate more effectively. A famous example is the late physicist Stephen Hawking, who relied on a speech synthesizer to deliver his groundbreaking ideas despite his condition. Beyond accessibility, speech synthesis plays a crucial role in various applications. It has applications in the realm of entertainment, dubbing, video conferencing, surveillance, etc. As the technology continues to evolve and incorporate additional modalities, we can expect even more transformative applications to emerge, shaping the future of communication and interaction.

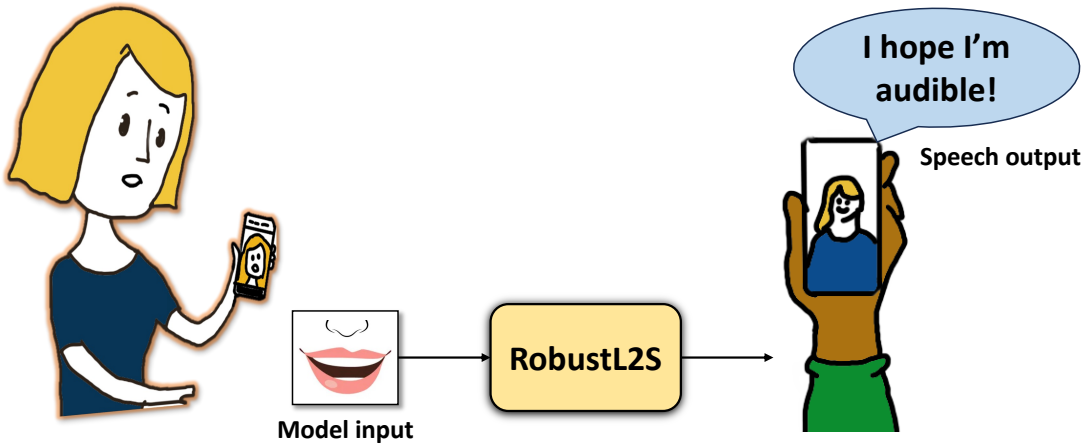

Figure 1.1: Speech synthesis based on lip movements captured by a camera.

1.1 Speech synthesis

Text-to-speech (TTS) synthesis is the process of generating speech from text. It has evolved significantly over the years, transitioning from mechanical methods to electrical and mathematical approaches and finally to neural network methods. Early attempts at TTS relied on mechanical methods, such as phonographs equipped with bellows and resonators or reeds to mimic the human vocal tract [1]. Later, electrical means were explored. Homer Dudley [2] developed keyboard-operated synthesizers called Voder. While ingenious for their time, these efforts produced artificial-sounding speech with limited expressiveness. Methods involving breaking down speech into its constituent components, such as phonemes, and using rules and algorithms to generate speech based on these components were developed. Concatenative synthesis gained popularity. While these approaches represented significant progress, they still struggled to produce natural-sounding speech, often lacking natural intonation and prosody.

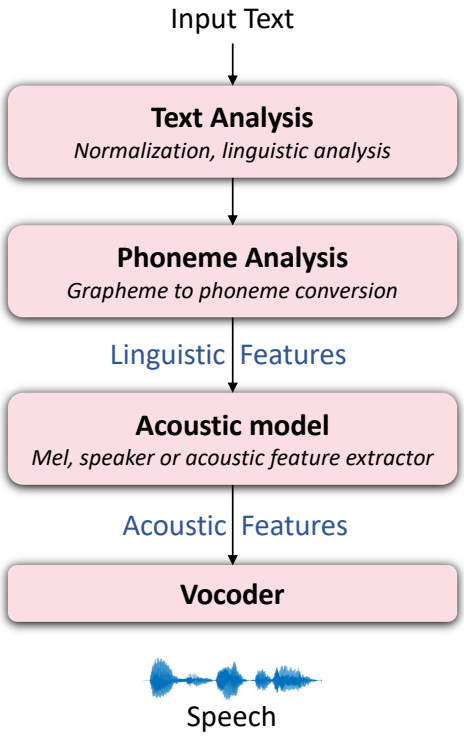

Figure 1.2: Block diagram depicting main components in a TTS system.

With the rise of machine learning, statistical methods emerged as a dominant force. Hidden Markov Models (HMMs) [3], [4] were employed to model the sequential nature of speech, leading to improvements in naturalness. However, these methods still struggled to capture the complex nuances of human speech, resulting in a robotic quality. The advent of deep learning, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), ushered in a new era for TTS. These powerful architectures excel at capturing complex relationships within data, making them ideal for learning the intricacies of human speech production. Pioneering work by van den Oord et al.[5] - WaveNet, a deep convolutional network, demonstrated the potential of neural networks for generating high-fidelity speech. WaveNet directly models the raw waveform of speech, resulting in unprecedented naturalness and expressiveness. Since then, various neural network architectures have been explored, each with its own strengths and weaknesses. A modern TTS system typically comprises three main components as depicted in Fig. 1.2:

- **Linguistic Model**: This component analyzes the text, understanding its grammatical structure, meaning, and intended emphasis. It can leverage techniques like natural language processing (NLP) to extract relevant information for speech generation.

- **Acoustic Model**: This component translates the linguistic representation into an acoustic representation, such as a spectrogram or mel-cepstral coefficients. Such accoustic representations are discussed in further sections.

- **Vocoder**: This component converts the acoustic representation back into an actual audio waveform. Recent advancements have seen the rise of Generative Adversarial Networks (GANs) [6] and Diffusion-based vocoders [7]. They demonstrate remarkable performance in synthesizing clear, human-sounding speech.

The future of speech synthesis extends beyond just text input. Researchers are actively exploring the possibility of synthesizing speech based on visual cues, like lip movements extracted from videos. For this, researchers have explored building linguistic models specifically tailored to these modalities. This entails learning relevant representations of lip movements using neural networks. This would be particularly valuable for applications like dubbing videos or creating speech interfaces for robots that rely on visual communication.

1.2 Speech Representation - Acoustic Modelling

Humans synthesize speech through a complex process involving the coordination of various speech organs, including the lungs, vocal cords, mouth, and tongue. Air from the lungs passes through the vocal cords, where vibrations create sound. The mouth and tongue shape this sound into specific phonemes, which are the building blocks of speech. Finally, the brain coordinates these movements and selects appropriate words and grammar to convey meaning. Speech, in the physical world, is a product of rapid air pressure fluctuations, resulting in audible sound waves. For computer processing and analysis, speech requires digital representation. This involves discretizing the continuous signal into a sequence of amplitude values sampled at a specific frequency (e.g. 8 kHz, 16 kHz). While convenient for storage and manipulation, this one-dimensional representation presents challenges for efficient processing due to the data volume and inherent redundancy within the speech signal itself. To address these limitations, researchers have explored various techniques for representing speech signals in a more compact and informative manner. These methods aim to capture the essential characteristics of speech while minimizing redundancy and data volume, ultimately facilitating more efficient processing and analysis. Some of the most prevalent traditional representations include:

- Linear predictive coding (LPC): This technique analyzes the speech signal as a sequence of vocal tract filters, extracting information about the formants (resonant frequencies) that shape the sound. LPC is particularly useful for representing voiced sounds, where formants play a key role in distinguishing different vowels.

- Mel-frequency cepstral coefficients (MFCCs): This method mimics the human auditory system by focusing on frequencies relevant to human speech perception. MFCCs capture the spectral en-

velope of the speech signal, which carries information about the vocal tract shape and contributes significantly to vowel sounds.

- Mel spectrogram: It captures how sound energy distributes across perceived frequencies. It is a representation of sound that visualizes the frequency content of an audio signal over time, with frequencies on the y-axis and time on the x-axis. Mel spectrogram is widely used in speech and audio processing tasks, such as speech recognition and music analysis, as it better approximates how humans perceive sound. This includes all the information in the audio signal, such as speech content, speaker voice information, background noise if present, etc.

These traditional representations laid the groundwork for significant advancements in speech processing. Their efficiency and interpretability continue to make them valuable tools in various applications, particularly in resource-constrained scenarios or when dealing with smaller datasets.

Deep learning techniques have revolutionized speech representation by automatically learning informative features directly from the raw audio data. This has led to the development of a diverse range of speech representations, each tailored for specific tasks and offering distinct advantages. i-vectors and d-vectors were among the early attempts to learn speaker embeddings for speaker verification tasks. These representations, derived from factor analysis techniques like Joint Factor Analysis and Deep Neural Networks (DNNs), have proven highly effective for speaker verification. Self-supervised learning techniques have emerged as a powerful approach for learning speech representations without requiring labelled data. HuBERT [8] is an example of a self-supervised model designed to capture rich content information from speech signals. By training on large amounts of unlabeled speech data, HuBERT learns to extract features that are informative for automatic speech recognition (ASR) downstream task while also disentangling content from other factors like speaker identity or background noise. This ”disentanglement” between speaker identity and the actual message spoken allows for superior performance in tasks like speech recognition and natural language processing when combined with downstream fine-tuning. Disentanglement refers to the ability of a representation to separate different factors of variation in the data. Better disentanglement between content and other information in speech audio enhances the robustness and generalization of speech processing systems. The development of these diverse speech representations offers several key benefits:

- Improved Performance: More informative representations improve performance in various speech processing tasks, including speaker recognition, speech synthesis, and automatic speech recognition (ASR).

- Reduced Computational Cost: Compact representations with lower dimensionality require less processing power, enabling more efficient algorithms and real-time applications.

- Enhanced Generalizability: Representations that capture the essence of speech content are less susceptible to variations in speaker characteristics or background noise, leading to more robust models.

The exploration of new speech representations continues to be an active area of research. As deep learning techniques evolve and incorporate new modalities beyond audio, we can expect even more sophisticated and versatile representations to emerge, further unlocking the potential of speech processing for a wide range of applications. Further research is needed to explore advanced techniques for disentanglement and representation learning in speech processing. Challenges include handling variability across languages and dialects, as well as developing models that are robust to environmental factors and speaker variations.

1.3 Video Representation

Videos have become a ubiquitous part of our lives. However, their inherent complexity, consisting of dynamic sequences and rich visual information, poses challenges for computers to process and understand them efficiently. This is where video representations come to the forefront, acting as a crucial bridge between the raw video data and the needs of various applications.

At the most fundamental level, a video can be simply represented as a sequence of individual frames, static images capturing a single moment in time. Videos can be represented as a sequence of images, which is a 2-dimensional matrix of intensity values. However, this basic approach fails to capture the essence of the moving world. More sophisticated techniques delve deeper, extracting features like motion information, object recognition, and even the semantic meaning conveyed within the video.

- Motion Information: Understanding movement is crucial for comprehending video sequences. This information is valuable for tasks like action recognition and video stabilization.
- Object Recognition: Identifying objects within the video frames is another critical step. This enables applications like video surveillance, where objects of interest can be automatically detected and tracked, or video editing, where specific objects can be easily selected and manipulated.
- Semantic Understanding: Extracting the higher-level meaning conveyed by the video involves techniques that go beyond individual frames and objects. Natural language processing (NLP) combined with deep learning approaches can analyze the video content in conjunction with any accompanying audio or text (captions, narration) to understand the overall message or story being told. This paves the way for tasks like video summarization, where key points and events are automatically extracted, or video question answering, where the system can answer questions about the video content.

Some of the methods used to learn these video representations include : Optical Flow: Estimates the apparent motion of pixels between frames, capturing the direction and speed of objects in the scene. 3D Convolutional Neural Networks (3D-CNNs): Process video data as a 3D volume, which considers each frame's spatial information (width and height) along with the temporal dimension (sequence of frames). By learning patterns across these 3 dimensions, 3D-CNNs can effectively capture both spatial

features (objects, shapes) and temporal information (motion) in videos. Residual Networks (ResNets): Introduced skip connections, facilitating better gradient flow and enabling training deeper networks for complex video tasks like action recognition or video segmentation. Self-supervised Learning (SSL): The model learns video representations by itself from unlabeled videos. Contrastive Learning: This approach focuses on pulling closer video clips from the same category while pushing apart those from different categories, allowing the model to learn discriminative representations. These are just a few examples, and the field of video representation is constantly evolving. By combining these techniques and exploring new avenues, researchers are continuously improving the ability of computers to understand the rich information within videos.

To synthesize speech from lip movements, we need to be able to identify the lip region in the video. This can be done by simple methods like keypoint detection. Then, we need to learn the movements and how these movements relate to sounds that are produced by mouth. Methods like 3D-CNNs, ResNets, SSL and transformers are explored for this.

1.4 Self-Supervised Learning

In machine learning, labelled data – meticulously categorized and annotated by human experts – plays an essential role. However, the process of acquiring such data can be a significant impediment to progress. The laborious and time-consuming nature of human labelling often creates a bottleneck, hindering the development and application of machine learning models. This is where self-supervised learning (SSL) emerges as a game-changer. Unlike supervised learning, which relies solely on labelled data for training, SSL cleverly leverages the vast amount of unlabeled data. By ingeniously crafting tasks and extracting implicit labels from the data itself, SSL empowers models to learn meaningful representations and perform well on downstream tasks that traditionally require labelled data.

Imagine a child learning to identify objects in the world. They don't need someone explicitly pointing and labelling every object they encounter. Instead, they learn by observing the world around them, making connections, and drawing inferences. SSL operates in a similar fashion. By setting up tasks like predicting the next word in a sentence, identifying the colour of an object in a distorted image, or reconstructing a masked portion of a video, the model essentially "plays" with the data, uncovering hidden patterns and relationships within it. Though not explicitly labelled by humans, these self-generated supervisory signals guide the model's learning process, enabling it to develop a strong understanding of the underlying data structure. The benefits of SSL are manifold. Firstly, it tackles the data scarcity problem, allowing us to utilize the abundance of unlabeled data that often goes untapped. Secondly, it offers increased learning efficiency, requiring significantly less labelled data compared to traditional supervised learning. Finally, by fostering the ability to learn from diverse and unconstrained data, SSL empowers models to develop a more generalizable understanding that can adapt to new situations and unseen data. SSL holds immense potential for revolutionizing various fields, including computer vision, natural language processing, and recommendation systems.

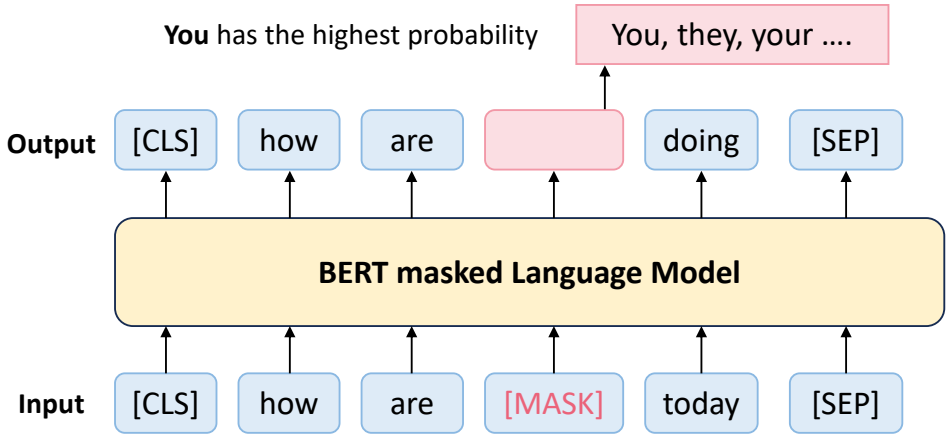

Figure 1.3: Pretraining of masked language model.

In Natural Language Processing (NLP), SSL leverages unlabeled text data to train models by crafting pretext tasks that guide the learning process. Here are some common examples of SSL techniques in NLP:

- Masked Language Modeling (MLM): This technique as depicted in Fig. 1.3 involves randomly masking words in a sentence and training the model to predict the masked words based on the surrounding context. This helps the model learn the relationships between words and their semantic meaning.
- Next Sentence Prediction (NSP): Given two sentences, the model is tasked with predicting whether the second sentence is a logical continuation of the first. This helps the model understand sentence coherence and relationships between ideas.
- Contrastive Learning: This is a specific SSL technique that focuses on pulling similar data points closer together in a latent representation space, while pushing dissimilar data points further apart. This allows the model to identify important features and relationships within the data, even without explicit labels.

SSL has emerged as a powerful tool in NLP, paving the way for advancements in various applications like machine translation, chatbots, and information retrieval. The success of self-supervised learning (SSL) in the realm of natural language processing (NLP) has spurred its exploration in other domains, including speech and vision. One promising approach in speech processing involves adapting the masked language modelling technique used in NLP. Here, the core concept remains the same: the model encounters masked portions of the data and attempts to predict the missing elements. However, in speech applications, the masked data points would be specific frames within the speech signal rather than words in a sentence. Instead of predicting the raw speech amplitude values of the masked frames, the model would operate on mel-spectrograms or iteratively learned speech representations. This allows the model to focus on the underlying structure and relationships within the speech data. As research in

SSL continues to flourish, we can expect even more innovative applications to emerge across various modalities, paving the way for a new era of robust and versatile machine-learning models.

1.5 Sequence Modelling

Sequence modelling is a technique designed to handle sequential data, which could be either at the input or at the output. Sequence-to-Sequence (Seq2Seq) models take in one sequence of input (words, letters, video frames) and output another sequence of output (next words, text in different language, caption). Seq2Seq models have revolutionized the field of machine learning by enabling tasks involving the conversion of one sequence of data to another. Seq2seq models have become a powerful tool for various time domain tasks. From machine translation and text summarization to chatbots, speech recognition and music synthesis, these models have become the backbone of numerous applications that interact with and understand human language. Their ability to handle sequential data and capture temporal dependencies makes them valuable for time-domain applications.

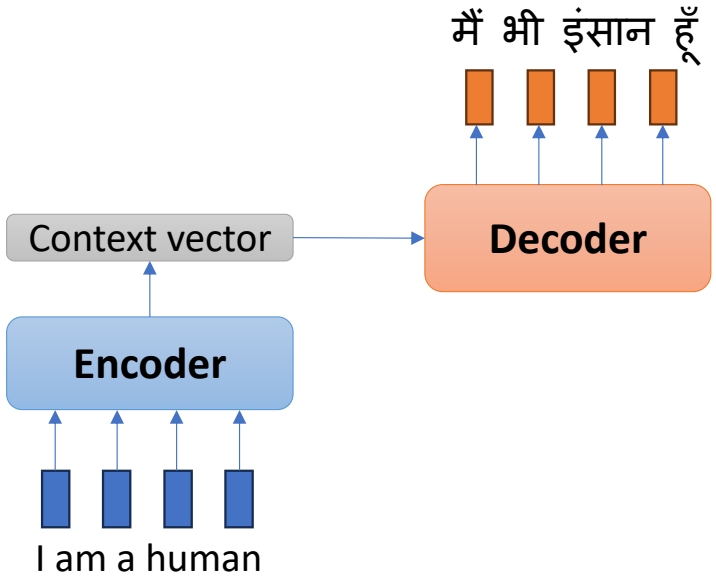

Figure 1.4: Block diagram of Encoder-Decoder architecture employed for machine translation task.

Some of the earlier approaches for seq2seq modelling include Hidden Markov Models (HMMs), Finite State Machines (FSMs), and rule-based systems.

- Hidden Markov Models (HMMs): HMMs were an early attempt at capturing sequential data. They model the probability of transitioning between states, allowing for the generation of sequences.

- Statistical Machine Translation (SMT): This rule-based approach relied on translation dictionaries and linguistic rules to translate sentences.

These approaches often struggled with capturing complex patterns and long-range dependencies in sequential data. However, with the introduction of recurrent neural networks (RNNs), particularly with the Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures, Seq2Seq modelling gained significant traction. These neural network-based approaches revolutionized sequence-to-sequence tasks by allowing for more effective learning of sequential patterns and handling of variable-length input and output sequences.

- Recurrent Neural Networks (RNNs): RNNs have internal loops that allow them to process sequential data and maintain information about previous elements in the sequence. However, traditional RNNs suffer from the vanishing gradient problem, making learning long-term dependencies in long sequences difficult. To overcome the limitations of traditional RNNs, several variants were developed.

- Long Short-Term Memory (LSTM) networks: LSTMs address the vanishing gradient problem by introducing memory cells that can store information for longer durations. This allows them to effectively capture long-range dependencies in sequences, making them well-suited for Seq2Seq tasks.

- Gated Recurrent Units (GRUs): Similar to LSTMs, GRUs are another RNN variant designed to address the vanishing gradient problem. They utilize a simpler gating mechanism compared to LSTMs, making them computationally more efficient.

- Encoder-Decoder Architecture: This core architecture, as depicted in Fig. 1.4, became the foundation for modern Seq2Seq models. The encoder takes the input sequence and processes it into a context vector, capturing the essential information. The decoder then utilizes this context vector to generate the output sequence one step at a time, relying on an attention mechanism to focus on relevant parts of the input sequence during the decoding process.

- Attention Mechanisms: Transformer models have significantly improved the ability of models to focus on specific parts of the input sequence, leading to more accurate and nuanced outputs.

- Conditional Seq2Seq Models: These models incorporate additional information beyond the input sequence, such as labels or specific tasks, to guide the generation process, enabling applications like text summarization with specific keywords or sentiment analysis.

- Pre-trained Language Models (PLMs): Large, pre-trained models like BERT [9] and GPT-3 [10] are being leveraged as encoders in Seq2Seq models, further enhancing their ability to capture complex semantic relationships and generate human-quality text.

Seq2Seq models play a great role in shaping the future of human-computer interaction and natural language processing. As research advances, we can expect even more powerful and versatile Seq2Seq models to emerge, shaping the future of various applications that rely on understanding and manipulating sequential data.

1.6 Objectives and contributions

This thesis investigates the potential of multimodal speech synthesis, where different modalities beyond just text are incorporated as input to the model. The key objective and contributions of the thesis are as follows:

- Delve into speech synthesis based on lip movements in a video, by leveraging self-supervised learning representations.
- Can multiple modalities be used as input to one speech synthesis model? How can we bring all modalities to one common space?
- What could be the potential applications of this?

By addressing these questions, this thesis aims to contribute to the advancement of speech synthesis by exploring the potential of multimodal fusion for speech generation. Developing techniques to effectively incorporate diverse modalities into the synthesis process. The findings from this research will provide valuable insights into how multimodal information can be leveraged to generate speech, opening doors for new applications and improved user experiences in various domains.

Chapter 2 details the related works in the field of speech synthesis, lip reading and SSL in the speech and video domain, which have been leveraged in the multimodal speech synthesis framework. This chapter lays the foundation for our research by exploring the existing body of work relevant to multimodal speech synthesis. Chapter 3 explores a method to utilize powerful SSL representations for Lip2Speech synthesis task. Chapter 4 delves into how different modalities can be used for speech synthesis in one common model. Concluding Chapter 5 summarizes the thesis and proposes possible directions for future research.

Chapter 2

Related Works

This chapter explores existing research on Lip-to-Speech (Lip2Speech) synthesis and multimodal speech synthesis. Lip2Speech tackles the challenging task of generating realistic speech by analyzing silent videos of a person speaking. This technology bridges the gap between visual and auditory information, offering potential applications like adding speech to silent videos or aiding individuals with speech difficulties. Early research in Lip2Speech often relied on constrained datasets where a limited number of speakers uttered specific sentences. They were recorded in controlled environments, with frontal faces always visible in the video and had limited vocabulary, hampering the model's ability to generalize to real-world scenarios. Datasets collected from real-world YouTube videos form the unconstrained datasets. These videos capture people speaking in real-world scenarios and have various head poses and a large vocabulary, with a natural speaking style containing fillers and pauses. This chapter will delve into existing methods for both constrained and unconstrained settings. These methods are usually end-to-end or supervised ML models. Furthermore, we will explore recent advancements in representation learning and Self-Supervised Learning (SSL) that have been applied to Lip2Speech synthesis. These techniques aim to achieve faster and more efficient training on large, unconstrained datasets, paving the way for more robust and scalable Lip2Speech models. Later, recent advancements in LLM-based TTS and multimodal speech synthesis frameworks are discussed briefly.

2.1 Lip-to-Speech synthesis

- Constrained Lip-to-Speech synthesis

  Constrained lip-to-speech synthesis tackles speech generation from videos with limited vocabulary and minimal head movement [11, 12]. Pioneering works focused on building models to translate silent video directly into speech features. Ephrat *et al.*[13] introduced a CNN-based approach to predict Linear Predictive Coding (LPC) features from silent talking videos. This approach laid the groundwork for Lip2Speech synthesis. They later enhanced their model to a two-tower CNN-based encoder-decoder architecture [14], encoding raw frames and optical flows separately. The CNN generates multiple mel-scale spectrograms, which are converted by a post-processing

network to a longer-range linear-scale spectrogram. Phase reconstruction is done to transform the long-range spectrogram into a waveform. With the development of autoencoders, which excel at compressing and decompressing information, researchers explored combining them with lipreading networks to better capture the nuances of audio and visual data. [15] propose a combination of an autoencoder for extracting bottleneck features from audio spectrograms and a lipreading network comprising CNN, LSTM, and fully connected layers, for visual feature extraction. On the other hand, [16] utilizes a stochastic modelling approach employing a variational autoencoder. A variational autoencoder is a type of autoencoder that injects some randomness for better generalization and a more probabilistic approach to synthesis. [17] achieves zero-shot lip-to-speech synthesis using a variational autoencoder to disentangle speaker and content information and a face identity encoder for unseen speakers. Other methods [18, 19, 20, 21] employ Generative Adversarial Networks (GANs) to synthesize speech from video frames. GANs are known for their ability to generate realistic data. [22] train an attention-based encoder-decoder model to reconstruct speech from silent facial movement sequences without human annotations. [23] performs multi-modal supervision, leveraging text and audio to complement the insufficient word representations to reconstruct speech with correct contents from the input lip movements. While these previous works work well on limited datasets, they struggle to handle unconstrained datasets with significant head movements and a wide vocabulary, limiting their real-world applications. In contrast, our model primarily focuses on unconstrained lip-to-speech synthesis while also demonstrating its capabilities on small constrained datasets like GRID-4S and TCD-TIMIT.

- Unconstrained Lip-to-Speech synthesis

  Unconstrained lip-to-speech synthesis aims to generate natural speech from real-world videos featuring extensive vocabulary, diverse speakers, and significant head movements. This section reviews recent advancements in this field, highlighting key methodologies and their strengths and limitations in the context of the thesis. Pioneering work Lip2Wav by Prajwal *et al.* [24] introduced a model comprising a 3D Convolutional Neural Network (3D-CNN) for video feature learning and an LSTM-based autoregressive (AR) sequence-to-sequence architecture inspired by Tacotron2 [25] for single-speaker lip-to-speech synthesis. Their model generates mel-spectrograms, a compressed representation of speech, based on the input video frames. They also released a dataset comprising YouTube lecture videos of five different speakers, the largest single-speaker unconstrained audio-visual dataset at that point. This paved the way for further exploration into unconstrained Lip2Speech synthesis. He et al. [26] proposed a non-autoregressive (NAR) architecture to address the slow inference times associated with autoregressive models like Lip2Wav. Their model utilizes 3D convolutional blocks, a transformer-based condition module, and a Glow decoder module [27] for efficient mel-spectrogram refinement. Flow-based generative models, like the Glow model, use invertible transformations to create complex data distributions from simpler ones, enabling both efficient training and high-quality sample generation. This approach prioritizes faster inference speed, a crucial factor for real-world applications. Varshney et

al. [28] built upon Lip2Wav work by training a transformer model to learn a joint latent distribution for speech generation. This approach aims to capture the shared characteristics between speech and visual features for improved synthesis. Additionally, the VV-Memory architecture [29] tackles speaker independence by combining audio and visual information using a key-value memory structure. This enables video-to-speech reconstruction and speaker-independent speech retrieval.

Recent Advancements - End-to-End NAR and Self-Supervised Learning: Wang et al. [30] proposed a novel end-to-end NAR transformer for directly synthesizing speech from unconstrained videos. Their architecture incorporates a visual encoder, an acoustic decoder, and a GAN-based vocoder for generating audio from mel-spectrograms. This eliminates the intermediate step of mel-spectrogram generation. LipSound2 [22] investigated a different approach focusing on cross-modal self-supervised pre-training. Their encoder-decoder architecture utilizes a location-aware attention mechanism to map face image sequences to mel-scale spectrograms. This leverages self-supervised learning for feature extraction without explicit audio supervision. [31] utilize a diffusion model to capture speaker characteristics, a lip-reading model to infer text, and employ a diffusion vocoder to synthesize audio by combining these pieces of information. In contrast, our proposed differs by using content-rich SSL representations [32] and learning target speech representations from lip sequences.

2.2 Self-supervised representation for Lip-to-Speechsynthesis

Traditional approaches to Lip-to-Speech synthesis typically involve encoding lip or facial sequences into hidden states, which are then decoded to generate Mel-spectrograms. These spectrograms are subsequently transformed into time-domain waves by an independently trained vocoder. However, this method faces challenges due to the high correlation among Mel frames along both the time and frequency axes, potentially leading to performance degradation in the overall Lip-to-Speech synthesis process [33]. Furthermore, these Mel-spectrograms exhibit higher variance than the quantised speech SSL representations. This increased variance complicates the training process of a Speech Synthesis Transformer model. Consequently, despite the recent emergence of SSL representations, their utilization in lip-to-speech synthesis remains limited. As researchers continue to explore and refine these techniques, addressing these challenges will be crucial for advancing the effectiveness and efficiency of Lip-to-Speech synthesis systems. Self-supervised learning (SSL) has emerged as a powerful technique for learning informative representations from unlabeled data. In the context of speech, SSL models can learn meaningful representations directly from audio waveforms, capturing the underlying structure of speech without requiring explicit labels. While SSL offers promising advantages, its application in lip-to-speech synthesis is still in its early stages. This limited adoption can be attributed to the challenges mentioned above with Mel-spectrograms, which are often used as an intermediate representation between lip movements and speech.

VCVTS (Vector Quantized Contrastive Predictive Coding Transformer System) by Wang et al. [34] offers a novel approach to lip-to-speech synthesis. VCVTS utilizes Vector Quantized Contrastive Predictive Coding (VQCPC) to extract informative features from lip movements. VQCPC can be understood as a two-part process: Vector Quantization: This step involves converting the raw lip image features into a sequence of discrete codes. Imagine a large dictionary containing various lip shape patterns. VQCPC identifies the code in the dictionary that most closely resembles the current lip image, essentially compressing the information into a more manageable form. Contrastive Predictive Coding: Here, the system attempts to predict the next code in the sequence based on the current one. This prediction encourages the model to learn the temporal relationships between different lip shapes, which are crucial for capturing the dynamics of speech. VCVTS incorporates a separate speaker encoder. This encoder analyzes additional information, likely speaker identity or voice characteristics, to ensure the generated speech aligns with the target speaker. Additionally, a pitch predictor estimates the fundamental frequency of the speech signal, which is important for controlling the perceived highness or lowness of the voice. The extracted lip features (VQCPC codes), speaker information, and predicted pitch are then fed into a decoder network. This decoder aims to infer Mel-spectrograms, which represent the speech signal in terms of frequency and power. As discussed earlier, Mel-spectrograms have limitations, but in VCVTS, they serve as an intermediate step for speech generation. VCVTS employs a distinct voice conversion model. This model likely transforms the Mel-spectrogram and speaker information into a speaker-specific representation. Finally, a vocoder converts the processed Mel-spectrogram into the final time-domain waveform, which is the actual audible speech.

Revise [35] explores using self-supervised learning (SSL) for speech enhancement. They tackle the challenge of making speech clearer and more understandable, regardless of background noise or distortions. Revise uses a pre-trained model called AV-HuBERT. Unlike traditional models that need labelled speech data, AV-HuBERT learns from vast amounts of unlabeled videos, capturing the connection between audio and lip movements. This AV-HuBERT model forms the basis for a module called P-AVSR (combined audio-visual speech recognition). Given silent video input, P-AVSR predicts a sequence of discrete units, like building blocks of speech, based on the combined audio and visual information. Another module, P-TTS (modified text-to-speech synthesis), takes these predicted units and reconstructs the original speech waveform using a modified HiFi-GAN architecture [32]. Here, P-TTS acts like a regular text-to-speech system, but it receives the predicted speech units instead of text. Overall, Revise aims to enhance speech quality by leveraging the power of SSL representations learned from unlabeled data. P-AVSR can potentially identify and compensate for noise or distortions, while P-TTS generates a cleaner speech waveform. While AV-HuBERT provides a strong foundation, Revise acknowledges its limitations. To further refine speech generation, P-TTS is fine-tuned on a speech dataset like LJSpeech [36], helping it adapt to the natural speech patterns in the data. By combining audio-visual information with predicted speech units, Revise demonstrates the potential of SSL representations for improving speech quality from silent videos.

Our research builds upon the promising application of self-supervised learning (SSL) representations for lip-to-speech synthesis, as demonstrated in Revise. Both approaches leverage the power of SSL models trained on vast amounts of unlabeled data to overcome limitations associated with traditional methods requiring labelled speech data. However, our work introduces key distinctions that aim to improve upon Revise: Disentangled Feature Extraction: While Revise utilizes a combined audio-visual SSL model (AV-HuBERT), we employ separate features extracted from disentangled AV-HuBERT and HuBERT models [8]. This allows us to potentially capture richer information compared to a combined model. AV-HuBERT, by design, might learn some modality-agnostic representations, while separate models like HuBERT (focusing on audio) and the audio stream of AV-HuBERT can provide more specific auditory details. This potentially leads to a more comprehensive understanding of the speech information present in silent videos. Decoupled Training Procedure: Revise employs a combined training approach for their P-AVSR and P-TTS modules. We propose a decoupled training procedure for our lip-to-speech synthesis model. This separation can offer several advantages. First, it allows for independent optimization of each module, potentially leading to improved performance. Second, it provides more flexibility for future modifications or integrations with different modules in the pipeline. Focus on Speaker-Specificity: While Revise demonstrates the effectiveness of SSL for speech enhancement, existing works, including Revise, haven't fully explored its potential for speaker-specific lip-to-speech generation. Our research specifically addresses this gap. We aim to achieve speaker-specific speech synthesis directly from silent videos using SSL representations. This focus on speaker-specificity can lead to more natural and realistic speech generation, as the synthesized speech will resemble the voice characteristics of the person in the video. By incorporating these advancements, our work seeks to contribute significantly to the field of lip-to-speech synthesis. We leverage the power of SSL representations while addressing the limitations of previous approaches, ultimately aiming to achieve more natural, speaker-faithful speech generation from silent videos.

2.3 Speech Synthesis

In recent years, self-supervised learning (SSL) has emerged as a game-changer in text-to-speech (TTS) tasks. This approach allows models to learn informative representations from vast amounts of unlabeled data, circumventing the limitations of traditional methods that require large quantities of labelled speech data. This shift towards SSL offers exciting possibilities for improved speech synthesis quality and efficiency. Several recent advancements have solidified the potential of SSL for TTS. Pioneering works like SoundStream [37] and Encodec [38] have demonstrated the effectiveness of SSL representations in generating high-fidelity speech. These methods, alongside established models like HuBERT and Wav2Vec [39], contribute to a growing trend where researchers leverage the power of unlabeled data to achieve superior TTS performance.

The recent work on Vall-E [40] showcases another exciting development. Vall-E employs discrete representations during training, allowing the model to utilize massive datasets effectively. This approach

is evident in Vall-E's training on a staggering 60,000 hours of audio data, highlighting the potential of large-scale unlabeled data for achieving impressive TTS quality.

ParrotTTS [41] further emphasizes the power of HuBERT representations in TTS. This work harnesses HuBERT's capabilities to enhance speech synthesis and conducts comprehensive comparisons with other SSL techniques, providing valuable insights into the effectiveness of different approaches. BASE TTS [42] pushes the boundaries of TTS by incorporating a GPT architecture into the training process. This innovative approach leverages a vast corpus of text and speech data exceeding 100,000 hours. Additionally, BASE TTS introduces a novel speech tokenization method, further expanding the capabilities of TTS models. By combining these advancements, BASE TTS exemplifies the continuous evolution of TTS methodologies.

The developments mentioned above showcase the dynamic landscape of TTS research. The increasing adoption of SSL, exploration of large-scale datasets, and integration of novel architectures like GPT-based approaches all hold immense promise for the future of TTS. As research continues to expand upon these trends, we can expect even more impressive advancements in speech synthesis, leading to more natural, human-like speech generation.

While self-supervised learning (SSL) is revolutionizing text-to-speech (TTS), another exciting area is AI dubbing. This technology focuses on automatically generating dubbed speech that aligns with the lip movements of a person in a video, often based on a text script. Several promising approaches have emerged in AI dubbing. VDTTS [43] stands out for its ability to generate high-quality speech while maintaining temporal coherence with the visual cues in the video. NeuralDubber [44] demonstrates the potential of deep learning for realistic and expressive dubbing. HPMDubbing [45] takes a learning-based approach, focusing on generating high-fidelity speech while preserving the speaker's emotional characteristics. However, a key limitation exists with many current AI dubbing methods. These models are often designed for specific input modalities. For example, a system trained on text scripts might struggle to adapt to other inputs. This lack of flexibility hinders the broad applicability of these techniques. Recent research is addressing this limitation.

In contrast, our research endeavours to pioneer a unified speech synthesis model capable of accommodating various input modalities. This model aims to combine the strengths of SSL representations for speech and Lip2Speech techniques, allowing it to handle various input modalities, including text alone, lip movements alone, or a combination of both. By exploring the integration of multiple modalities and employing mechanisms such as multi-head attention (MHA), we aim to develop a versatile speech synthesis framework. This offers greater flexibility and applicability compared to existing methods, which are limited to specific modalities.

Chapter 3

RobustL2S: Speaker-Specific Lip-to-Speech Synthesis exploiting Self-Supervised Representations

3.1 Introduction

Understanding lip movements offers a distinct advantage in situations where auditory cues are unavailable. It proves particularly valuable for individuals with hearing impairments, and speech disorders and aids in speech rehabilitation by providing visual feedback [46]. The synthesis of accurate speech from lip movements can assist in tasks such as movie dubbing [45], language learning, forensic investigations [47], video conferencing in noisy conditions (Fig. 1.1), voice inpainting [48] or giving artificial voice to people who cannot produce intelligible sound.

The problem of Lip-to-Speech synthesis is inherently ill-posed because a sequence of lip movements can correspond to multiple possible speech utterances [24]. Additional challenges arise from factors such as head pose movements, non-verbal facial expressions, variations in capture quality, and ambient noise, which further complicate the problem. Reliance on contextual information, such as environment, place, topic, etc., can help alleviate the Lipreading challenges [49, 50]; however, such information may not always be available.

Most existing approaches constitute an encoder-decoder architecture; the encoder maps the lip sequence to intermediate representations, which are then directly decoded into mel-spectrograms. The major drawback of this approach is that apart from speech content, the decoder is also forced to predict the time-varying speaker and ambient noise characteristics present in the ground truth Mel. We hypothesize that this dependence hurts the model's performance in terms of speech intelligibility, reducing its usability for various downstream applications [51]. Our work addresses these limitations by taking a modularized approach, exploiting the advances in Self-supervised learning (SSL) in audio and audio-visual scenarios.

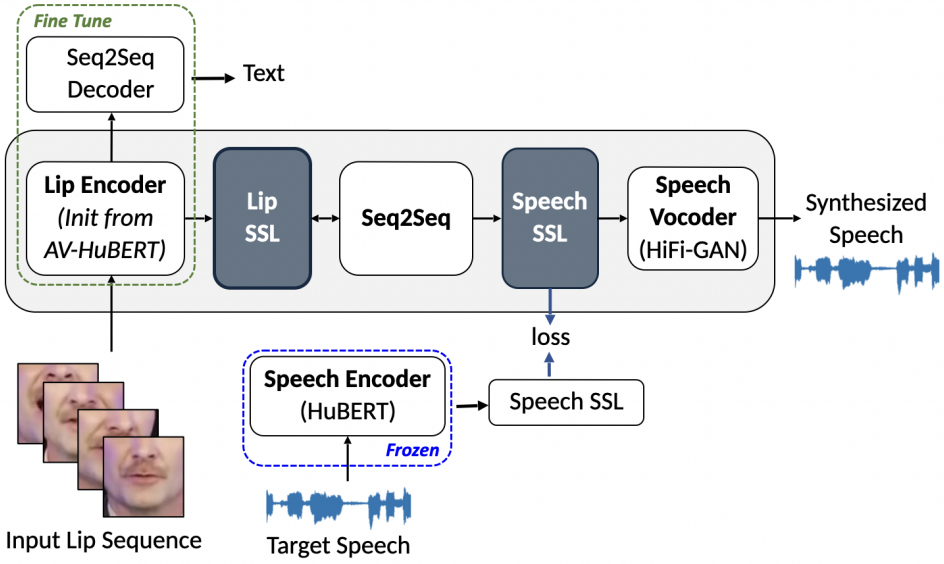

Figure 3.1: The proposed RobustL2S model utilizes lip encoder and speech encoder to extract SSL representations from lip sequences and their corresponding speech. A Seq2Seq model maps the lip representations to speech representations, which are then decoded to synthesize speech.

In contrast to direct mel prediction from lip features, we take a two-staged approach as depicted in the Fig. 3.1. The first step extracts SSL representations of lip sequences and maps them to corresponding speech SSL representations using a sequence-to-sequence (Seq2Seq) model. The key idea is to use speech embeddings that disentangle the content from the speaker and ambient information. The second stage maps the content-rich speech embeddings to raw speech using a speaker-conditioned vocoder. The proposed RobustL2S framework simplifies training and brings robustness to variations in head-pose, ambient noise, and time-varying speaker characteristics, leading to significant gains in speech intelligibility. To validate the efficacy of our approach, we perform comprehensive experiments on GRID [11], TCD-TIMIT [12] and Lip2Wav [24] datasets. The quantitative measures and MOS scores show that the synthesized speech generated by our method accurately represents the intended content and improves on the intelligibility/naturalness compared to current state-of-the-art methods [30, 34] on all three datasets.

More formally, our work makes the following contributions: (1) We propose a novel modularized framework for Lip-to-Speech synthesis exploiting self-supervised embeddings for both lip and speech sequences (2) A Seq2Seq network for cross-modal knowledge transfer to map lip SSL representations to speech SSL representations; and (3) Thorough experimental results demonstrating that RobustL2S is capable of synthesizing high-quality speech, achieving state-of-the-art results in objective and subjective evaluation without requiring additional data augmentation [30].

3.2 Method

3.2.1 Preliminaries

Fig. 3.1 illustrates the proposed RobustL2S framework. RobustL2S consists of four modules: two encoders - an encoder that extracts lip representation from video sequences and a pre-trained HuBERT model that generates speech features for corresponding speech sequences, a Seq2Seq model that maps lip representations to speech representations, and a modified HiFi-GAN vocoder that synthesize speech using the speech representations. We introduce four functions as follows:

- $f_{\text{l}} : L^{T \times W \times H} \mapsto L_{\text{ssl}}$, which maps the input lip sequence to its corresponding SSL representation. Here, T represents the number of time steps (frames), and H and W correspond to the spatial dimensions of the frames.
- $f_{\text{s}} : X \mapsto S_{\text{ssl}}$, which maps the ground-truth raw speech to its corresponding SSL representation.
- $f_{\text{s2s}} : L_{\text{ssl}} \mapsto S_{\text{ssl}}$, which maps the lip representation to its corresponding speech representation.
- $f_{\text{voc}} : S_{\text{ssl}} \mapsto \widehat{X}$, which maps the speech representation to the synthesized speech $\widehat{X}$.

3.2.2 Encoder

We use HuBERT [8] for f_{s}, to extract speech representation of target speech signals. HuBERT representation is content rich [32] and agnostic to other variations. Although our framework is compatible with various off-the-shelf SSL models, we specifically utilize AV-HuBERT [52] for f_{l}, our video encoder to extract lip representations. This choice is based on the similarity in training methods between AV-HuBERT and HuBERT, ensuring content-rich representations for improved intelligibility, which aligns with our goals.

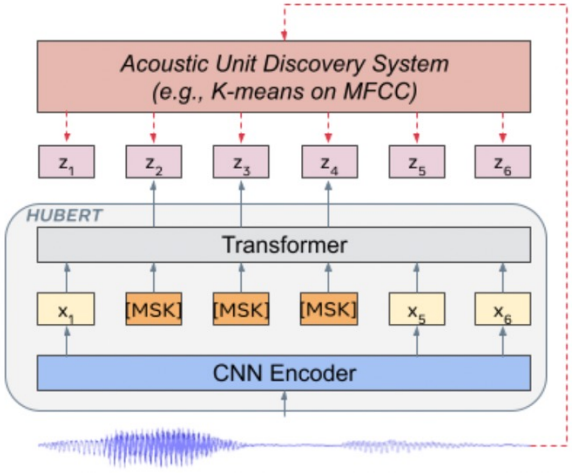

Figure 3.2: Masked prediction based SSL pretraining for speech - HuBERT pretraining model architecture [8].

HuBERT pretraining framework is depicted in Fig 3.2. The HuBERT and AV-HuBERT models are trained using a masked-prediction loss to predict cluster IDs, which are learned using k-means clustering. The labels for the first iteration are derived by clustering MFCC features derived from acoustic frames. For subsequent iterations, more complex features derived from an audio or audio-visual encoder, depending on the specific model being used, are clustered. AV-HuBERT incorporates a modified ResNet [53, 54] as its frontend, coupled with a transformer encoder. We also finetune the pretrained AV-HuBERT using an attention-based Seq2Seq cross-entropy loss as in [52].

3.2.3 Seq2Seq model

In recent years, Seq2Seq models have gained significant attention in the field of cross-domain generation. The core concept of our approach is to align representations from two different domains - visual and audio, that share a common generating process. By recovering correspondences between these domains, we facilitate the transfer of knowledge from one domain to the other.

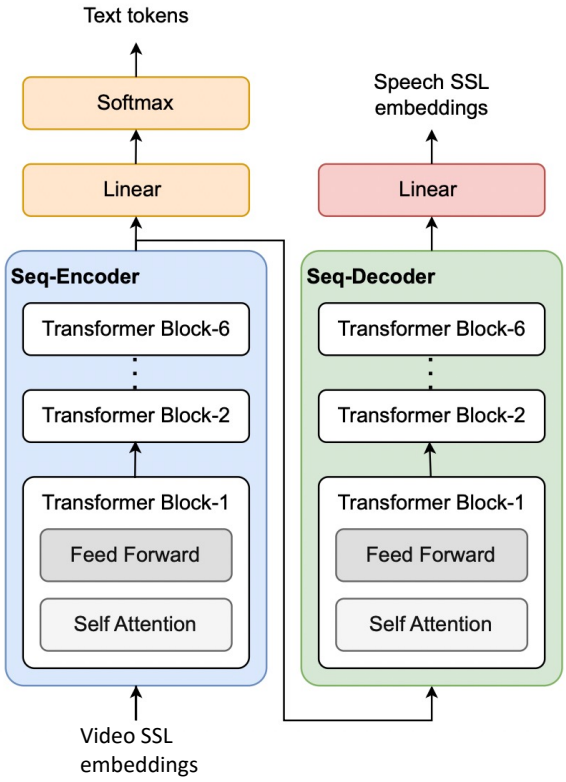

Figure 3.3: Architecture of proposed seq2seq model

Our Seq2Seq model 3.3, denoted as f_{s2s}, adopts a non-autoregressive-based encoder-decoder architecture to map lip representations to their corresponding speech representations. The encoder and decoder consist of feed-forward transformer blocks with self-attention [55], along with 1-dimensional

convolutions inspired by Fastspeech2 [56]. A transposed convolution layer is used at the encoder to match the rate of video and audio representations. The encoder takes the lip representation L_{ssl} and encodes it into a sequence of fixed-dimensional vectors. The decoder generates predictions for all representations of S_{ssl} simultaneously. We train three versions of the Seq2Seq model:

- $f_{\text{s2s-units}}$: This encoder-decoder architecture utilizes Cross-Entropy (CE) loss to train the model on the decoded speech units. The input to the architecture consists of cluster IDs from the video encoder, and the decoder predicts the corresponding HuBERT cluster IDs for the audio. The objective can be written as:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{N} S_{\text{ssl\_units\_i}} \log(\hat{S}_{\text{ssl\_units\_i}}), \tag{3.1}$$

  where $S_{\text{ssl\_units}}$ are the ground-truth speech units, $\hat{S}_{\text{ssl\_units}}$ are the decoded speech units, and N are the number of HuBERT units.

- $f_{\text{s2s-features}}$: Here the model learns mapping from audio-visual feature to corresponding speech feature vectors. This model utilizes L1 loss, quantifying the difference between the decoded features and ground-truth speech features. The objective can be written as:

$$\mathcal{L}_{\text{L1}} = \frac{1}{T}\sum_{i=1}^{T} |S_{\text{ssl\_features\_i}} - \hat{S}_{\text{ssl\_features\_i}}|, \tag{3.2}$$

  where $S_{\text{ssl\_features}}$ are the ground-truth speech features, $\hat{S}_{\text{ssl\_features}}$ are the decoded speech features, and T is the time-steps.

- $f_{\text{s2s-features-ctc}}$: This architecture follows the same structure as $f_{\text{s2s-features}}$, but also includes an additional fully connected linear head to predict CTC tokens after the encoder layer. For given input lip representation $L_{\text{ssl}} \in \mathbb{R}^{TxD}$ of length T and dimension D, let Enc_{ssl} be the output of the encoder. The goal is to minimize the negative log-likelihood by using $P_{CTC}(S_{\text{ssl}}|Enc_{\text{ssl}})$ to train the model effectively using the CTC approach and is defined as:

$$\mathcal{L}_{CTC} := -\log P_{CTC}(S_{\text{ssl}}|Enc_{\text{ssl}}). \tag{3.3}$$

  By weighted summing the L1 and CTC loss functions, the objective function can be formulated as:

$$\mathcal{L}_{\text{Tot}} = \alpha_{CTC} * \mathcal{L}_{\text{CTC}} + \alpha_{L1} * \mathcal{L}_{\text{L1}}, \tag{3.4}$$

  where $\alpha_{\text{CTC}} \in \mathbb{R}$ and $\alpha_{\text{L1}} \in \mathbb{R}$ are the hyperparameter that balances the influence between two loss.

3.2.4 Speech Vocoder

We use a modified version of HiFiGAN-v2 [6] to synthesize speech. It has a generator G and a discriminator D. G runs S_{ssl} through transposed convolutions for upsampling to recover the original sampling rate followed by residual block with dilations to increase the receptive field to synthesize the signal, $\widehat{X} := G(S_{\text{ssl}})$.

The discriminator in our model has the task of distinguishing the synthesized signal $\widehat{X}$ from the original signal X. It is evaluated using two sets of discriminator networks. The multi-period discriminators operate on equally spaced samples of the signals, focusing on capturing temporal patterns and characteristics. On the other hand, the multi-scale discriminators analyze the input signal at different scales, enabling the model to capture both fine-grained details and global structure. The primary objective of the model is to minimize the discrepancy, measured by $D(X, \widehat{X})$, between the original signal and the synthesized signal. This optimization process applies to all the parameters of the speech decoder, improving its overall performance and fidelity.

3.3 Experiments

3.3.1 Datasets

- **Lip2Wav**: The Lip2Wav dataset [24] is a large, person-specific, unconstrained dataset, commonly used for learning Lip-to-Speech synthesis for individual speakers. It consists of real-world lecture videos featuring 5 different speakers. Each speaker has approximately 20 hours of video data, and the vocabulary size exceeds 5000 words for each speaker. We do experiments on all five speakers: Chess Analysis (chess), Chemistry Lectures (chem), Hardware Security (hs), Deep Learning (dl), and Ethical Hacking (eh).

- **GRID-4S**: The GRID-4S is a subset of the GRID audio-visual dataset [11] specifically designed for constrained Lip-to-Speech synthesis. This subset includes two male speakers ($s1$, $s2$) and two female speakers ($s4$, $s29$), which are frequently used in the literature [20, 24]. The videos in the dataset were captured in an artificial environment. The vocabulary used in GRID-4S is limited to only 51 words. The sentences in the dataset follow a restricted grammar, with each sentence containing 6 to 10 words.

- **TCD-TIMIT-3S**: The TCD-TIMIT-3S is a subset of the TCD-TIMIT dataset, which comprises recordings of 62 speakers captured under studio conditions. Among these speakers, three are trained lip-speakers. The primary objective of selecting this subset was to enable comparison with previous studies [20, 24]. Our focus was solely on the audio-visual data generated by these three lip-speakers. Each lip-speaker delivers 375 distinct sentences that exhibit phonetic diversity. Additionally, all three lip-speakers speak two sentences.

3.3.2 Implementation details

- **Data preparation**: For the GRID-4S and TCD-TIMIT-3S datasets, we adhere to the convention of randomly selecting 90% data for training, 5% for validation, and 5% for testing, as established in previous works [57, 24, 18, 30]. For Lip2Wav, we adopt the official data split [24]. For consistency, in line with previous works, we evaluated RobustL2S on the Lip2Wav dataset using a speaker-dependent setting [24, 14, 29]. This involved training the network separately with individual speakers. However, for the GRID-4S and TCD-TIMIT-3S datasets, we evaluated RobustL2S in a constrained (seen) speaker setting [57, 24, 18]. The video sequences are resampled to a frame rate of 25 frames per second (fps), while the raw audio is sampled at 16 kHz. We utilize the SFD [58] face detector to detect 68 key points, allowing us to crop a mouth-centered region-of-interest measuring 96×96 pixels. In order to solely assess the advantages of using SSL representation in our proposed setup, we opt not to employ any data augmentation techniques to enhance the quality of synthesized speech. The Lip2Wav dataset does not provide transcripts, so we rely on the Whisper *small* model [59] to extract transcripts. These transcripts are then used to fine-tune the AV-HuBERT model.

- **SSL representation**: We utilize the official fairseq repository implementation of the BASE models AV-HuBERT [52] and HuBERT [8] for our experiments. We fine-tune the AV-HuBERT pre-trained model with an attention-based STS cross-entropy loss for visual speech recognition [52]. To achieve this, a transformer decoder is added to the pre-trained model, which autoregressively decodes the AV-HuBERT features to target character probabilities. The fine-tuned AV-HuBERT model extracts SSL representations for lip sequences, while HuBERT is employed to extract representations from speech signals. Both models provide 768-dimensional features. For $f_{\text{s2s-units}}$ model, following the approach in [52, 8], the lip features are clustered into 2000 AV-HuBERT units, while the speech features are clustered into 100 HuBERT units, using k-means clustering. For $f_{\text{s2s-features}}$ model, the output features from HuBERT and AV-HuBERT models are used directly.

- **Seq2Seq model**: Our model comprises a 6-layer transformer encoder and decoder with a hidden dimension of 512 and 2 attention heads. we set the batch size to 32 and the maximum number of steps to $20,000$. We employ the Adam optimizer with an initial learning rate of 4.4 x 10^{-2}, along with an annealing rate of 0.3 and annealing steps at $[3000, 4000, 5000]$. The HuBERT model encodes speech into features at a frame rate of 50 Hz, while the SSL unit from AV-HuBERT is encoded at 25 Hz. To match these rates, we incorporate a lightweight transposed convolution layer with a kernel size of $[4, 3]$ and a stride length of $[2, 1]$. We set α_{CTC} and α_{L1} mentioned in (3.4) to 0.001 and 1, respectively.

- **Speech Vocoder**: We train modified HiFiGAN-v2[1] to generate audio from speech SSL representations for all speakers. This model employs encoding of raw audio into a sequence of discrete tokens from a set of 100 possible HuBERT tokens, with a code hop size of 160 raw audio samples. We set the batch size to 16, the learning rate to $2\text{x}10^{-4}$, the number of embeddings to 100, the embedding dimension to 128, and the model input dimension to 256. We train the model up to $300k$ steps. Following the approach in [60], F0 is not used as a feature in our training process. The aforementioned vocoder configuration is effective for speech units. However, for our investigated feature-based models, $f_{\text{s2s-features}}$ and $f_{\text{s2s-features-ctc}}$, we apply a pre-trained k-means[2] clustering model trained on HuBERT features. During the inference phase, the generated features undergo k-means clustering to obtain discrete speech units, which are then passed through the speech vocoder. We train the vocoders for the three datasets separately.

3.3.3 Evaluation metric

During our evaluation, we employ several metrics to assess the quality of the synthesized speech. These include: Word Error Rate (WER), Short-Time Objective Intelligibility (STOI) [61], and Extended Short-Time Objective Intelligibility (ESTOI) [62]. Additionally, we conduct subjective evaluations using Mean Opinion Score (MOS), where human evaluators rate the quality, intelligibility, and naturalness of the synthesized speech based on their subjective perception.

3.4 Results

3.4.1 Need for Seq2Seq model

Table 3.1: Performance comparison: Seq2Seq model vs. evaluated variations vs. no Seq2Seq model employed on chemistry speaker of Lip2Wav dataset.

Baseline (Ours)	**STOI ↑**	**ESTOI ↑**
f_l(pre-trained) + f_{voc}	0.447	0.22
f_l(finetuned) + f_{voc}	0.50	0.27
f_l(finetuned) + $f_{\text{s2s-units}}$ + f_{voc}	0.18	0.013
f_l**(finetuned) +** $f_{\textbf{s2s-features}}$ **+** f_{voc}	**0.583**	**0.397**
f_l(finetuned) + $f_{\text{s2s-features-ctc}}$ + f_{voc}	0.557	0.368

[1] https://github.com/facebookresearch/speech-resynthesis

[2] https://github.com/facebookresearch/fairseq/tree/main/examples/textless_nlp/gslm/speech2unit

We tested our hypothesis of using a Seq2Seq model on the Lip2Wav dataset, specifically for a chemistry speaker, and report our findings in Table 3.1. Fine-tuning the AV-HuBERT model using transcripts consistently improved objective metrics by approximately 0.05 units on both the metrics compared to the pre-trained version. Deploying the Seq2Seq model on the finetuned AV-HuBERT features (f_l(finetuned) + $f_{\text{s2s-features}}$ + f_{voc}) resulted in an increase of approximately 0.08 and 0.12 units in STOI and ESTOI metrics, respectively, compared to not using the Seq2Seq model (f_l(finetuned) + f_{voc}). These results highlight the effectiveness of our Seq2Seq approach using SSL representations for Lip-to-Speech synthesis. The significant performance gap (approximately 0.39 units on both metrics) between the Seq2Seq model using SSL features and the model using SSL units on both evaluated metrics approximates the amount of information lost in speech reconstruction when audio-visual sequences are represented as SSL units instead of SSL features. From now on, we will refer to f_l(finetuned) + $f_{\text{s2s-features}}$ + f_{voc} as RobustL2S. The inclusion of CTC loss in our f_l(finetuned) + $f_{\text{s2s-features-ctc}}$ model resulted in a statistically insignificant decrease of approximately 0.02 units in STOI compared to the model without CTC loss, f_l(finetuned) + $f_{\text{s2s-features}}$. The decrease may be due to the lack of ground-truth transcripts in the Lip2Wav dataset. However, when evaluating our model using CTC loss on datasets (GRID-4S and TCD-TIMIT-3S) with ground-truth transcripts, we observed a slight increase of 0.02 units. Nevertheless, our focus is on working with datasets in the wild that generally lack ground-truth transcripts, so we proceed with experiments excluding the CTC loss.

3.4.2 RobustL2S in Constrained settings

Table 3.2: Performance comparison in constrained-speaker setting on GRID-4S dataset

Method	**STOI** ↑	**ESTOI** ↑	**WER** ↓
Vid2speech [13]	0.491	0.335	44.92 %
Lip2AudSpec [15]	0.513	0.352	32.51 %
1D GAN-based [18]	0.564	0.361	26.64 %
Vocoder-based [57]	0.648	0.455	23.33 %
Ephrat *et al.* [14]	0.659	0.376	27.83 %
Lip2Wav [24]	0.731	0.535	14.08 %
VAE-based [16]	0.724	0.540	-
VCA-GAN [20]	0.724	**0.609**	12.25 %
kim *et al.* [29, 63]	0.738	0.579	-
RobustL2S	**0.754**	0.571	**11.21** %

Table 3.3: Performance comparison in constrained-speaker setting on TCD-TIMIT-3S dataset

Method	**STOI** ↑	**ESTOI** ↑	**WER** ↓
Vid2speech [13]	0.451	0.298	75.52 %
Lip2AudSpec [15]	0.450	0.316	61.86 %
1D GAN-based [18]	0.511	0.321	49.13 %
Ephrat *et al.* [14]	0.487	0.310	53.52 %
Lip2Wav [24]	0.558	0.365	31.26 %
VCA-GAN [20]	0.584	0.401	-
RobustL2S	**0.596**	**0.452**	**29.03** %

Table 3.2 and 3.3 summarizes the performance of our RobustL2S in the context of Lip-to-Speech synthesis using constrained datasets: GRID-4S and TCD-TIMIT-3S. We compare our results with existing Lip-to-Speech synthesis works, including state-of-the-art approaches. We report the mean test scores on all four speakers of the GRID-4S dataset and all three speakers of the TCD-TIMIT-3S dataset, as documented in previous works. Remarkably, our RobustL2S approach demonstrates significant improvements in terms of STOI and WER metrics when compared to other approaches. This improvement is particularly noticeable on the TCD-TIMIT-3S dataset, which contains a larger number of novel words that were unseen during training. This observation highlights the ability of our RobustL2S to accurately pronounce new words and effectively capture semantic information from lip movements, resulting in the generation of more intelligible speech.

3.4.3 RobustL2S in Unconstrained settings

Table 3.4 provides a synopsis of RobustL2S's performance on the Lip2Wav dataset. This dataset includes a significant amount of silences between words, and RobustL2S shows a notable improvement across all metrics. Despite Lip2Wav's data asynchrony issues, which may affect the quality of the generated speech, RobustL2S demonstrates substantial performance gains in objective metrics, highlighting its overall superiority in producing intelligible speech. However, it is worth noting that RobustL2S performs similarly or slightly worse than [29] on the *hs* and *eh* ESTOI metric. This could potentially be attributed to the poor resolution ($480p$ and $360p$) of the original videos, making it challenging to accurately recognize lip regions.

Table 3.4: Performance comparison in speaker-dependent setting on Lip2Wav dataset

Speaker	**Method**	**STOI ↑**	**ESTOI ↑**
Chemistry Lectures (chem)	Ephrat *et al.* [24]	0.165	0.087
	GAN-based [64]	0.192	0.132
	Lip2Wav [24]	0.416	0.284
	Hong et al. [29]	0.566	**0.429**
	RobustL2S	**0.583**	0.397
Chess Analysis (chess)	Ephrat *et al.* [24]	0.184	0.098
	GAN-based [64]	0.195	0.104
	Lip2Wav [24]	0.418	0.290
	Hong et al. [29]	0.506	0.334
	RobustL2S	**0.517**	**0.340**
Deep Learning (dl)	Ephrat *et al.* [24]	0.112	0.043
	GAN-based [64]	0.144	0.070
	Lip2Wav [24]	0.282	0.183
	Hong et al. [29]	0.576	0.402
	RobustL2S	**0.627**	**0.419**
Hardware Security (hs)	Ephrat *et al.* [24]	0.192	0.064
	GAN-based [64]	0.251	0.110
	Lip2Wav [24]	0.446	0.311
	Hong et al. [29]	0.504	0.337
	RobustL2S	**0.511**	**0.337**
Ethical Hacking (eh)	Ephrat *et al.* [24]	0.143	0.064
	GAN-based [64]	0.171	0.089
	Lip2Wav [24]	0.369	0.220
	Hong et al. [29]	0.463	**0.304**
	RobustL2S	**0.493**	0.277

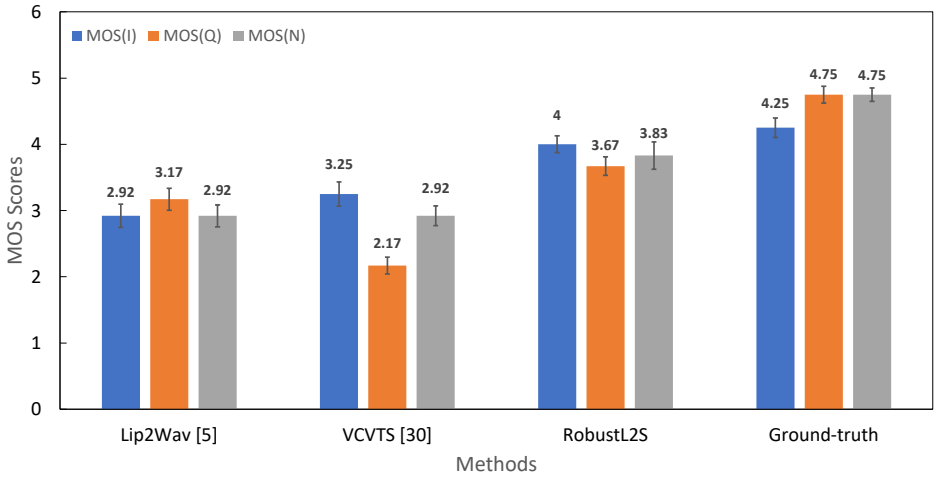

Figure 3.4: MOS scores on Intelligibility, Quality, and Naturalness with their 95 % confidence interval computed from their t-distribution on GRID-4S dataset

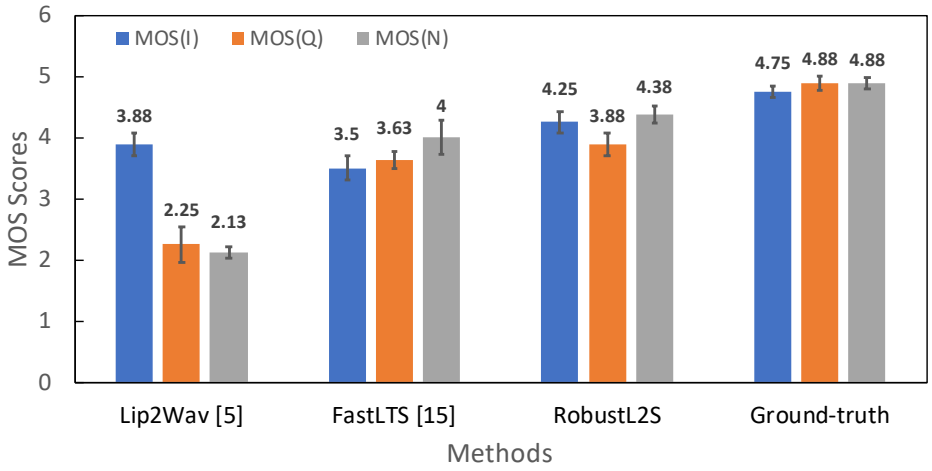

Figure 3.5: MOS scores on Intelligibility, Quality and Naturalness with their 95 % confidence interval computed from their t-distribution on Lip2Wav dataset

3.4.4 Subjective evaluation

Fig. 3.4 and Fig. 3.5 shows the MOS scores on intelligibility (MOS(I)), quality (MOS(Q)), and naturalness (MOS(N)) of synthesized speech from evaluated methods on Grid-4S and Lip2Wav datasets. We requested ten English-proficient subjects to score five randomly selected samples from different meth-

ods on the Lip2Wav and GRID-4S datasets. It can be observed that our model outperforms the evaluated methods, exhibiting higher Mean Opinion Score (MOS) values. This demonstrates that the proposed approach inherits the advantages of disentangled SSL features and the mapping of lip sequences to content-specific information. As a result, our model not only inherently improves the intelligibility aspect of synthesized speech but also generates speech that is highly natural and of high quality.

3.5 Conclusions

We propose a novel framework for Lip-to-Speech system, called RobustL2S, which accurately synthesizes spoken content from silent videos. This is accomplished by utilizing a non-autoregressive based sequence-to-sequence model to establish an inter-modality mapping, allowing us to learn a suitable decoding space from the lips' self-supervised (SSL) representations. We further demonstrate the effectiveness of mapping SSL features rather than SSL units for synthesizing intelligible speech. Both quantitative and qualitative results showcase state-of-the-art performance in constrained settings (such as GRID and TCD-TIMIT) and unconstrained settings (like Lip2Wav). In our future work, we aim to introduce emotive effects in the synthesized speech, considering that HuBERT embeddings are known to lack prosody information. Additionally, we plan to explore diffusion-based speech vocoders and their application in a multi-lingual setup.

Chapter 4

OmniSpeak: Towards a Unified Speech Generation Model

4.1 Introduction

The human experience with language is multifaceted. We effortlessly integrate textual information with visual cues, like lip movements, to create clear and nuanced communication. In dubbing, for example, actors flawlessly synchronize their speech with the on-screen character's lip movements based on the script. This remarkable ability highlights the interconnectedness of speech and visual information in our communication. Natural Language Processing (NLP) has mirrored this human capability with advancements in Text-to-Speech (TTS) and Lip-to-Speech (Lip2Speech) models. However, these models currently function independently, hindering their potential in applications that demand seamless text and visual information merging.

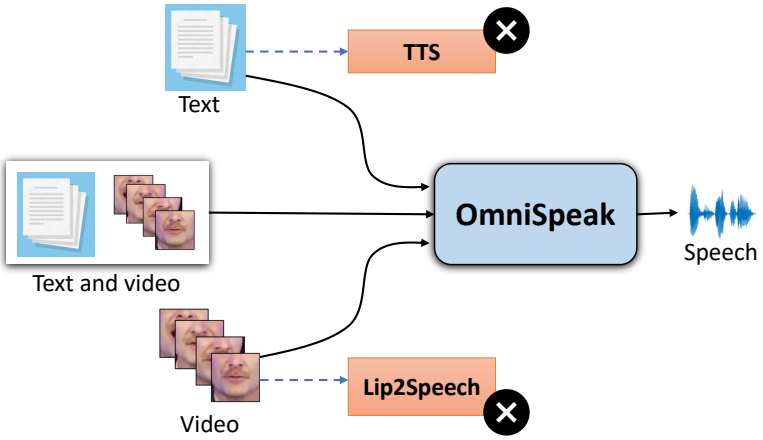

Figure 4.1: Proposing a unified model for multimodal input - speech synthesis tasks. Instead of training separate TTS or Lip2Speech or Combined input-based models, train one unified model to handle all input types.

The recent advancements in Self-Supervised Learning (SSL) have yielded powerful representations that are transforming various Natural Language Processing (NLP) tasks. This progress has also spurred significant development in speech generation, with Text-to-Speech (TTS) and Lip-to-Speech (Lip2Speech) models achieving impressive results. However, these models typically operate in isolation, requiring separate training and storage resources. The possibility of creating a unified speech generation model that integrates these modalities into a single, cohesive system capable of bridging the gap between textual input, visual cues, and spoken output is explored here. This model would be capable of: Text-to-Speech (TTS): Generating natural-sounding speech from a given text input. Lip-to-Speech (Lip2Speech): Synthesizing speech based solely on visual information of lip movements. Combined (Text & Lips)-to-Speech: Combining both text and lip movements to generate synchronized speech.

The proposed combined model as depicted in Fig. 4.1, holds immense potential for various applications across industries. Applications like automatic voice-over in dubbing can significantly benefit from a model that synchronizes speech with lip movements, with or without a given script. This model could automate the generation of realistic-sounding dubbed speech, eliminating the need for laborious post-production adjustments. Moreover, in educational settings, the model could facilitate language learning by providing feedback on pronunciation and lip synchronization, enhancing the efficacy of language acquisition tools. By combining multiple speech generation modalities, this model would eliminate the need for separate TTS and Lip2Speech systems in various applications. This would improve efficiency by reducing storage and training resources. Recent research suggests the possibility of creating lightweight TTS models like ParrotTTS [41]. Expanding on such a model to integrate visual cues ensures a unified yet lightweight model. The lightweight nature of the proposed model ensures that it can be deployed across a wide range of devices, democratizing access to advanced speech synthesis technology. This expands the reach of such technology and opens avenues for innovative applications in fields such as accessibility, entertainment, and human-computer interaction.

4.2 Method

4.2.1 Both text and video as input

A crucial aspect of creating a unified speech generation model is effectively combining text and video inputs. This allows the model to leverage the complementary information present in both modalities, ultimately leading to improved lip synchronization and speech quality. Inspired by the HPMDubbing [45], we propose using a Multi-Head Attention (MHA) layer to fuse the textual and visual features, as depicted in Figure 4.2. This layer learns to selectively attend to specific parts of the text (keys and values) based on the relevant video information (queries). In this way, the model can focus on textual segments corresponding to the video's visualised lip movements.

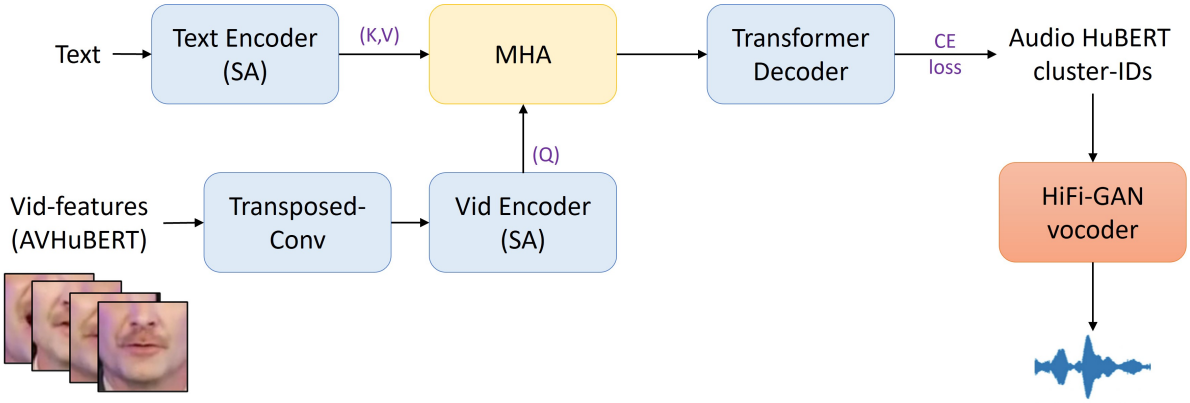

Figure 4.2: Proposed architecture to merge text and video at input for combined speech synthesis

Prior to feeding the text into the MHA layer, we perform essential preprocessing steps. This includes cleaning the text to remove unnecessary characters or punctuation. Additionally, for the English language, we convert the cleaned text into phonemes, which are the basic units of speech sound. This step helps the model better understand the pronunciation of the text and map it to the corresponding visual cues in the video. For the video input, we utilize AV-HuBERT features as used in RobustL2S in the previous chapter. These features capture the relationship between audio and visual information in the video, making them ideal for our task. Similar to existing works, we employ a transposed convolution layer to project these features onto the desired audio duration. This ensures that the generated speech aligns temporally with the visual cues in the video.

Drawing inspiration from ParrotTTS, we utilize HuBERT cluster IDs instead of raw HuBERT features. Since phonemes are involved along with video, this approach provides a more compact and content-rich representation of speech compared to raw features.

4.2.2 Text input

The text-to-speech framework is similar to ParrotTTS. The text input is cleaned to remove any unnecessary characters or punctuation. In the case of English text, phonemization is performed. Phonemes are the basic building blocks of spoken language, representing the individual sounds that make up words. By converting the text into phonemes, we provide the model with a more granular understanding of the pronunciation. We leverage the Montreal Forced Aligner (MFA) tool to achieve accurate phoneme extraction. This tool aligns a text transcript with a corresponding audio recording, providing the phonemes and their individual durations. Post this, based on the number of samples in a HuBERT frame, the duration (in terms of the number of times the phoneme needs to be repeated to achieve the desired audio duration) corresponding to each phoneme is extracted and saved. Here too, HuBERT cluster IDs are used as speech representation. Inspired by FastSpeech2 [56], we incorporate a duration prediction module within our model.

4.2.3 Video as input

Similar to the approach taken in RobustL2S, we extract video features using AV-HuBERT. Inspired by ParrotTTS, and as we are trying to build a unified model, we use HuBERT cluster IDs as speech representation. This provides a more content-rich representation compared to raw features. This enriched representation ultimately leads to a more accurate and natural-sounding output when combined with the text information. Since we need to combine the modalities at the input, the duration predictor from the previous section is used here also instead of using transposed-convolution layers as in some of the previous works, to match the input video dimension to that of the output speech dimension.

By employing the duration predictor and content-rich HuBERT cluster IDs, our model establishes a stronger link between the visual cues in the video and the generated speech. The duration predictor ensures that the generated speech aligns temporally with the lip movements, while the HuBERT cluster IDs provide crucial semantic context, leading to a more natural and cohesive final output.

4.2.4 Unified Model

A transformer encoder-decoder architecture similar to the one employed in ParrotTTS is used for learning sequence-to-sequence mapping between the input modalities and the output speech modality mentioned in the above sections. This architecture is well-suited for tasks involving sequence-to-sequence learning.

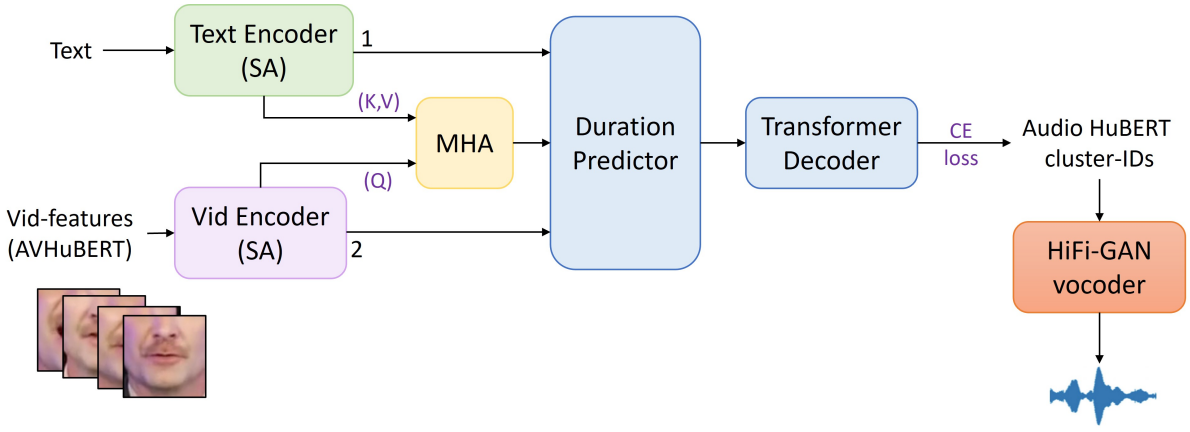

Figure 4.3: Proposed architecture for unified speech synthesis model. Model alternates between different inputs (1,2,3) during training.

To facilitate effective learning from both text and video modalities, our model employs an alternating training strategy. A random number predictor is used to dynamically choose which modality (text or video or combined) serves as the primary input for each training iteration. This approach encourages the model to learn independently from each modality while also fostering the ability to integrate them seamlessly.

When the model encounters video input, either alone or combined with text, the desired duration for the generated speech is directly derived from the video length. Similar to the approach used in RobustL2S, we can achieve this by replicating each input feature within the video twice. This ensures the generated speech temporally aligns with the visual cues in the video. On the other hand, when only text input is provided, the model relies on the duration information extracted during the text preprocessing stage, as detailed in the "Text Input" section. This information is typically obtained using tools like Montreal Forced Aligner (MFA).

To optimize the training process, we employ a combination of loss functions. For predicting HuBERT cluster IDs, we utilize cross-entropy loss. This loss function measures the dissimilarity between predicted and actual probability distributions of speech representation. It penalizes incorrect predictions with higher magnitudes, guiding the model towards more accurate cluster ID prediction. Furthermore, we employ the mean squared error (MSE) loss function for duration prediction. The loss is employed on log duration predictions. This choice is appropriate because we aim for the predicted log duration to closely match the desired log duration. By minimizing the MSE loss, the model learns to accurately predict the temporal requirements for the generated speech. Therefore the loss function is:

$$L = \alpha * L_{CE} + \beta * L_{MSE} \tag{4.1}$$

$$L_{CE} = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(p_{i,c}) \tag{4.2}$$

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(d_i - \hat{d}_i)^2 \tag{4.3}$$

α and β are scaling factors, d_i is log-duration, y_i represents the true HuBERT unit and p_i represents the predicted unit, and C in this case is 100, it is the distinct HuBERT IDs present in data.

Modified HiFi-GAN based vocoder is trained to convert the discrete speech representations to speech signal.

In summary, the alternating training strategy, combined with the integration of the duration predictor and appropriate loss functions, allows our unified model to effectively learn from text, video and combined modalities.

4.3 Experiments

4.3.1 Datasets

English Dataset: To train and evaluate our unified speech generation model for English, we utilize the Lip2Wav-Chemistry dataset released by neuralDubber. This dataset comprises YouTube lecture video recordings, providing a rich source of synchronized audio and visual speech information. It serves as a valuable resource for training a model capable of generating speech that aligns with lip movements.

During our initial experiments, we observed that relying solely on the Lip2Wav-Chemistry dataset was insufficient for training a robust TTS system in English. This limitation could be attributed to two potential factors: Data Quantity and quality (transcript inaccuracy). The Lip2Wav-Chemistry dataset might not contain a sufficient volume of data to fully train the model and capture the complexities of natural speech. There might be inconsistencies between the provided transcripts and the actual audio content. Perfect alignment is crucial for training the model to map text to accurate speech representations. To address these limitations, we incorporate an additional dataset, LJSpeech, into the training process. LJSpeech is a well-established English speech dataset, of a person reading newspaper, known for its high quality and accurate transcripts. By including LJSpeech, we provide the model with additional high-quality training data for TTS.

Multilingual Dataset: To explore the multilingual capabilities of our approach, we acquire a Hindi dataset. Video lectures on Deep Learning from NPTEL channel were downloaded. The videos were clipped to create 5 second long video clips. Audio was extracted and these audio files were then transcribed using Whisper Hindi ASR, an automatic speech recognition tool.

4.3.2 Dataset Preprocessong

The videos are sampled to 25 frames per second (fps). For the video datasets, we employ a preprocessing step to extract lip regions. We leverage the s3fd face detection algorithm to locate faces within the video frames. Once the face is detected, we focus on the mouth area and crop a region centred around the mouth. This cropped region is then used to extract AV-HuBERT features. The extraction process is based on the implementation provided in the fairseq GitHub repository, ensuring compatibility with existing frameworks.

Audio is sampled at 16 kHz, and its HuBERT features are extracted using implementation available in fairseq repository. A pre-trained K-means clustering model is used to cluster the HuBERT features, to get the HuBERT cluster ids.

For English language, the text is phonemized. For the Hindi language, the characters themselves are used as the textual input. Due to high error rate in Hindi data transcripts, we did not train a TTS for this explicitly for the experiments here. The whisper Hindi ASR itself has a WER of 17. Montreal Forced aligner is used to extract phoneme level durations. This is then mapped to the durations corresponding to HuBERT representations and saved.

4.3.3 Evaluation metrics

To objectively assess the quality and effectiveness of our unified model, we employ several established metrics:

- Word Error Rate (WER): This metric provides a quantitative measure of the intelligibility of the generated speech. WER calculates the number of errors made when comparing the generated

word sequence against the ground truth reference transcript. A lower WER indicates better speech intelligibility, signifying that the generated speech closely matches the intended words.

- Speech Quality Assessment: While WER focuses on intelligibility, we also evaluate the overall quality and naturalness of the generated speech. We achieve this by utilizing three common speech quality assessment metrics:

  - Short-Time Objective Intelligibility (STOI): This metric specifically measures the intelligibility of speech. It analyzes the similarity between ground truth and synthesized speech signals, offering insights into how well the generated speech can be understood.
  - Extended Short-Time Objective Intelligibility (ESTOI): Building upon STOI, ESTOI incorporates additional features to provide a more comprehensive assessment of speech intelligibility. It considers factors like spectral distortion and reverberation, offering a more nuanced evaluation of quality.
  - Perceptual Evaluation of Speech Quality (PESQ): This metric aims to simulate human perception of speech quality. PESQ provides a score that reflects how natural and pleasant the generated speech sounds to human listeners by comparing the generated speech to a high-quality reference signal.

By considering both WER and these speech quality metrics, we understand how well our model generates intelligible and quality speech that aligns with the visual cues in the video.

4.4 Results and Discussion

4.4.1 Analysis

Table 4.1: WER evaluations on different combinations of datasets and input combinations.

Language	**Training**	**Inference**	**WER**		
	Dataset	Lang — Dataset	Text-input	Video-input	(text & video) input
English	LJSpeech (23.9h), Chem(9.2h)	Eng. — Chem	**15.55**	54.49	51.77
	LJSpeech (13.56h), Chem(9.2h)	Eng. — Chem	19.17	**51.40**	**42.49**
English and Hindi	LJSpeech (13.56h), Chem (9.2h), Hindi (18.03h)	Eng. — Chem	19.62	56.83	57.51
		Hindi	117.57	71.05	72.04

The presented Table 4.1 provides valuable insights into the performance of our unified speech generation model, as measured by Word Error Rate (WER). We can analyze these results through the lens of data usage and modality combination.

English Language Models:

The first two rows explore the impact of data discrepancy between LJSpeech and Lip2Wav-Chemistry datasets. Training on full LJSpeech data (more text-speech data) leads to better TTS performance compared to using similar durations of text and text-video data. In the first row, where the entire 24 hours of LJSpeceh data and only 9 hours of Chem dataset are used, the model prioritizes learning TTS due to the abundance of textual information. This can be interpreted as a form of overfitting to the text modality. However, the second row presents a more balanced approach, where instead of using the entire LJSpeech dataset, only 13 hours of LJSpeech dataset along with 9 hours of Chem dataset is used for training. By incorporating this, a slight decline in TTS performance (increase in WER) is observed. This is likely due to the model having to learn from a wider variety of inputs, potentially leading to a trade-off between pure TTS and video-based speech generation. Nevertheless, video-based performance (reduced WER for video-only input) improves in this scenario. This highlights the model's ability to leverage visual cues more effectively when trained with a balanced dataset. Furthermore, the significant reduction in WER when introducing text alongside video (compared to video-only) underscores the model's capacity to leverage textual information to enhance the accuracy of speech generation even when aligned with visual cues.

Multilingual Performance:

The inclusion of Hindi for combined training introduces a new language element. This can explain the decline in English results compared to English-only models. The model must adapt to the additional language, potentially leading to a performance decrease compared to a well-trained English model. Moreover, as noted earlier, the absence of Hindi data for text-only training likely contributes to the very poor TTS performance observed for Hindi. This finding suggests that when video is present, the model prioritizes the visual modality even with the alternating training strategy. If explicit text-based training is not conducted for a specific language, the model struggles to perform TTS in that language. This highlights the importance of balanced training data across all modalities for optimal performance in multilingual settings.

These results demonstrate the importance of balanced training data for optimal performance in multimodal TTS systems. Data balance and modality distribution within the training dataset play a crucial role in influencing the model's focus on text-based TTS or video-based speech generation. When video data is present, the model seems to prioritize visual cues over textual information, particularly for languages not extensively trained for text-to-speech generation. Introducing a new language requires additional training to achieve optimal performance. These findings underscore the need for carefully constructed training datasets and potentially exploring language-specific training strategies to maximize the effectiveness of our unified speech generation model across multiple languages. Future work could

explore methods to encourage a more balanced attention mechanism across modalities for improved performance in multilingual and multimodal settings.

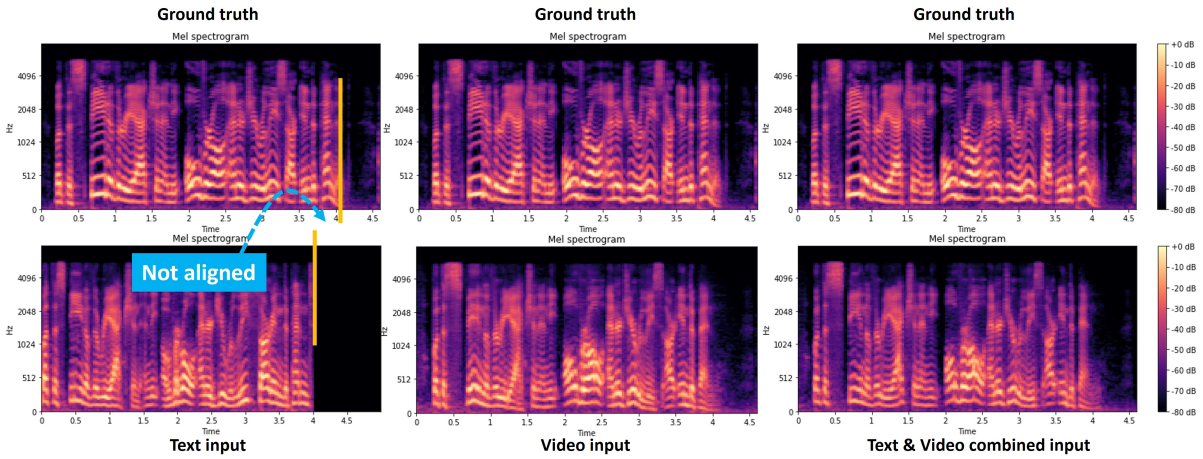

Figure 4.4: Spectrograms of ground-truth and synthesised speech for text, video and for combined speech synthesis

Another point to note is the working of duration predictor module. A spectrogram 4.4 is displayed showcasing that for the same text, without reference video, there is an alignment mismatch between the synthesised speech and ground truth speech, whereas in the case of Lip2Speech or Combined to speech, the ground-truth mel spectrogram and the synthesised speech spectrograms are aligned, as they are video guided synthesis. The model might struggle to predict precise timing based on text alone, for a corresponding speaker style. However, when video information is available (Lip2Speech or Combined mode), the duration prediction becomes more robust. The model can leverage visual cues from the video to estimate phoneme durations more accurately. In the spectrogram, you'll see a closer match between the formants (darker regions) of the synthesized and natural speech, indicating temporally accurate speech generation. This video-guided approach allows the model to synthesize speech that is temporally aligned with the lip movements in the video, creating a more natural and believable experience.

4.4.2 Comparison on Lip2Wav-Chem dataset

This section delves into the comparative analysis of our unified speech synthesis model, OmniSpeak, against existing models. The presented Table 4.2 summarizes the performance across various approaches, including:

- ParrotTTS: A state-of-the-art text-to-speech model,
- Lip2Speech Models: Lip2Wav, RobustL2S and combined text and video-to-speech synthesis model, NeuralDubber.

Table 4.2: Comparison of proposed OmniSpeak results to existing single-modality based speech synthesis models.

Input Modality	**Model**	**STOI↑**	**ESTOI↑**	**PESQ↑**	**WER↓**
Text	Ground truth	1	1	4.548	2.42%
	ParrotTTS	**0.170**	0.005	1.135	**14.79%**
	OmniSpeak (Proposed)	0.168	**0.010**	**1.141**	19.17%
Video	Lip2Wav	0.282	0.176	1.194	72.70%
	RobustL2S	**0.583**	**0.429**	1.120	**32.03%**
	OmniSpeak (Proposed)	0.415	0.229	**1.194**	51.40
Text and Video Combined	NeuralDubber	**0.467**	**0.308**	**1.250**	**18.01%**
	OmniSpeak (Proposed)	0.423	0.232	1.192	42.49

- Text-to-Speech Performance: The results indicate that OmniSpeak achieves comparable performance with existing, dedicated TTS models like ParrotTTS in the text-only domain. This demonstrates that OmniSpeak effectively learns the core functionalities of text-to-speech conversion, generating high-quality speech from textual input.

- Video-to-Speech Performance: When evaluating performance on video-only input, OmniSpeak significantly outperforms 3D-CNN and LSTM-based Lip2Speech models like Lip2Wav. This superiority is evident across the intelligibility-related metrics, highlighting OmniSpeak's ability to accurately map visual cues from videos to corresponding speech. However, RobustL2S, another Lip2Speech model, exhibits slightly better performance than OmniSpeak. Several factors might contribute to this observation. Video Feature Quality: RobustL2S potentially leverages superior video features due to a fine-tuning process not implemented in OmniSpeak. This could lead to more robust visual information extraction and ultimately impact speech generation accuracy. Loss Function: OmniSpeak utilizes HuBERT speech features with a Mean Squared Error (MSE) loss function. Further exploration of loss functions specifically designed for this task could potentially improve OmniSpeak's performance.

- Combined Input Performance: While OmniSpeak demonstrates promising results in both text-only and video-only scenarios, it doesn't yet surpass the current state-of-the-art (SOTA) models in the combined text-and-video input-to-speech domain.

The benchmarking results demonstrate that OmniSpeak effectively performs TTS and video-to-speech models on video-only input. While not yet surpassing the current SOTA in combined input

scenarios, the model exhibits significant potential. By addressing potential limitations through video feature fine-tuning and exploring alternative loss functions, OmniSpeak can be further optimized to become a leading model in the unified speech synthesis domain.

4.5 Conclusion

This work presents the development and evaluation of OmniSpeak, a unified speech generation model. OmniSpeak tackles the challenge of synthesizing speech from various input modalities, including text alone, video alone, or a combination of both. The model achieved comparable performance to existing text-to-speech models in the text-only domain, and lip-to-speech models in the video-only domain, demonstrating its proficiency. While OmniSpeak also achieved promising results with combined text and video input, it did not yet surpass the current state-of-the-art models in this area. Further exploration of video feature fine-tuning techniques and alternative loss functions specifically designed for multimodal training has the potential to bridge this gap. Additionally, investigating the impact of different languages and the amount of training data for each language could lead to enhanced multilingual performance through language-specific training strategies.

OmniSpeak represents a significant advancement in the development of unified speech generation models. Its ability to learn from textual and visual information opens doors to a wide range of applications. By addressing its limitations through further research and development, OmniSpeak has the potential to become a cornerstone technology for generating high-quality, natural-sounding speech across diverse modalities.

Chapter 5

Conclusions and Future Work

This thesis has explored two innovative approaches to speech synthesis. The first part focused on a method for synthesizing speech from lip movements in videos. It leveraged the power of Self-Supervised Learning (SSL) to capture meaningful representations from videos and speech. By employing a transformer architecture, the model effectively mapped the visual features to the corresponding speech information. Furthermore, the utilization of a modified HiFi-GAN vocoder ensured the generated speech maintained high fidelity and naturalness. This approach achieved state-of-the-art results in terms of intelligibility and speech quality, demonstrating its effectiveness in converting visual cues into high-quality audio.

The second part of the thesis ventured into the development of a universal speech synthesis model, with the potential to handle various input modalities. This model aimed at achieving broader versatility, functioning as a Text-To-Speech (TTS) system, a Lip2Speech system, and even a combined (text and lip) to speech system. To achieve this multimodal capability, the model employed Multi-Head Attention (MHA) as a powerful tool for fusing information from different modalities. Additionally, an alternating training strategy was implemented, where the model focused on learning from any of the modalities during each training iteration. While further research is needed to refine the multimodal model, its initial success demonstrates the exciting possibilities for creating a truly flexible and versatile speech synthesis system.

5.1 Future Works

The presented approaches hold immense potential for further exploration. The lip-to-speech synthesis method could be enhanced by incorporating speaker identity information, allowing for the generation of speech with unique characteristics. Additionally, exploring emotional speech synthesis could enable the model to convey a range of emotions. The future of the universal speech synthesis model lies in expanding its capabilities. Delving into more sophisticated multimodal training strategies could significantly improve the model's ability to learn from a combination of text and video input. Language

expansion would allow the model to handle multiple languages effectively. Finally, optimizing the model for real-time speech synthesis would enable exciting applications.

These advancements in speech synthesis pave the way for a more inclusive and interactive world. The potential applications extend to the entertainment industry. This thesis has laid a strong foundation for further exploration, and the future of speech synthesis holds immense promise for the creation of more accessible and interactive technologies.

Related Publications

Accepted Publications

1. **Neha Sahipjohn**, Neil Shah, Vishal Tambrahalli and Vineet Gandhi, "**RobustL2S: Speaker-Specific Lip-to-Speech Synthesis exploiting Self-Supervised Representations**," 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Taipei, Taiwan, 2023, pp. 1492-1499, doi: 10.1109/APSIPAASC58517.2023.10317357.

Other works in MS, not part of thesis:

1. **StethoSpeech: Speech Generation Through a Clinical Stethoscope Attached To The Skin**, **Neha Sahipjohn***, Neil Shah*, Vishal Tambrahalli and Vineet Gandhi (under review)
   **Equal contribution*

2. **ParrotTTS: Text-to-speech synthesis exploiting disentangled self-supervised representations**, EACL Findings, 20024, Neil Shah, Saiteja Kosgi, Vishal Tambrahalli, **Neha Sahipjohn**, Niranjan Pedanekar and Vineet Gandhi

3. Ritu Srivastava, Saiteja Kosgi, Sarath Sivaprasad, **Neha Sahipjohn** and Vineet Gandhi, "**Adversarial Robustness of Mel Based Speaker Recognition Systems**," 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Taipei, Taiwan, 2023, pp. 145-150, doi: 10.1109/APSIPAASC58517.2023.10317404.

Bibliography

[1] Wolfgang Von Kempelen. *Mechanismus der menschlichen Sprache: nebst der Beschreibung seiner sprechenden Maschine...* JV Degen, 1791.

[2] Homer Dudley. "The carrier nature of speech". In: *Bell System Technical Journal* 19.4 (1940), pp. 495–515.

[3] Lawrence Rabiner and Biinghwang Juang. "An introduction to hidden Markov models". In: *ieee assp magazine* 3.1 (1986), pp. 4–16.

[4] Qingqing Zhang, Frank Soong, Yao Qian, Zhijie Yan, Jielin Pan, and Yonghong Yan. "Improved modeling for F0 generation and V/U decision in HMM-based TTS". In: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2010, pp. 4606–4609.

[5] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. "Wavenet: A generative model for raw audio". In: *arXiv preprint arXiv:1609.03499* (2016).

[6] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis". In: *Advances in neural information processing systems* 33 (2020), pp. 17022–17033.

[7] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. "Diffwave: A versatile diffusion model for audio synthesis". In: *arXiv preprint arXiv:2009.09761* (2020).

[8] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 3451–3460.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[11] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. “An audio-visual corpus for speech perception and automatic speech recognition”. In: *The Journal of the Acoustical Society of America* 120.5 (2006), pp. 2421–2424.

[12] Naomi Harte and Eoin Gillen. “TCD-TIMIT: An audio-visual corpus of continuous speech”. In: *IEEE Transactions on Multimedia* 17.5 (2015), pp. 603–615.

[13] Ariel Ephrat and Shmuel Peleg. “Vid2speech: speech reconstruction from silent video”. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2017, pp. 5095–5099.

[14] Ariel Ephrat, Tavi Halperin, and Shmuel Peleg. “Improved speech reconstruction from silent video”. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2017, pp. 455–462.

[15] Hassan Akbari, Himani Arora, Liangliang Cao, and Nima Mesgarani. “Lip2audspec: Speech reconstruction from silent lip movements video”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 2516–2520.

[16] Ravindra Yadav, Ashish Sardana, Vinay P Namboodiri, and Rajesh M Hegde. “Speech prediction in silent videos using variational autoencoders”. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, pp. 7048–7052.

[17] Zheng-Yan Sheng, Yang Ai, and Zhen-Hua Ling. “Zero-shot personalized lip-to-speech synthesis with face image based voice control”. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5.

[18] Konstantinos Vougioukas, Pingchuan Ma, Stavros Petridis, and Maja Pantic. “Video-driven speech reconstruction using generative adversarial networks”. In: *arXiv preprint arXiv:1906.06301* (2019).

[19] Rodrigo Mira, Konstantinos Vougioukas, Pingchuan Ma, Stavros Petridis, Björn W Schuller, and Maja Pantic. “End-to-end video-to-speech synthesis using generative adversarial networks”. In: *IEEE Transactions on Cybernetics* (2022).

[20] Minsu Kim, Joanna Hong, and Yong Man Ro. “Lip to speech synthesis with visual context attentional GAN”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 2758–2770.

[21] Rodrigo Mira, Alexandros Haliassos, Stavros Petridis, Björn W Schuller, and Maja Pantic. “SVTS: scalable video-to-speech synthesis”. In: *arXiv preprint arXiv:2205.02058* (2022).

[22] Leyuan Qu, Cornelius Weber, and Stefan Wermter. “LipSound2: Self-Supervised Pre-Training for Lip-to-Speech Reconstruction and Lip Reading”. In: *IEEE transactions on neural networks and learning systems* (2022).

[23] Minsu Kim, Joanna Hong, and Yong Man 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.

[24] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV 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.

[25] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions”. In: *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2018, pp. 4779–4783.

[26] Jinzheng He, Zhou Zhao, Yi Ren, Jinglin Liu, Baoxing Huai, and Nicholas Yuan. “Flow-Based Unconstrained Lip to Speech Generation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 1. 2022, pp. 843–851.

[27] Durk P Kingma and Prafulla Dhariwal. “Glow: Generative flow with invertible 1x1 convolutions”. In: *Advances in neural information processing systems* 31 (2018).

[28] Munender Varshney, Ravindra Yadav, Vinay P Namboodiri, and Rajesh M Hegde. “Learning Speaker-specific Lip-to-Speech Generation”. In: *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE. 2022, pp. 491–498.

[29] Joanna Hong, Minsu Kim, Se Jin Park, and Yong Man Ro. “Speech reconstruction with reminiscent sound via visual voice memory”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 3654–3667.

[30] Yongqi Wang and Zhou Zhao. “FastLTS: Non-Autoregressive End-to-End Unconstrained Lip-to-Speech Synthesis”. In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 5678–5687.

[31] Yochai Yemini, Aviv Shamsian, Lior Bracha, Sharon Gannot, and Ethan Fetaya. “LipVoicer: Generating Speech from Silent Videos Guided by Lip Reading”. In: *arXiv preprint arXiv:2306.03258* (2023).

[32] Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. “Speech resynthesis from discrete disentangled self-supervised representations”. In: *arXiv preprint arXiv:2104.00355* (2021).

[33] Chenpeng Du, Yiwei Guo, Xie Chen, and Kai Yu. “VQTTS: high-fidelity text-to-speech synthesis with self-supervised VQ acoustic feature”. In: *arXiv preprint arXiv:2204.00768* (2022).

[34] Disong Wang, Shan Yang, Dan Su, Xunying Liu, Dong Yu, and Helen Meng. “VCVTS: Multi-speaker Video-to-Speech synthesis via cross-modal knowledge transfer from voice conversion”. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 7252–7256.

[35] Wei-Ning Hsu, Tal Remez, Bowen Shi, Jacob Donley, and Yossi Adi. “ReVISE: Self-Supervised Speech Resynthesis With Visual Input for Universal and Generalized Speech Regeneration”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 18795–18805.

[36] Keith Ito and Linda Johnson. *The LJ Speech Dataset*. `https://keithito.com/LJ-Speech-Dataset/`. 2017.

[37] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. “Soundstream: An end-to-end neural audio codec”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2021), pp. 495–507.

[38] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. “High fidelity neural audio compression”. In: *arXiv preprint arXiv:2210.13438* (2022).

[39] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. “wav2vec 2.0: A framework for self-supervised learning of speech representations”. In: *Advances in neural information processing systems* 33 (2020), pp. 12449–12460.

[40] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. “Neural codec language models are zero-shot text to speech synthesizers.(2023)”. In: *arXiv preprint arXiv:2301.02111* (2023).

[41] Neil Shah, Saiteja Kosgi, Vishal Tambrahalli, Neha Sahipjohn, Anil Kumar Nelakanti, and Vineet Gandhi. “ParrotTTS: Text-to-speech synthesis exploiting disentangled self-supervised representations”. In: (2024).

[42] Mateusz Lajszczak, Guillermo Cambara Ruiz, Yang Li, Fatih Beyhan, Arent van Korlaar, Fan Yang, Arnaud Joly, Álvaro Martín Cortinas, Ammar Abbas, Adam Michalski, Alexis Moinet, Sri Karlapati, Ewa Muszynska, Haohan Guo, Bartosz Putrycz, Soledad López Gambino, Kayeon Yoo, Elena Sokolova, and Thomas Drugman. “BASE TTS: Lessons from building a billion-parameter text-to-speech model on 100K hours of data”. In: *arXiv* (Unknown). URL: https://www.amazon.science/publications/base-tts-lessons-from-building-a-billion-parameter-text-to-speech-model-on-100k-hours-of-data.

[43] Michael Hassid, Michelle Tadmor Ramanovich, Brendan Shillingford, Miaosen Wang, Ye Jia, and Tal Remez. “More than words: In-the-wild visually-driven prosody for text-to-speech”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10587–10597.

[44] Chenxu Hu, Qiao Tian, Tingle Li, Wang Yuping, Yuxuan Wang, and Hang Zhao. “Neural dubber: Dubbing for videos according to scripts”. In: *Advances in neural information processing systems* 34 (2021), pp. 16582–16595.

[45] Gaoxiang Cong, Liang Li, Yuankai Qi, Zheng-Jun Zha, Qi Wu, Wenyu Wang, Bin Jiang, Ming-Hsuan Yang, and Qingming Huang. “Learning to dub movies via hierarchical prosody models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 14687–14697.

[46] Zixiong Su, Shitao Fang, and Jun Rekimoto. “LipLearner: Customizable Silent Speech Interactions on Mobile Devices”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–21.

[47] Randa El-Bialy, Daqing Chen, Souheil Fenghour, Walid Hussein, Perry Xiao, Omar H Karam, and Bo Li. "Developing phoneme-based lip-reading sentences system for silent speech recognition". In: *CAAI Transactions on Intelligence Technology* 8.1 (2023), pp. 129–138.

[48] Hang Zhou, Ziwei Liu, Xudong Xu, Ping Luo, and Xiaogang Wang. "Vision-infused deep audio inpainting". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 283–292.

[49] William H Sumby and Irwin Pollack. "Visual contribution to speech intelligibility in noise". In: *The journal of the acoustical society of america* 26.2 (1954), pp. 212–215.

[50] Lynne E Bernstein, Nicole Jordan, Edward T Auer, and Silvio P Eberhardt. "Lipreading: A review of its continuing importance for speech recognition with an acquired hearing loss and possibilities for effective training". In: *American Journal of Audiology* 31.2 (2022), pp. 453–469.

[51] Jeongsoo Choi, Minsu Kim, and Yong Man Ro. "Intelligible Lip-to-Speech Synthesis with Speech Units". In: *arXiv preprint arXiv:2305.19603* (2023).

[52] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. "Learning audio-visual speech representation by masked multimodal cluster prediction". In: *arXiv preprint arXiv:2201.02184* (2022).

[53] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[54] Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Georgios Tzimiropoulos, and Maja Pantic. "Audio-visual speech recognition with a hybrid ctc/attention architecture". In: *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE. 2018, pp. 513–520.

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[56] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. "Fastspeech 2: Fast and high-quality end-to-end text to speech". In: *arXiv preprint arXiv:2006.04558* (2020).

[57] Daniel Michelsanti, Olga Slizovskaia, Gloria Haro, Emilia Gómez, Zheng-Hua Tan, and Jesper Jensen. "Vocoder-based speech synthesis from silent videos". In: *arXiv preprint arXiv:2004.02541* (2020).

[58] Adrian Bulat and Georgios Tzimiropoulos. "How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks)". In: *International Conference on Computer Vision*. 2017.

[59] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. "Robust speech recognition via large-scale weak supervision". In: *OpenAI Blog* (2022).

[60] Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, and Wei-Ning Hsu. "Direct speech-to-speech translation with discrete units". In: *arXiv preprint arXiv:2107.05604* (2021).

[61] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. "An algorithm for intelligibility prediction of time–frequency weighted noisy speech". In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.7 (2011), pp. 2125–2136.

[62] Jesper Jensen and Cees H Taal. "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.11 (2016), pp. 2009–2022.

[63] Minsu Kim, Joanna Hong, Se Jin Park, and Yong Man Ro. "Multi-modality associative bridging through memory: Speech sound recollected from face video". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 296–306.

[64] Douglas Burnham, Ruth Campbell, G Away, and BJ Dodd. *Hearing eye II: the psychology of speechreading and auditory-visual speech*. Routledge, 2013.