Visual Speech Recognition For Seamless Communication With Hearing Immpared Persons

A Report Submitted in
Partial Fulfilment of the Requirements for the Degree of
Master of Technology

by

Md. Laraib Ahmad
Registration No. 2322207

Under the Supervision of

Dr. Debbrota Paul Chowdhury

Computer Science & Engineering Department
NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR
May, 2025

© NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR, MAY, 2025
ALL RIGHTS RESERVED

COMPUTER SCIENCE & ENGINEERING DEPARTMENT
NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR
(An Institute of National Importance)
SILCHAR, ASSAM, INDIA – 788010
Fax: (03842) 224797 **Website:** http://www.nits.ac.in

Declaration

Thesis Title: **Visual Speech Recognition For Seamless Communication With Hearing Immpaired Persons**
Degree for which the Thesis is submitted: **Master of Technology**

I declare that the presented thesis represents largely my own ideas and work in my own words. Where others ideas or words have been included, I have adequately cited and listed in the reference materials. The thesis has been prepared without resorting to plagiarism. I have adhered to all principles of academic honesty and integrity. No falsified or fabricated data have been presented in the thesis. I understand that any violation of the above will cause for disciplinary action by the Institute, including revoking the conferred degree, if conferred, and can also evoke penal action from the sources which have not been properly cited or from whom proper permission has not been taken.

Signed: _____

Date: _____

Md. Laraib Ahmad
2322207

COMPUTER SCIENCE & ENGINEERING DEPARTMENT
NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR
(An Institute of National Importance)
SILCHAR, ASSAM, INDIA – 788010
Fax: (03842) 224797 **Website:** http://www.nits.ac.in

Certificate

It is certified that the work contained in this thesis entitled "**Visual Speech Recognition For Seamless Communication With Hearing Immpaired Persons**" submitted by **Md. Laraib Ahmad**, Registration no **2322207** for the M.Tech is absolutely based on his own work carried out under my supervision.

Place:
Date:

Dr. Debbrota Paul Chowdhury
Computer Science & Engineering
National Institute of Technology Silchar

Abstract

Visual Speech Recognition (VSR) is an innovative approach aimed at enabling seamless communication with hearing-impaired individuals. This project focuses on lip reading as a primary tool for understanding spoken words without relying on auditory cues. The developed methodology integrates lip localization, feature extraction, and Euclidean distance-based prediction to recognize spoken words accurately. A custom dataset, consisting of recorded and curated videos, serves as the foundation for evaluating the proposed algorithm. Achieving over 78.33% accuracy in word prediction, the system minimizes computational overhead by using deep learning solely for lip feature extraction, while prediction relies on a custom-designed algorithm. This research highlights the potential of VSR in accessibility, particularly for hearing-impaired individuals, and sets the stage for advancements through accent-independent datasets, 3D lip feature analysis, and deep learning comparisons. Future work includes enhancing the model to tackle homophones and expanding its applicability across diverse datasets and real-world conditions.

Acknowledgements

I take this opportunity to express my sincere gratitude and heartily thank my supervisor **Dr. Debbrota Paul Chowdhury**, Computer Science and Engineering, National Institute of Technology, Silchar, for his continuous inspiration and valuable guidance at every stage of my research work. I have learned so much from him. Whatever I know about Lip Reading is just because of his encouragement. I do consider myself lucky for working under such a helpful gentle person.

Date :

Md. Laraib Ahmad
Reg. No. 2322207

Contents

List of Figures

CHAPTER 1

Introduction

FIGURE 1.1: Lip Reading

Lip reading is a technique for comprehending speech by visually understanding the motions of the lips. Lip reading is a difficult skill for humans. To anticipate spoken words, it is necessary to know the underlying language and visual cues. To interpret spoken words, experts must have a particular amount of expertise and awareness of visual expressions. It becomes more difficult for a machine due to the different accents used for the same word. For individuals who are deaf or hard-of-hearing, lip reading can be a primary method of understanding spoken language, especially when they do not have access to or choose not to use sign language. Accurate lip reading can significantly enhance the effectiveness of hearing aids and cochlear implants, helping users to understand better speech in noisy environments or when the audio signal is weak. laryngeal cancer can result in loss of speech or changes in the voice, especially if the cancer or its treatment affects the vocal cords or requires the removal

of parts of the larynx. In such a case, the speaker can only move their lips to communicate. Visual speech recognition has a wide range of applications across various fields such as hearing impairment assistance, forensic analysis, silent communication, dubbing animation, noise-rich environments, cross-language communication, and cognitive science etc. With recent advancements in the area of deep learning, the translation of lip sequences into meaningful words has become more accurate. Despite significant progress, visual speech recognition faces challenges such as variability in lip movements [1], diverse speaking styles [2], and the need for large and labeled datasets for training [3]. Ongoing research aims to address these challenges and further refine technology, making it more accurate, reliable, and widely applicable. This work looks at the advancement of lip identification, to create a dataset of different accents, and to develop an algorithm that can recognize words correctly, independent of accents.

1.1 Why Lip Reading is Important?

1.1.1 Primary communication tool:

FIGURE 1.2: sign language

For individuals who are deaf or hard-of-hearing, lip reading is often a vital communication strategy, enabling them to comprehend spoken language by visually interpreting the movements of the lips, tongue, and facial expressions of the speaker. This method becomes especially important in situations where access to sign language interpreters is unavailable or when sign language is not commonly used by the surrounding community. Unlike sign language, which requires both the speaker and listener to be familiar with the same signing system, lip reading allows individuals to directly engage with

spoken conversations, bridging communication gaps in diverse settings such as social interactions, educational environments, and professional scenarios. By relying on visual speech cues, lip reading empowers individuals to participate in conversations and fosters inclusivity, enhancing their ability to connect with others and navigate the hearing world.

1.1.2 Complement to hearing aids and cochlear implants

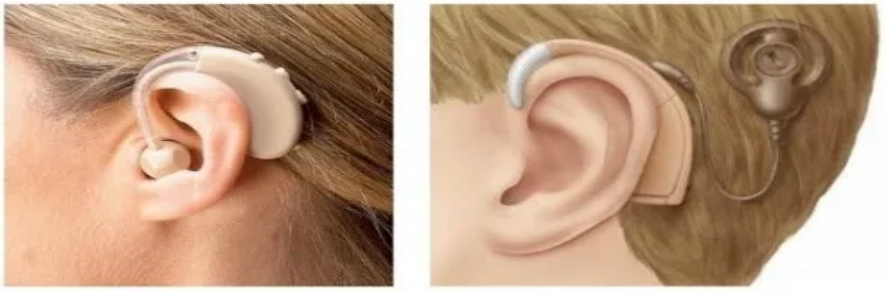

FIGURE 1.3: cochlear implants

Lip reading is a useful addition to cochlear implants and hearing aids, increasing their efficacy in a range of communication situations. Even though these assistive devices directly stimulate the auditory nerve or enhance auditory signals, they frequently have trouble working in settings with a lot of background noise, conversations that overlap, or poor audio quality. Even with the aid of technology, people may still have trouble telling spoken words apart in these circumstances.

Users can close the gap between speech's visual and aural components by using lip reading. Together with the increased audio information, the visual clues from lip movements offer further context to aid with word and sentence decoding. This dual-channel technique works especially well in situations where clarity is impaired, such busy areas, windy outdoor settings, or during

Incorporating lip reading with cochlear implants and hearing aids ultimately improves communication skills and promotes independence and inclusivity in a variety of social and professional contexts.

1.2 Laryngeal cancer

FIGURE 1.4: Laryngeal cancer

Laryngeal cancer, which impacts the tissues of the voice box or larynx, can greatly affect a person's vocal abilities. This illness, especially when it involves the vocal cords, often results in alterations to voice quality, such as hoarseness, diminished volume, or even a total loss of the ability to speak. In more advanced cases, treatment may require the surgical removal of some or all of the larynx, a procedure referred to as laryngectomy. This can permanently eliminate the individual's capacity to produce natural speech.

Following these interventions, people face the difficulty of discovering different ways to communicate. Often, they depend on their lip movements as a key method of expression. By silently shaping words with their lips, they allow others to understand their intended speech through lip reading. This technique is crucial for communication, particularly for those who do not utilize sophisticated speech prosthetics or assistive technologies such as electronic voice devices.

Reading lips is not only useful but also gives those who have experienced such significant treatments the ability to retain their independence and participate in conversations with family, friends, and coworkers. Additionally, visual speech recognition technologies, which can decode lip movements and convert them into text or synthetic voice, present a promising opportunity to advance communication for individuals impacted

by laryngeal cancer. These innovations offer hope for improved accessibility, aiding individuals in navigating the challenges brought by the loss of their natural speech.1.4.

1.3 Reasearch Advancement

1.3.1 Early 20th Century

The early 20th century was a significant era for establishing lip reading as a skill, especially for the deaf and hard-of-hearing communities [4]. During this period, major advancements were made in comprehending and teaching the fundamentals of lip reading, propelled by the larger movement aimed at including individuals with hearing impairments in society. Schools focused on educating the deaf started to integrate lip-reading courses into their offerings, acknowledging its potential as an effective means of communication.

These educational institutions sought to empower learners by showing them how to visually understand spoken language through observing the movement of the speaker's lips, facial cues, and various non-verbal signals. This method was frequently paired with speech training, creating a combination of skills that improved communication capabilities. Furthermore, innovations in teaching methods during this period set the stage for the creation of specialized training resources and techniques, many of which continue to impact contemporary deaf education.

This time also brought greater awareness of the importance of accessible communication, encouraging researchers and educators to enhance their comprehension of the cognitive and physiological elements of lip reading. These initiatives not only elevated the quality of teaching but also played a role in the wider societal acknowledgment of the abilities and rights of those with hearing difficulties.

1.3.2 1970s

The 1970s saw the beginning of automated lip reading as a research domain, fueled by progress in computer science and an increasing curiosity in human-computer interaction. Investigators started to examine the visual components of speech, with the goal of comprehending how the motions of the lips, mouth, and adjacent facial areas could be systematically recorded, processed, and analyzed by machines[5]. This era witnessed the establishment of essential theories and techniques for visual speech recognition, drawing inspiration from the human capability to read lips.

The initial approaches were mainly theoretical, emphasizing the modeling of how speech is articulated and finding methods to extract significant features from video recordings. Researchers employed fundamental image processing techniques to capture and monitor lip movements, frequently relying on hand-annotated datasets because of the absence of advanced automation tools. The goal of these studies was to uncover patterns in the forms, locations, and movements of the lips that relate to particular phonemes or sounds produced in speech.

The difficulties of the time were considerable, considering the restricted computing capabilities and the early development of machine learning techniques. Nevertheless, these initial attempts established a foundation for future advancements by showing that visual data could be used to enhance or even substitute for auditory signals in speech recognition systems. The study also highlighted the promising uses of automated lip reading, including support for individuals with hearing loss and improving speech recognition technology in noisy settings.

1.3.3 1980s and 1990s

The advancements in the 1980s and 1990s significantly changed the landscape of automated lip reading, thanks to improvements in computer vision, pattern recognition, and artificial intelligence that broadened the field's capabilities. Expanding upon the groundwork laid in the 1970s, scientists created systems capable of examining visual aspects of speech with enhanced precision by utilizing the growing computational power and innovative algorithms[6].

During the 1980s, the emphasis was on analyzing and monitoring lip movements, achieving important breakthroughs such as feature extraction methods including edge detection, contour modeling, and motion analysis. The goal of these techniques was to correlate lip shapes and movements with particular phonemes, and they were frequently executed on dedicated hardware due to the constraints of general-purpose computing capabilities.

In the 1990s, statistical frameworks like Hidden Markov Models (HMMs) became crucial for identifying temporal patterns in speech, aided by the availability of larger datasets and advancements in machine learning methods. Additionally, multimodal strategies became more popular, combining visual and auditory data to achieve stronger results in challenging acoustic settings.

During this time, practical uses emerged, such as assisting those with hearing disabilities, improving interactions between humans and computers, and advancing surveillance techniques. This era solidified automated lip reading as an important area of research, setting the stage for significant advancements in deep learning in the 21st century[7].

1.3.4 2000s

The 2000s saw notable advancements in automated lip reading, propelled by improvements in AI and deep learning technologies. Researchers expanded on statistical models such as Hidden Markov Models (HMMs).[8] to examine the time-based patterns of visual speech, frequently employing characteristics like lip shapes, optical movement, and shape-oriented descriptions. These techniques were enhanced by the access to more extensive datasets and advancements in computational technology, allowing for more efficient training.

The latter part of the decade witnessed the rise of deep learning as a groundbreaking method. Convolutional Neural Networks (CNNs) were modified for visual speech recognition, enabling the automatic extraction of features from unprocessed video frames. When paired with Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) networks, these models were able to grasp both spatial and temporal

dynamics of lip movements. Multimodal strategies also became popular, combining auditory and visual data to enhance recognition in difficult situations. By the decade's conclusion, these innovations had notably improved the precision and usefulness of automated lip-reading technologies, paving the way for upcoming advancements.

1.3.5 2010s

The decade of the 2010s represented a significant turning point for automated lip reading, fueled by advanced deep learning methods including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) [1]. Convolutional Neural Networks (CNNs) are particularly effective at identifying spatial features in video frames, adeptly detecting detailed lip movements. Meanwhile, Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM) networks, focus on modeling the temporal patterns present in speech sequences. The combination of these technologies resulted in end-to-end architectures that greatly enhanced the accuracy of lip-reading.

The access to extensive datasets like GRID, TCD-TIMIT, and LRW has been essential for training deep learning models. These datasets provided a variety of speakers, accents, and settings, which improved the generalizability of the systems. The rise in computational power from GPUs and cloud computing also significantly sped up advancements.

Applications grew to include assistive technologies for individuals with hearing loss, silent speech communication systems, and security solutions. By the end of the decade, advances in deep learning had made it possible to create effective lip-reading systems that function reliably in various real-world situations.

1.3.6 2016

In 2016, Google DeepMind introduced LipNet [1], a pioneering model that demonstrated remarkable precision in automated lip reading by utilizing deep learning methods. LipNet marked a substantial shift from conventional techniques, being the first

comprehensive deep learning model tailored for sentence-level lip reading, as opposed to concentrating on individual words or phonemes.

LipNet integrated Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) into a comprehensive framework. CNNs served to extract spatial characteristics from sequences of video frames, effectively capturing the detailed movements of the lips. These characteristics were subsequently fed into an RNN based on Gated Recurrent Units (GRUs), which modeled the temporal dynamics and interrelations among frames, successfully learning how lip movements developed over time.

One of the main breakthroughs of LipNet was its implementation of the Connectionist Temporal Classification (CTC) loss function, enabling the model to directly match lip movements with textual sequences without the need for frame-by-frame annotations. This approach streamlined the training process and removed the reliance on time-consuming manual labeling.

Based on datasets such as GRID, LipNet reached new heights in accuracy for lip reading at the sentence level, surpassing conventional statistical models and previous machine learning techniques. The model showed proficiency in understanding continuous speech, even under difficult circumstances, representing an important achievement in the discipline.

LipNet's achievements underscored the possibilities of end-to-end deep learning in the field of lip reading and motivated additional studies into more sophisticated architectures and uses, establishing a new benchmark for visual speech recognition technologies.

1.3.7 2020s

In the 2020s, advancements in lip reading technologies have greatly enhanced their precision, effectiveness, and practical use in real-world situations [3]. The incorporation of state-of-the-art deep learning methods, such as Transformer architectures and multimodal strategies, has improved the resilience of visual speech recognition systems, allowing them to better suit various environments and difficult conditions.

Lip reading technologies are becoming more prevalent in assistive devices, offering immediate transcription for individuals who are deaf or hard-of-hearing, and enhancing

communication in various fields like education, healthcare, and public services. In the realm of security, lip reading is utilized for discreet surveillance and forensic examination, allowing observation in situations where capturing audio is not feasible or acceptable.

In the field of human-computer interaction (HCI), lip reading technologies are incorporated into voice-activated applications, enabling users to communicate with devices in loud or busy settings. Additionally, these systems enhance privacy by facilitating speech recognition silently, which is especially beneficial in public or sensitive areas.

Additionally, studies are concentrating on improving the effectiveness of lip reading models in practical situations, such as training with a broader range of datasets to enhance recognition across multiple languages and different accents. The ongoing development towards instantaneous lip reading and better adaptability to various lighting conditions, backgrounds, and speaker traits is propelling the discipline ahead. Consequently, lip reading technologies are being adopted more widely across various sectors, showing considerable promise for enhancing accessibility, security, and communication.

1.4 Motivation

1.4.1 Accessibility for Hearing-Impaired Individuals

Reading lips is an essential resource for individuals with hearing disabilities, allowing them to understand spoken words by watching the movements of the speaker's lips. The objective of this research is to enhance the precision and effectiveness of automated lip-reading systems, thereby promoting improved communication across various environments, including classrooms, workplaces, and social interactions. Advanced lip-reading technology can create a more inclusive world for individuals who depend on visual signals for communication.

1.4.2 Speech Enhancement in Noisy Environments:

In locations where there is considerable background noise (such as busy public areas, industrial settings, or airports), standard audio-based speech recognition can have difficulty accurately detecting speech. Lip reading presents a possible solution by leveraging visual indicators to augment or improve the audio input. By combining lip reading with current audio technologies, we can develop more effective speech enhancement systems, enhancing communication and speech recognition in difficult environments.

1.4.3 Multimodal Systems:

Lip reading can be integrated with various forms of communication, including audio, gestures, and facial expressions, to develop multimodal systems. These systems can utilize diverse sources of information to enhance comprehension and interaction in human-computer interfaces. By incorporating lip reading into multimodal systems, we can improve the precision, dependability, and overall user experience of applications in fields like virtual assistants, augmented reality, and human-robot interaction, particularly in noisy or ever-changing environments.

1.4.4 Advances in AI and Machine Learning:

The swift progress in AI and machine learning, especially in deep learning, has resulted in major advances in automated lip reading. Cutting-edge algorithms are now capable of examining visual speech data with enhanced accuracy, paving the way for real-time lip reading systems that operate efficiently in intricate, uncontrolled settings. These improvements allow the application of lip reading across various fields, from assistive technologies to security and surveillance, broadening its possible influence.

1.5 Problem Statement

Although substantial progress has been made in visual speech recognition, numerous obstacles still exist that limit its broad usage and efficiency. These obstacles consist of:

1.5.1 Variability in Lip Movements

Lip movements can differ greatly among individuals because of variations in facial anatomy, speech styles, and articulation. Additionally, a single person might display different lip movements influenced by factors such as the speed of speech, emotional state, and context. This inconsistency presents challenges for lip-reading algorithms in accurately interpreting speech across various speakers and scenarios. Creating systems that can adapt to a wide range of lip shapes and movements continues to be a significant obstacle.

1.5.2 Diverse Speaking Styles

People speak in various styles, accents, and dialects, and these factors influence lip movements. For example, individuals with different regional accents or those who speak rapidly or slowly may display subtle but significant differences in how they form sounds, making it harder for visual speech recognition systems to accurately decode speech. The ability to effectively handle such diversity in speaking styles is crucial for the robustness of lip-reading systems.

1.5.3 Need for Large and Labeled Datasets

To train deep learning models effectively, vast amounts of labeled data are needed, particularly for tasks like visual speech recognition. Creating large datasets that capture a wide range of speakers, speaking styles, and environmental conditions is resource-intensive and time-consuming. Moreover, labeling these datasets with accurate transcriptions can be labor-intensive and expensive, limiting the availability of high-quality datasets that are crucial for improving model performance.

Ongoing research is actively working to address these challenges. Approaches such as using synthetic data, improving transfer learning techniques, and developing more sophisticated algorithms for feature extraction are being explored. Additionally, efforts are being made to create more diverse and larger datasets that better capture the variability in lip movements and speaking styles. These advancements aim to make visual speech recognition systems more accurate, adaptable, and reliable across different contexts, ultimately expanding their applicability in real-world scenarios, from assistive technologies to surveillance and human-computer interaction.

1.6 Objectives of the Thesis

The objectives of the thesis are :

i. To create a dataset of different accents.

ii. To develop an algorithm that can recognize words correctly independent of accents.

1.7 Organization of the Thesis

The main focus of this thesis is to develop and refine a lip reading system for predicting spoken words based on visual features of lip movements. The thesis is organized into six chapters, as outlined below, with a brief explanation of their contents.

Chapter 1: Introduction This chapter provides an overview of the field of lip reading, focusing on its importance and potential applications, particularly in speech recognition and assistive technologies. The chapter discusses the challenges of visual speech recognition, including variability in lip movements, diverse speaking styles, and the need for large annotated datasets. It also introduces the key research questions and objectives addressed in the thesis.

Chapter 2: Literature Review This chapter offers a detailed survey of state-of-the-art techniques and methods in automated lip reading. It covers the evolution of visual speech recognition from early approaches to recent deep learning advancements,

highlighting key milestones and challenges in the field. The chapter also reviews different datasets, feature extraction methods, and the application of machine learning techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs).

Chapter 3: Feature Extraction In this chapter, we discuss the feature extraction process used in lip reading systems. It focuses on the key visual features, such as lip shapes, movements, and facial landmarks, that are essential for recognizing spoken words. The chapter emphasizes the role of feature extraction methods like Local Binary Patterns (LBP) and their use in capturing facial movements, making the system invariant to changes in lighting and orientation.

Chapter 4: Proposed System This chapter introduces the proposed lip reading system, explaining the architecture and the steps involved in the prediction of spoken words from visual lip movements. The system is divided into stages: the initial stage captures the features from video frames, while the subsequent stages involve matching these features to predefined word signatures. We also discuss the use of adaptive techniques, such as thresholding and dynamic adjustments, to handle variations in lighting and speaking styles.

Chapter 5: Experimental Results and Discussion This chapter provides details about the experimental setup, including the datasets used for training and testing the system. We present the evaluation metrics and performance results, discussing the effectiveness of the proposed system in recognizing words from lip movements. The chapter also examines the impact of factors like illumination changes and variations in speaker styles, analyzing how the system performs under different conditions.

Chapter 6: Conclusion In this chapter, we summarize the findings of the thesis, highlighting the contributions made towards improving automated lip reading systems. The chapter concludes by discussing the limitations of the current approach and suggests directions for future research, including potential improvements in model accuracy, dataset expansion, and real-time lip reading applications.

CHAPTER 2

Literature Review

2.1 Sentence-level prediction using spatiotemporal convolutions, recurrent neural networks (RNNs), and a connectionist temporal classification (CTC) Loss.

The paper titled [1] "LipNet: End-to-End Sentence-Level Lipreading" introduces LipNet, a deep learning model designed to improve the task of lipreading by predicting entire sentences from sequences of video frames. Traditional methods for lipreading separated feature extraction and sequence prediction, focusing mostly on word classification. LipNet, however, is an end-to-end model capable of sentence-level prediction using spatiotemporal convolutions, recurrent neural networks (RNNs), and a connectionist temporal classification (CTC) loss, enabling it to learn both spatiotemporal features and sequential patterns without requiring manual frame-to-text alignment.

Key contributions of the paper include:

1. End-to-End Model: LipNet is the first model to make full sentence-level predictions from lip movements, rather than individual word recognition.

2. Performance: LipNet outperforms previous state-of-the-art models, achieving a 95.2% accuracy on the GRID corpus, a dataset of sentence-level lipreading. It also surpasses human lipreaders significantly, demonstrating a word error rate (WER) of 4.8% on overlapping speakers.

3. Model Architecture: LipNet uses a combination of spatiotemporal convolutions and bidirectional gated recurrent units (Bi-GRUs) for processing video input, followed by a softmax output layer with CTC for predicting text sequences.

4. Human Comparison: In experiments, LipNet showed far superior performance compared to experienced human lipreaders.

5. Generalization: The model was able to generalize well to unseen speakers, showcasing its robustness across different individuals.

The paper also provides a thorough analysis of learned visual features and phonetic confusions (visemes), enhancing the understanding of how the model interprets speech visually. This work demonstrates the potential of end-to-end neural networks in automating complex tasks like lipreading and suggests that further improvements can be made with more data and advanced architectures.

2.2 VSR models to recognize words and phrases that were not part of the original training set

The paper titled[9] "Harnessing GANs for Zero-Shot Learning of New Classes in Visual Speech Recognition" presents a novel approach to overcoming the challenge of limited training data in Visual Speech Recognition (VSR) using Generative Adversarial Networks (GANs). VSR, also known as automated lip reading, involves recognizing speech by interpreting lip movements from video.

Key Contributions:

1. Zero-Shot Learning with GANs: The paper introduces a method to generate training data for unseen utterances using Temporal Conditional GANs (TC-GANs). This allows VSR models to recognize words and phrases that were not part of the original training set, improving prediction accuracy by 27%.

2. Data Augmentation: By generating synthetic lip movement videos for both seen and unseen phrases, the model significantly improves accuracy for unseen classes (out-of-vocabulary utterances). The synthetic data enhances the model's ability to generalize, even for new languages (e.g., generating Hindi phrases using English-trained models).

3. Cold-Start Problem: The paper addresses the cold-start issue in lip-reading, where no prior data is available for training. The GAN-based approach shows success in generating realistic videos from audio inputs without any prior training data, leading to a large increase in accuracy over traditional methods.

4. Language-Agnostic GAN Model: The GAN model can generate lip movements for different languages, showcasing the method's flexibility and generalization across language boundaries. The paper demonstrates this by generating realistic Hindi lip movements using English data.

5. Methodology used in this paper

   (a) Temporal Conditional GAN (TC-GAN): The GAN is conditioned on both an audio input and a still image of the speaker, which generates a sequence of video frames representing lip movements. This approach allows the model to interpolate between visemes (basic visual units in speech), ensuring smooth and continuous lip movements.

   (b) Viseme-Concatenation Approach: In addition to TC-GAN, the authors also explore a viseme-concatenation approach, where pre-annotated visemes are stitched together to generate synthetic videos. However, this approach performs worse compared to the TC-GAN.

6. Experiments: The paper evaluates the proposed GAN-based approach by training VSR models on synthetic data generated for unseen classes. The experiments show that the model trained with GAN-generated videos outperforms the baseline, especially for unseen phrases and new languages.

7. Results: The VSR model trained with TC-GAN-generated data achieves higher accuracy (up to 27%) compared to traditional models trained only on available data. In a cold-start scenario where 100% of the test classes are unseen, the GAN-based model still outperforms traditional approaches. The model demonstrates robust performance for both seen and unseen classes, with negligible degradation in accuracy for seen classes.

8. Conclusion: The paper demonstrates that GANs can effectively be used for zero-shot learning in VSR by generating realistic lip movement videos for unseen utterances. This method provides a significant boost in accuracy and can be extended to new languages with little adaptation. The authors suggest that future work could explore extending this approach to continuous VSR and other phonetically distant languages. Overall, the research shows promising advancements in the field of automated lip reading, especially for resource-constrained scenarios where training data is scarce.

2.3 Viseme-based Classification

The paper titled [10]"Lip Reading Sentences Using Deep Learning With Only Visual Cues" presents a neural network-based system designed for lip reading using purely visual cues (without audio). The proposed system aims to improve the accuracy of lip reading for sentences, which is more challenging than word-based lip reading. Key contributions of the paper include:

1. Viseme-based Classification: The system uses visemes (visual representations of phonemes) instead of words or ASCII characters to reduce the number of classes and allow for better generalization to unseen words.

2. Deep Learning Model: A specially designed transformer model is used to classify visemes and convert them into words using perplexity analysis, which improves the accuracy of sentence-level lip reading.

3. Robustness to Lighting Variations: The system performs well under varying lighting conditions, demonstrating robustness in different environments.

4. Performance Improvement: The system achieves a 15% lower word error rate compared to state-of-the-art models, as validated on the challenging BBC LRS2 dataset.

5. Major Challenges Addressed:

   (a) Word-based systems are limited to words seen during training.

   (b) The ambiguity of visemes (one viseme can represent multiple phonemes) is managed by perplexity analysis to convert visemes into words.

   (c) The system shows improved accuracy in sentence prediction with robust handling of noisy or varied lighting conditions.

   The paper emphasizes the importance of viseme-to-word conversion in sentence-level lip reading and explores future enhancements in this area

2.4 multiple data augmentation techniques, including speed perturbation, rotation, flipping, and color transformation.

The paper [11]"The NPU-ASLP-LiAuto System Description for Visual Speech Recognition in CNVSRC 2023" describes the visual speech recognition (VSR) system developed by the NPU-ASLP-LiAuto team for the Chinese Continuous Visual Speech Recognition Challenge (CNVSRC) 2023. The system was applied to both the Single-Speaker and Multi-Speaker VSR Tasks in the challenge. The team leveraged lip motion extraction and multiple data augmentation techniques, including speed perturbation, rotation, flipping, and color transformation.

1. Data Processing: Lip motion videos were extracted at multiple scales (48, 64, 96, and 112 pixel sizes) from the CNVSRC dataset. Speed perturbation was applied to augment the training data.

2. VSR Model Architecture: The system uses an end-to-end architecture with a joint CTC (Connectionist Temporal Classification) and attention loss. - The

model includes a ResNet3D visual frontend for feature extraction and an E-Branchformer encoder. The decoder is a standard Transformer-based module.

3. Multi-System Fusion: The team developed multiple systems using different encoders (E-Branchformer, Branchformer, and Conformer) to facilitate multi-system fusion. Post-fusion was achieved using ROVER (Recognizer Output Voting Error Reduction).

   Single-Speaker VSR Task: The system achieved 34.76% Character Error Rate (CER) in the evaluation set.Multi-Speaker VSR Task: The system achieved 41.06% CER in the evaluation set.The system ranked first place in all three tracks it participated in, demonstrating its strong performance across tasks.

   The NPU-ASLP-LiAuto team's VSR system achieved state-of-the-art results in the CNVSRC 2023 by leveraging multi-scale data processing, a robust deep learning architecture, and multi-system fusion, marking significant progress in visual speech recognition.

2.5 Time Masking Technique:

The paper [12] "Visual Speech Recognition for Multiple Languages in the Wild" presents a novel approach to Visual Speech Recognition (VSR), also known as lipreading, focusing on recognizing speech solely from lip movements without audio input. The authors aim to demonstrate that careful model design, rather than just increasing training data size, can lead to significant performance improvements.

1. Model Design: The paper introduces prediction-based auxiliary tasks to a VSR model, which jointly predicts both audio and visual features. The authors also optimize hyperparameters and introduce time-masking, a temporal augmentation technique commonly used in Automatic Speech Recognition (ASR) models. Time-masking helps the model rely more on contextual information, improving its ability to distinguish between similar lip movements.

2. Data Augmentation and Optimization: The model uses various data augmentations, such as random cropping and image flipping, and is trained using publicly

available datasets (LRS2, LRS3). Despite using smaller datasets than previous methods, the authors achieved state-of-the-art performance.

3. Multi-Language Evaluation: Unlike most previous works that focus solely on English, the model is tested on multiple languages, including Spanish, Mandarin, Italian, French, and Portuguese, achieving state-of-the-art results across these languages.

4. Efficiency with Smaller Datasets: The model achieves superior performance compared to models trained on much larger non-public datasets, demonstrating that model architecture and optimization can be as impactful as dataset size. When combined with additional training data, even from other languages or using automatically generated transcriptions, further performance improvements are achieved.

5. Key Results: The proposed model outperforms previous state-of-the-art VSR models on benchmarks like LRS2, LRS3, and CMLR (Mandarin) datasets, reducing Word Error Rate (WER) and Character Error Rate (CER) by a large margin. The approach achieves up to 12.4% improvement on LRS2 and a 12.9% improvement on the CMLR dataset, showcasing its effectiveness across languages.

6. Conclusion: The paper argues that **model design**, optimization, and the use of prediction-based auxiliary tasks are as important as dataset size in improving VSR performance. The model is also shown to be effective across multiple languages, making it a strong candidate for real-world, multi-lingual VSR applications. The study highlights challenges in VSR and ethical considerations before the wide application of such technology.

2.6 Disentangling Homophemes in Lip Reading using Perplexity Analysis

1. Preprocessing: The process begins with mapping words to phonemes using the Carnegie Mellon Pronouncing Dictionary, followed by converting phonemes to visemes using a specific mapping convention (Lee and Yook's approach).

2. Word Detection: A viseme-to-word converter is employed, taking a sequence of visemes as input and predicting the most likely sentence as output. The word detector first performs a word lookup using the mapped visemes, then calculates perplexity scores to determine the most probable word combinations.

3. Perplexity Analysis: The perplexity score, a measure of how grammatically sound a sentence is, is used to identify the most likely word sequence. The lower the perplexity, the more grammatically correct the sentence.

4. Chunkification: In scenarios where word boundaries are unknown, the viseme sequence undergoes "chunkification," where sequences of visemes are recursively segmented into possible clusters, each corresponding to a potential word.

5. Iterative Search: A beam search algorithm with a width of 50 is implemented to iteratively calculate the perplexity of different word combinations. This reduces the computational overhead, limiting the search to the top 50 most probable sentence combinations.

6. Performance Evaluation: The system's performance is measured using various metrics, such as Viseme Error Rate (VER), Character Error Rate (CER), Word Error Rate (WER), and Sentence Accuracy Rate (SAR), based on the edit distance between the predicted and ground-truth sentences.

   This methodology leverages deep learning and language modeling techniques to tackle the challenges posed by homophemes in lip reading[13].

2.7 Training Strategies for Improved Lip-Reading

The paper [14] "Training Strategies for Improved Lip-Reading" systematically explores various training strategies and temporal models for isolated word lip-reading to improve classification accuracy.

1. Problem Addressed: Despite recent advancements in lip-reading, many approaches evaluate data augmentation methods, temporal models, and training strategies in isolation. This paper combines state-of-the-art methods to assess their individual and collective contributions.

2. Temporal Models

   (a) Compared three models: Bidirectional Gated Recurrent Units (BGRU), Multi-Scale Temporal Convolutional Networks (MS-TCN), and Densely-Connected Temporal Convolutional Networks (DC-TCN).

   (b) Found DC-TCN performed best due to dense connections and attention mechanisms.

3. Data Augmentation:

   (a) Evaluated several techniques like random cropping, flipping, mixup (combining inputs), and Time Masking.

   (b) Time Masking was the most effective, followed by mixup.

4. Additional Techniques:

   (a) Word Boundary Indicators: Using binary vectors to indicate word boundaries improved accuracy.

   (b) Self-Distillation: Sequentially trained models as both teacher and student networks, yielding incremental performance gains.

5. Experimental Results:

   - Tested on the LRW dataset (500 isolated words, 488k+ samples).
   - Best model accuracy:
     - 93.4% with all methods combined (ensemble).
     - 94.1% when pre-trained on additional datasets.
   - Ablation studies revealed individual contributions of each component.
   - Performance significantly improved for "hard-to-recognize" words.

6. Conclusion: The study highlights how combining advanced training strategies, temporal models, and data augmentations achieves state-of-the-art performance in isolated word lip-reading. Time Masking and DC-TCN were particularly impactful, and the model's success in classifying challenging words underscores its robustness.

2.8 Attention Is All You Need

This seminal 2017 paper [15] introduces the Transformer architecture, a novel deep learning model designed for sequence transduction tasks like machine translation. Unlike prior models, the Transformer eliminates recurrence and instead relies entirely on a mechanism called self-attention to draw global dependencies between input and output.

1. **Self-Attention Mechanism:** Allows the model to weigh the importance of different words in a sentence, regardless of their position, enabling better understanding of context.
2. **Parallelization:** By removing recurrence, the model significantly improves training efficiency and scalability.
3. **Positional Encoding:** Injects information about word order into the model since it has no recurrence or convolution.
4. **Encoder-Decoder Architecture:** Both encoder and decoder are composed of stacked layers with multi-head self-attention and feed-forward networks.
5. The Transformer achieved state-of-the-art performance on machine translation tasks (e.g., English-to-German and English-to-French).
6. It set the stage for later breakthroughs like BERT, GPT, and other large-scale language models.

CHAPTER 3

Feature Extraction

3.1 Lip Feature Extraction Using dlib

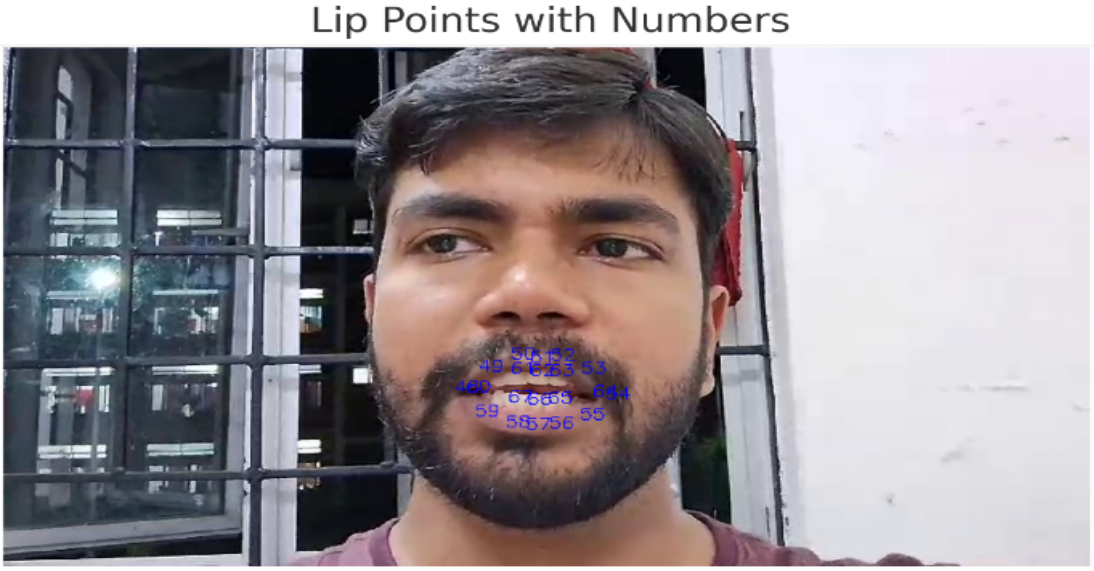

FIGURE 3.1: lip points with numbers

Lip feature extraction is a critical step in visual speech recognition, where the goal is to isolate and analyze the region of interest (ROI) corresponding to the lips. This section outlines the methodology and technical details of lip feature extraction using the dlib library, a robust toolkit for machine learning and computer vision.

1. Overview of dlib Dlib provides a pre-trained 68-point facial landmark detector that uses a shape predictor model based on ensemble regression trees. This model detects key facial landmarks, including points that define the contours of the lips.

2. Lip Landmark Points The 68 facial landmarks detected by Dlib include 20 points specifically representing the lips:

   - Outer Lip Contour: Points 48 to 59.
   - Inner Lip Contour: Points 60 to 67.

   These landmarks correspond to key positions that define the shape and motion of the lips during speech. By focusing on points 48 to 67, we can effectively track and analyze lip movements.

3. Feature Extraction Steps

   (a) Step 1: Input Video Processing

   - Frame Extraction: Using a video processing library like OpenCV, the input video is decomposed into individual frames.
   - Grayscale Conversion: Each frame is converted to grayscale to reduce computational overhead, as Dlib does not require color information for landmark detection.

   (b) Face Detection

   - Frontal Face Detector: Dlib's Histogram of Oriented Gradients (HOG)-based face detector identifies the bounding box of the face in each frame.

   (c) Facial Landmark Detection:

   - The detected face is passed to the shape predictor model (e.g., shape_predictor_68_fa
     to locate the 68 facial landmarks, including points 48 to 67.

   (d) Lip Landmark Extraction The lip region is isolated by extracting the coordinates of points 48 to 67:

   - Outer lip (48-59): Defines the boundary of the outer lip contour.
   - Inner Lip (60–67): Defines the boundary of the inner lip contour.

   (e) Normalization To account for variations in face size, orientation, or position within the frame:

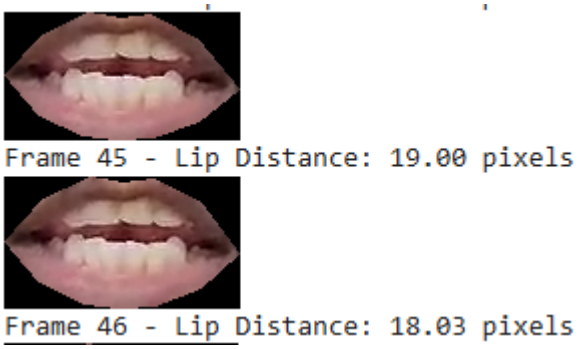

FIGURE 3.2: ROI

- The coordinates are normalized relative to a reference point, typically point 48 (the left corner of the mouth).
- The formula for normalization: xnormalized = x - x48 ynormalized = y - y48

This ensures that lip features are invariant to translation.

(f) Feature Representation: The extracted lip points are represented as a set of 2D coordinates, forming a feature vector for each frame:

$$\text{Feature Vector} = [(x_{48}, y_{48}), (x_{49}, y_{49}), \ldots, (x_{67}, y_{67})]$$

Each coordinate pair captures the position of a specific landmark, allowing precise modeling of the lip shape and motion.

(g) Applications of Lip Features

- Motion Analysis: The temporal changes in landmark positions across frames capture lip motion, crucial for speech recognition.
- Similarity Calculation: Euclidean distance between corresponding lip points of different frames is used for comparing lip movements (e.g., for word prediction).

(h) Implementation: Below is a high-level Python code snippet illustrating the feature extraction process:python code

Python Code for Lip Feature Extraction using `cv2` and `dlib`:

```
1 import cv2
2 import dlib
```

```
3
4  # Load dlib's face detector and shape predictor
5  detector = dlib.get_frontal_face_detector()
6  predictor = dlib.shape_predictor("
       shape_predictor_68_face_landmarks.dat")
7
8  # Initialize video capture
9  video_path = "input_video.mp4"
10 cap = cv2.VideoCapture(video_path)
11
12 lip_features = []
13
14 while cap.isOpened():
15     ret, frame = cap.read()
16     if not ret:
17         break
18
19     # Convert to grayscale
20     gray = cv2.cvtColor(frame, cv2.COLOR_BGR2GRAY)
21
22     # Detect faces
23     faces = detector(gray)
24     for face in faces:
25         # Detect landmarks
26         landmarks = predictor(gray, face)
27
28         # Extract lip points (48-67)
29         lip_points = [(landmarks.part(i).x, landmarks.
       part(i).y) for i in range(48, 68)]
30
31         # Normalize points relative to point 48
32         origin = lip_points[0]
33         normalized_points = [(x - origin[0], y - origin
       [1]) for x, y in lip_points]
34
35         lip_features.append(normalized_points)
```

```
36
37 cap.release()
```

(i) Advantages of dlib-Based Lip Feature Extraction

- Efficiency: The HOG-based face detection and landmark detection algorithms are fast and reliable.
- Robustness: Works well under varying lighting and facial orientations.
- Non-Intrusiveness: Requires only a standard video input, avoiding the need for additional sensors or hardware.

(j) Limitations

- Sensitive to occlusions (e.g., hand covering the mouth).
- Requires clear video frames with a well-lit and frontal view of the face for optimal landmark detection.
- Does not account for 3D lip motion unless combined with 3D reconstruction methods.

CHAPTER 4

Proposed System

4.1 Generalised Block Diagram

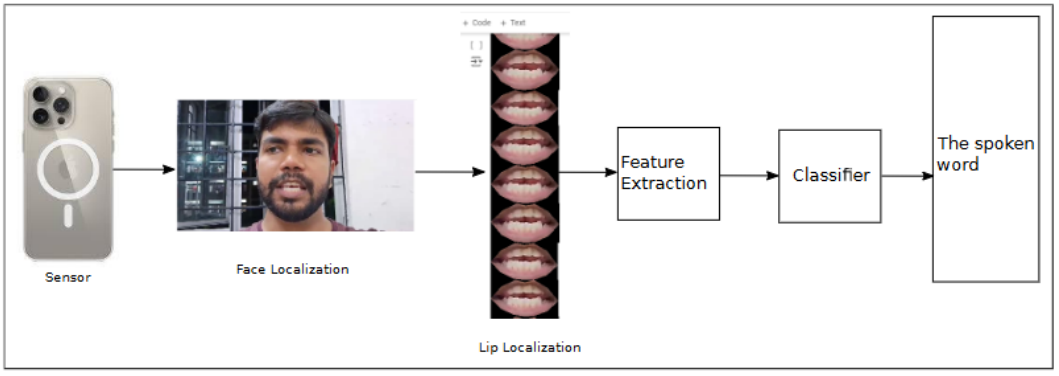

FIGURE 4.1: generalised block diagram

The block diagram in Fig.4.1 illustrates the overall architecture of a Visual Speech Recognition (VSR) system. Each block represents a key stage in the process of converting visual information from lip movements into recognized spoken words. Below is a detailed explanation of each component:

1. Sensor: The process begins with the acquisition of input data using a sensor, such as a camera or a smartphone.

- Role: Captures video frames of the speaker's face while they are speaking.
- Details: Modern cameras can record high-resolution video, ensuring that lip movements are clearly visible. A stable and well-lit environment is ideal for accurate lip reading.

2. Face Localization Once the video frames are captured, the system identifies and localizes the face region in each frame.
   - Role: Detects the location of the face to isolate it from the background or irrelevant parts of the video.
   - Implementation: This step typically uses face detection algorithms, such as
     - dlib's HOG-based face detector.
     - Deep learning models like the Multi-task Cascaded Convolutional Networks (MTCNN).
   - Outcome: Produces a bounding box around the face for further processing.
3. Lip Localization: After the face is localized, the system isolates the lip region to focus exclusively on the area responsible for speech articulation.
   - Role: Extracts the lip area (region of interest) from the detected face.
   - Implementation:
     - dlib's facial landmark detector identifies key landmarks on the face.
     - Points 48 to 67 (the outer and inner contours of the lips) are used to extract the lip region.
     - The extracted region is cropped and may be resized or normalized for consistency.
   - Outcome: A sequence of lip images (frames) is generated, capturing the shape and motion of the lips as the speaker articulates.
4. Classifier: The feature vectors are passed through a classifier, which predicts the spoken word based on the observed lip movements. The extracted lip region frames are analyzed to derive meaningful features that represent the shape, motion, and dynamics of the lips.

- Role: Converts visual data into a numerical representation suitable for analysis and classification.
- Implementation:
  - Lip landmarks (points 48–67) are used to create a feature vector for each frame.
  - Additional preprocessing, such as normalization and noise removal, ensures robustness.
  - Temporal changes in lip features across frames capture motion patterns critical for word recognition.
- Outcome: A sequence of feature vectors that encode lip movements corresponding to spoken phonemes or words.

5. Classifier: The feature vectors are passed through a classifier, which predicts the spoken word based on the observed lip movements.
   - Role: Maps the extracted features to the corresponding word in the vocabulary.
   - Implementation: Euclidean distance-based matching.
   - A predicted word or phrase based on the visual input.
6. Output: The Spoken Word. The final output is the recognized spoken word, which is derived solely from visual information (lip movements).
   - Role: Provides the end-user with the text representation of the speech.
   - Application: Useful for hearing-impaired individuals or scenarios where audio is unavailable or noisy.

This block diagram represents a pipeline where raw video input is processed step-by-step to extract and analyze lip movements, ultimately resulting in word recognition. Each stage is carefully designed to ensure accuracy, efficiency, and robustness, making it a foundational framework for developing practical visual speech recognition systems.

4.2 Algorithm

Algorithm 1 Normalize Lip Points

Require: A map $lipPoints$ of integers to coordinate pairs (x, y)

Ensure: A normalized map of lip points

if $48 \in lipPoints$ **then**
  $origin \leftarrow lipPoints[48]$ ▷ Set the origin point
  $normalizedPoints \leftarrow \{\}$ ▷ Initialize an empty map for normalized points
  for all $(key, (x, y)) \in lipPoints$ **do** ▷ Iterate over each point
    $normalizedPoints[key] \leftarrow (x - origin.first, y - origin.second)$
  end for
  return $normalizedPoints$
else
  return $lipPoints$ ▷ Return original points if key 48 is not found
end if

Algorithm 2 Euclidean Distance Calculation

Require: $x1, y1, x2, y2$

Ensure: $dx \leftarrow x2 - x1$,

$dy \leftarrow y2 - y1$
$distance \leftarrow \sqrt{dx^2 + dy^2}$

Algorithm 3 Calculate Average Lip Point Distance

Require: Two maps $lipPoints1$ and $lipPoints2$ of integers to coordinate pairs (x, y)

Ensure: The average Euclidean distance between corresponding points in $lipPoints1$ and $lipPoints2$

$totalDistance \leftarrow 0.0$
for $i = 48$ **to** 67 **do**
  if $i \in lipPoints1$ and $i \in lipPoints2$ **then**
    $totalDistance \leftarrow totalDistance$+euclideanDistance$(lipPoints1[i], lipPoints2[i])$
  end if
end for
return $totalDistance/20$

Algorithm 4 Extract and Normalize Lip Points from a Video

Require: `videoPath`: Path to the video file
`detector`: Frontal face detector
`predictor`: Shape predictor
Ensure: `lipPointsAllFrames`: Vector of normalized lip points for each frame

Initialize `lipPointsAllFrames` as an empty vector
Open the video using `VideoCapture`
if the video cannot be opened **then**
    Print error and return `lipPointsAllFrames`
end if
while frames are available **do**
    Read the current frame and convert to grayscale
    Detect faces using `detector`
    if at least one face is detected **then**
        Extract landmarks using `predictor`
        for each index i from 48 to 67 **do**
            Store landmarks in `lipPoints`
        end for
        Normalize `lipPoints`
        Append `lipPoints` to `lipPointsAllFrames`
    else
        Append empty map to `lipPointsAllFrames`
    end if
end while
Release the video resource
return `lipPointsAllFrames`

Algorithm 5 Calculate Average Distance to Reference Videos

Require: `newVideoLipPoints`: Vector of lip points for the new video
`referenceVideosLipPoints`: Vector of lip points for the reference videos
Ensure: `averageDistance`: The average distance to the reference videos

Initialize `minFrames` as the size of `newVideoLipPoints`
for each reference video `refVideo` in `referenceVideosLipPoints` **do**
    Update `minFrames` as the minimum between `minFrames` and the size of `refVideo`
end for
Initialize `totalDistances` as an empty vector
for each frame index i from 0 to `minFrames` - 1 **do**
    Get the lip points for the current frame from `newVideoLipPoints`
    if lip points for the current frame are not empty **then**
        Initialize `distancesForFrame` as an empty vector
        for each reference video `refVideo` in `referenceVideosLipPoints` **do**
            if i is less than the size of `refVideo` and lip points for this frame are not empty **then**
                Calculate the distance between `lipPointsFrame` and the corresponding frame in `refVideo` using `calculateLipPointDistance`
                Add the calculated distance to `distancesForFrame`
            end if
        end for
        if `distancesForFrame` is not empty **then**
            Calculate `avgDistanceForFrame` as the average of `distancesForFrame`
            Append `avgDistanceForFrame` to `totalDistances`
        end if
    end if
end for
Calculate `averageDistance` as the average of `totalDistances`
return `averageDistance`

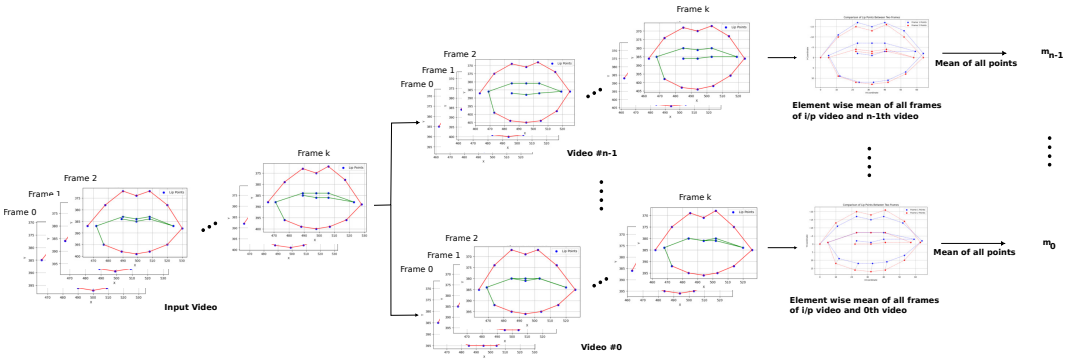

FIGURE 4.2: visual representation of methodology

4.3 Visual Representation of Methodology

The following elaboration details each step in the pipeline for identifying spoken words by comparing lip motion patterns from an input video with reference videos.

Step 1: Input Video and Reference Videos

- **Input Video:** A video containing the lip movements corresponding to an unknown spoken word. This is the test video to be matched against reference videos.
- **Reference Videos:** A dataset of pre-labeled videos, each containing lip movements for known spoken words (e.g., "A," "B," or specific words like "NO").

Step 2: Lip Landmark Normalization:

Using facial landmark detection (e.g., Dlib's facial landmark detector), lip points (points 48–67) are extracted for each frame of both the input and reference videos. We normalize the lip landmarks by translating all points such that the 48th landmark (typically corresponding to the right corner of the mouth) is positioned at the origin (0,0). This transformation aligns the lips of each frame to the origin. We did this because we want every lip 48th landmark mark start from the origin. Because we don't know where the face is positioned on the screen. By doing this, we are sure that every

lip of the frame is placed at the origin for the next task to identify the motion of lips by calculating the Euclidean Distance.

Let $P_i = (x_i, y_i)$ represent the original coordinates of the i-th lip landmark in a given frame, and let $P_{48} = (x_{48}, y_{48})$ be the coordinate of the 48th point. The normalized coordinates $\hat{P}_i$ are computed as:

$$\hat{P}_i = P_i - P_{48} = (x_i - x_{48},\ y_i - y_{48})$$

After this transformation, the 48th point becomes the origin:

$$\hat{P}_{48} = (0, 0)$$

And all other points are expressed relative to it. This normalized representation is then used to calculate the Euclidean Distance between corresponding points.

Step 3: Frame-by-Frame Lip Point Comparison

Using facial landmark detection (e.g., dlib's facial landmark detector), lip points (points 48–67) are extracted for each frame of both the input video and reference videos.

For each frame in the input video, corresponding frames in all reference videos are compared. The comparison involves computing the Euclidean Distance (ED) between each pair of lip points from the input frame and the corresponding frame in a reference video:

$$\text{ED} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}, \tag{4.1}$$

where (x_1, y_1) and (x_2, y_2) are the coordinates of a corresponding lip point in the input video and the reference video, respectively.

Step 4: Frame-Wise Average Euclidean Distance

For each frame, the distances between corresponding lip points (48–67) are averaged to compute a single value: the frame-wise or element-wise average Euclidean distance.

This step summarizes the similarity for each frame. The calculation is repeated for every frame in the input and reference videos.

Mathematically:

$$d_{ij} = \frac{\sum_{j=48}^{67} ED_{ij}}{20}, \tag{4.2}$$

- d_{ij} is the average distance for reference frame i.
- ED_{ij} is the Euclidean Distance between two landmark points. Where j is the lip landmark number and i is the ith frame of a reference video.

Step 5: Video-Level Average Euclidean Distance

After obtaining the frame-wise distances, the next step involves averaging the distances across all frames in the video. This generates a single average Euclidean distance value for the entire input video with respect to each reference video.

Mathematically:

$$X_i = \frac{\sum_{j=1}^{N} d_{ij}}{N}, \tag{4.3}$$

where:

- X_i is the average distance for reference video i.
- d_{ij} is the frame-wise average distance for frame j in reference video i.
- N is the total number of frames in the video.

Step 6: Identify the Matched Word

After calculating the average Euclidean distance for all reference videos, the reference video with the minimum distance to the input video is identified:

$$\text{Matched Word} = \min(X_0, X_1, \ldots, X_{n-1}), \tag{4.4}$$

where $X_0, X_1, \ldots, X_{n-1}$ are the average distances for reference videos $0, 1, \ldots, n-1$. The reference video with the minimum distance corresponds to the predicted spoken word.

4.3.1 Advantages of This Method

- **Robust Feature Representation:** Euclidean distance effectively captures the spatial variations of lip points.
- **Temporal Consistency:** Frame-wise and video-level averaging ensure that both individual frame differences and overall lip movement trends are considered.
- **Scalability:** The method can be extended to larger datasets of reference videos for recognizing a wider vocabulary.

4.3.2 Application Context

This approach is particularly suited for word-level visual speech recognition, where lip motions are unique for each word. It avoids the need for complex deep learning models, leveraging geometry-based feature matching for accurate predictions.

4.4 Silent Detection

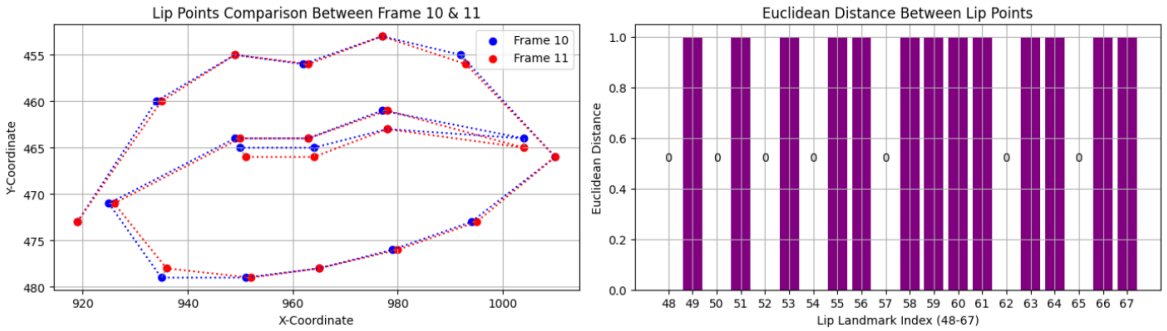

FIGURE 4.3: ED calculation of same video's frame no 10 and 11

For silent detection, we are comparing the Euclidean distance(ED) from the current frame lip points to the next frame lip points of the same video. If a speaker is not speaking, then the lip points of the current frame will overlap with those of the next frame, Fig. 4.3 and the average Euclidean distance will be below 1 pixel as can be seen in the speech activity graph Fig. 4.6. In this video, the speaker is silent until frame no 20, refer to Table.4.5, and starts speaking from frame no 21 to frame no 45. If you see the bar graph Fig. 4.3, which represents the ED between each corresponding lip point because the speaker is silent, either lip points overlap or differ with a negligible distance. If the lip point overlaps, then ED will be zero, as can be seen in the bar graph Fig.4.6. The lip landmark numbers 48, 50, 52, 54, 57, 62, and 65 are zero. On the other hand, other lip-points 49, 51, 53, 55, 56, 58, 59, 60, 61, 63, 64, 66, and 67 differ by only 1 pixel. Therefore, we set the threshold to 1 pixel. If the average ED is below 1 pixel means the speaker is silent, but if it is more than 1 pixel, the speaker is in speaking mode Fig.4.4.

When the speaker starts speaking calculated average ED from the current frame to the next frame will be more than 1 pixel, refer to Fig.4.4. The landmark numbers 48, 49, 51, and 52 have ED 0 Fig.4.4 bar graph. Landmark numbers 50, 53, 54, 60, 61, 62, and 63 have ED 2, the landmark numbers 56, 57, 58 have 4 ED. The landmark number 48 is the origin for every frame's lip points; the ED for comparison with frames for point 48 will always be zero.

The speech activity graph Fig. 4.6 shows the frame-wise mean ED of the current frame to the next frame. Since the speaker is silent till frame no 20, the average ED is below 1 pixel. Since the speaker starts speaking from frame no 21, therefore average ED is more than 1 pixel. The speaker is speaking till frame no 45. Afterward, in frame no 45, the Average Ed starts getting down because the speaker stops speaking.

Table.4.5 demonstrates each current and next frame correspondence, the lip points average Euclidean distance. There are a total of 60 frames in the video. Till framing no 21 speaker is silent; therefore calculated average Euclidean distance is lower than 1 unit. From frame no 21 to frame no 46 speaker is in speaking mode; therefore, the average Euclidean distance is greater than 1 pixel. After frame no 46, the speaker remains silent; therefore average Euclidean distance again went lower than 1 pixel.

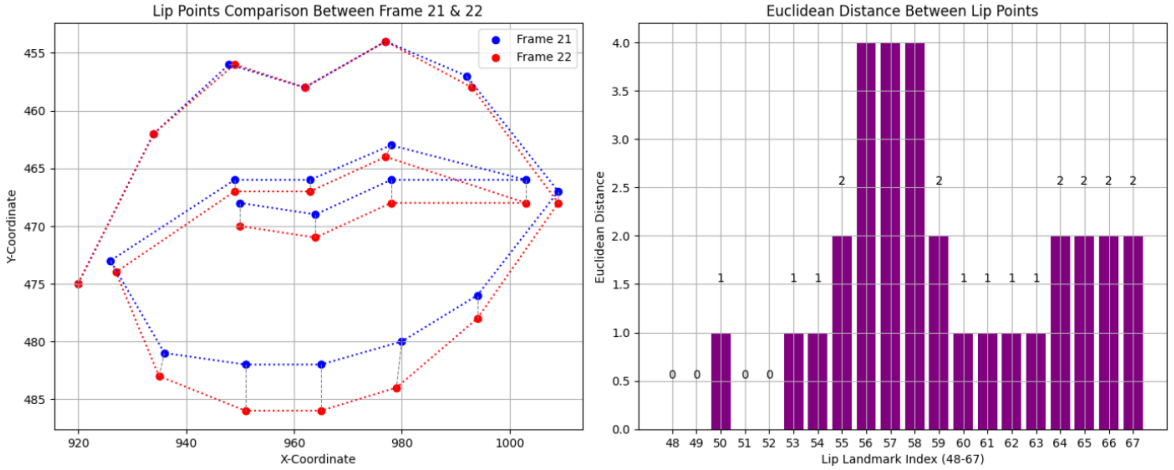

FIGURE 4.4: ED calculation of the same video's frame no 22 and 23

By calculating the Euclidean Distance of consecutive lips that appear on frames, this can be determined whether the speaker is speaking or silent.

Algorithm 6 Silent Detection Using Euclidean Distance

Require: Video with N frames, threshold $T = 1$ pixel

Ensure: Speech activity per frame: *Silent* or *Speaking*

for $i = 1$ to $N - 1$ **do**
  Extract lip landmarks from frame i: $L_i = \{(x_{i,j}, y_{i,j})\}_{j=1}^{M}$
  Extract lip landmarks from frame $i+1$: $L_{i+1} = \{(x_{i+1,j}, y_{i+1,j})\}_{j=1}^{M}$
  $sumED \leftarrow 0$
  for $j = 1$ to M **do** ▷ M = number of lip landmarks
    $ED_j \leftarrow \sqrt{(x_{i,j} - x_{i+1,j})^2 + (y_{i,j} - y_{i+1,j})^2}$
    $sumED \leftarrow sumED + ED_j$
  end for
  $avgED \leftarrow \frac{sumED}{M}$
  if $avgED < T$ **then**
    Label frame i as **Silent**
  else
    Label frame i as **Speaking**
  end if
end for

Current Frame	Next Frame	Average Euclidean Distance
1	2	0.41
2	3	0.42
3	4	0.40
4	5	0.44
5	6	0.53
6	7	0.55
7	8	0.67
8	9	0.30
9	10	0.42
10	11	0.71
11	12	0.70
12	13	0.65
13	14	0.39
14	15	0.87
15	16	0.57
16	17	0.71
17	18	1.06
18	19	0.79
19	20	1.38
20	21	0.54
21	22	1.63
22	23	2.49
23	24	3.11
24	25	2.15
25	26	1.40
26	27	1.62
27	28	1.93
28	29	2.30
29	30	2.11
30	31	1.17
31	32	4.54
32	33	2.40
33	34	3.88
34	35	2.79
35	36	2.12
36	37	1.60
37	38	4.02
38	39	0.85
39	40	4.80
40	41	2.04
41	42	0.44
42	43	2.21
43	44	2.41
44	45	0.74
45	46	1.69
46	47	0.57
47	48	0.88
48	49	1.21
49	50	1.29
50	51	0.49
51	52	0.87
52	53	1.42
53	54	1.02
54	55	0.32
55	56	1.10
56	57	0.15
57	58	0.48
58	59	0.35
59	60	0.20

FIGURE 4.5: Average ED of consecutive frames of the cat pronounced video

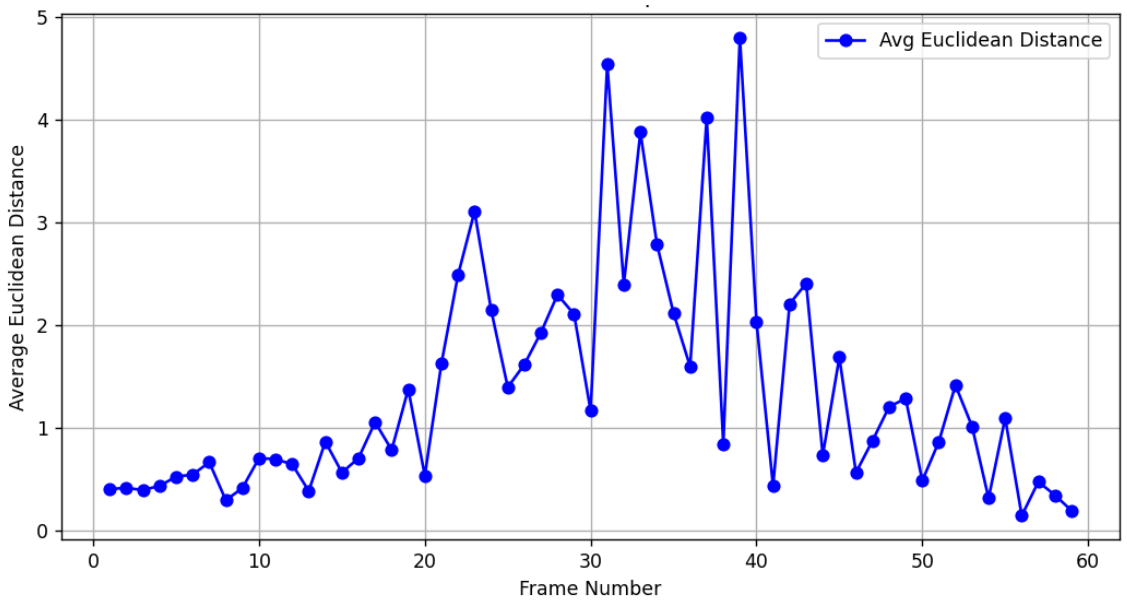

FIGURE 4.6: speech activity graph

CHAPTER 5

Experimental Results and Discussions

5.1 Dataset Preparation for Visual Speech Recognition System

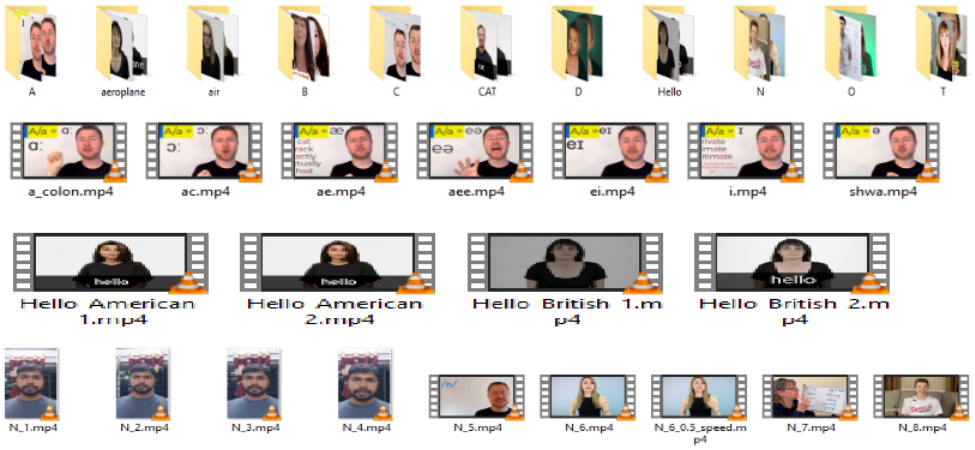

FIGURE 5.1: dataset created by me

The dataset is a crucial part of developing a reliable and accurate visual speech recognition system. The dataset used in this project was created by combining resources from online platforms and custom-recorded videos. Below is a detailed description of the dataset preparation process:

Sources of the Dataset

English with Collins Dictionary

- A notable resource was Collins' [16] pronunciation videos, which provide high-quality recordings of word pronunciations in different accents.

Self-Recorded Videos:

- In addition to publicly available content, custom videos were recorded to create a more diverse and personalized dataset.
- These recordings involved clear articulation of letters, syllables, and words to ensure precise lip movement capture.

Structure of the Dataset

The dataset was organized into folders for better accessibility and manageability, categorized as follows:

Folders by Letters:

- Individual letters (e.g., `A`, `B`, `C`) were stored in dedicated folders, each containing multiple videos of people pronouncing that specific letter.
- **Example:** Folder `A` includes files such as `a_colon.mp4`, `ae.mp4`, and `aee.mp4`, showcasing different pronunciations of the letter `A`.

Folders by Words:

- Words such as `CAT` or `HELLO` were grouped into folders, where each folder contains several videos of people pronouncing the word.
- **Example:** The `HELLO` folder contains `Hello_American_1.mp4`, `Hello_British_1.mp4`, and others, covering variations in pronunciation and accents.

Specialized Folders:

- Some folders were labeled with phonetic transcriptions or sounds, such as `/N/`, `shwa.mp4`, and `ei.mp4`, to accommodate different speech sounds or phonemes.

Video Specifications

Clarity and Quality:

- High-quality videos ensured accurate lip movement tracking and feature extraction.
- Attention was given to lighting, facial visibility, and lip movement clarity in custom-recorded videos.

Variations in Speech:

- Videos included diverse accents (e.g., British and American) to make the dataset robust and inclusive.
- Speed variations in speech were captured in some videos (e.g., `N_6_0.5_speed.mp4`) to simulate natural speaking patterns.

Dataset Utilization

The dataset serves as the foundation for the following tasks:

- **Lip Feature Extraction:** The videos are used to extract landmarks (points 48–67) representing the lip region for every frame.
- **Frame Matching:** Frame-by-frame Euclidean distance comparisons are performed to match input videos with reference videos.
- **Training and Testing:** The dataset enables training the visual speech recognition system and evaluating its accuracy in predicting words based on lip motion.

Advantages of the Dataset

- **Comprehensive Content:** The dataset includes a variety of pronunciations, accents, and speeds, improving model robustness.
- **Customizability:** By including self-recorded videos, the dataset was tailored to specific project requirements.
- **Phoneme and Word Level:** Both phoneme-level (letters) and word-level recordings were included, enabling flexibility in system development.

This combination of resources ensures that the dataset is diverse, well-organized, and suitable for lip-reading research and applications.

5.2 Result

This diagram Fig.5.2 illustrates the workflow for predicting a spoken word by comparing an input video with a dataset of reference videos using Euclidean distance as the similarity metric. The following detailed steps elaborate on the process:

Step 1: Dataset Preparation

- The dataset consists of pre-recorded videos containing lip movements for specific words (e.g., `Hello`).
- Each video in the dataset is analyzed frame-by-frame, and lip features are extracted using landmarks corresponding to points 48–67 (as defined by dlib).
- These extracted lip features are stored for each frame to create a reference for word recognition.

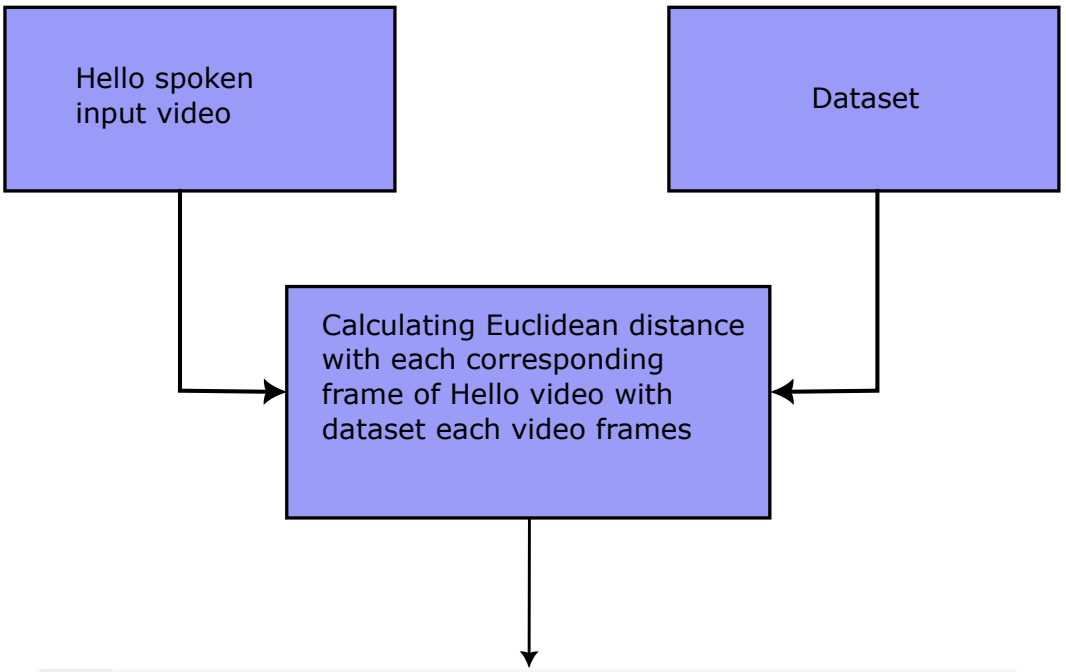

FIGURE 5.2: result

Step 2: Input Video Processing

- The input video of a spoken word (e.g., `Hello`) undergoes similar processing.
- Lip landmarks (points 48–67) are extracted for each frame of the input video.
- These features form a sequence of lip shape data representing the spoken word in the input video.

Step 3: Frame-by-Frame Comparison

- For each frame in the input video, the Euclidean distance is calculated between the corresponding lip landmarks of the input video and each frame of every video in the dataset.

- The Euclidean distance serves as a measure of similarity between the lip shapes in the input video and those in the dataset videos.

Step 4: Averaging Distances

- The distances calculated for all frames are averaged to provide a single similarity score for the input video against each video in the dataset.
- The process is repeated for all dataset videos to generate a list of average distances corresponding to each reference video.

Step 5: Word Prediction

- The word associated with the dataset video that has the minimum average Euclidean distance to the input video is selected as the predicted word.
- **Example:**
  - The distances for `Hello` and other words (`Cat`, `Aeroplane`, etc.) are calculated and compared.
  - The dataset video with the lowest average distance indicates the closest match, and the word `Hello` is predicted.

Advantages of the Process

- **Frame-by-Frame Accuracy:** Comparing individual frames ensures high granularity in the analysis.
- **Statistical Robustness:** Averaging distances over frames minimizes errors due to variations in individual frames.
- **Scalability:** This method can be extended to include multiple words and videos for larger datasets.

Results Displayed

The results include:

- The average Euclidean distance for each dataset video.
- The predicted word based on the minimum average distance.

Example output:

```
Average distance to 'Hello': 3.844
Predicted word: Hello.
```

5.3 Detection of whether speaker is speaking or silent

FIGURE 5.3: silent frame

When the speaker is not speaking, the average Euclidean distance will be less than compared to the speaker when speaking. It is obvious that if the speaker is not speaking, the lips will not be in motion. In this scenario, the average Euclidean distance will not vary. As we are calculating the corresponding lip point distance between the current frame to the next frame. If the current frame lip points and the next frame lip points are appearing at the same coordinates, then there will not be any lip movement, and

FIGURE 5.4: Beginning of Speaking Frame

the Average Euclidean Distance will be below 1 pixel, which signifies that the speaker is not speaking Fig.5.3. Until frame no. 21, the speaker was silent, so the Euclidean Distance below frame 21 was less than 1 pixel. However, from frame no. 21, the speaker starts speaking, so the Euclidean distance between frame no 21 to 22 is 1.63, which shows that the mouth is going to open. From frame no. 21, the speaker starts speaking, and the Euclidean Distance starts increasing, which shows that the speaker is now in speaking mode, Fig.5.4.

FIGURE 5.5: Middle of the speaking

Between frame no 31 and 32 in fig 8 Fig.5.5, the Average Euclidean distance is 4.54, which shows the lip has travelled with a good margin between consecutive frames. This identifies that the user is in speaking mode.

FIGURE 5.6: silent frame

In frame no. 46, the speaker has completed his speaking and closes his mouth. Therefore, for the Average Euclidean Distance between frame no. 46 to 47 is 0.56 pixels, denoting that the speaker is going in silent mode Fig.5.6.

CHAPTER 6

Conclusion and Future Work

6.1 Conclusion

The developed algorithm successfully recognizes spoken words by analyzing lip movements without relying on Deep Learning for word prediction. Instead, it utilizes a custom-designed methodology that emphasizes efficiency, accuracy, and computational simplicity.

Key Features of the Algorithm

Custom Dataset Application

- The algorithm has been exclusively tested on a custom dataset created specifically for this project.
- The dataset includes carefully recorded videos representing various words, ensuring diversity in lip movement patterns.

Performance

- The algorithm achieves a word recognition accuracy of 75% on the custom dataset.

- The accuracy drop is primarily due to homophones—words with similar lip movements but different meanings (e.g., `bat` and `pat`).

Strengths of the Algorithm

No Deep Learning for Prediction

- Unlike conventional approaches that employ deep learning end-to-end, this algorithm uses deep learning solely for lip feature extraction (via landmark points 48–67).
- For word prediction, a lightweight and computationally efficient algorithm designed specifically for this project is employed.

Reduced Computational Overhead

- By minimizing the reliance on deep learning models for prediction, the algorithm significantly reduces computational requirements.
- This approach is particularly beneficial for systems with limited hardware resources or real-time applications where processing speed is critical.

Focused Optimization

- The algorithm's design is tailored to analyze lip movements with precision, leveraging Euclidean distance-based comparisons to achieve reliable predictions.

Limitations

- The primary challenge lies in distinguishing between words with similar lip shapes (homophones) due to the inherent ambiguity in visual speech recognition.
- Expanding the dataset and incorporating additional features (e.g., temporal dynamics, facial muscle movement) could further enhance accuracy and robustness.

Implications

This algorithm represents a novel approach to spoken word recognition, balancing accuracy and efficiency. By reducing the dependency on deep learning for prediction, it offers a resource-friendly alternative for lip-reading systems, paving the way for applications in assistive technologies, human-computer interaction, and silent communication tools.

This approach can be expanded for larger datasets and further optimized to address challenges associated with homophones, potentially improving its utility in real-world scenarios.

6.2 Future Work and Proposed Improvements

While the current algorithm demonstrates significant potential for lip-reading-based word recognition, several avenues for further enhancement and evaluation have been identified. The following steps outline future work that could improve accuracy, robustness, and generalization of the system:

6.2.1 Acquire Videos of the Same Word with Different Accents

Objective: Enhance the system's ability to generalize across diverse speakers and accents.

Plan:

- Collect and analyze videos of the same word spoken by individuals with different accents.
- Investigate how lip movements vary across accents and incorporate these variations into the algorithm to improve recognition accuracy.

- Create a diverse dataset that includes regional and international accents for better system training and evaluation.

6.2.2 Extract 3D Lip Features for Improved Accuracy

Objective: Overcome the limitations of 2D feature extraction by leveraging 3D data for more detailed analysis.

Plan:

- Use 3D models to capture lip depth, curvature, and subtle movements that are not visible in 2D projections.
- Incorporate depth cameras or stereo vision techniques to extract 3D data from video recordings.
- Develop algorithms that integrate 3D features with existing methods, potentially improving recognition rates, especially for challenging cases like homophones.

6.2.3 Apply the Algorithm to Publicly Available Datasets

Objective: Test the algorithm's performance on standardized datasets to benchmark against other methods.

Plan:

- Evaluate the system on publicly available datasets such as GRID, TCD-TIMIT, or Lip Reading Sentences (LRS).
- Compare the results on these datasets with those achieved on the custom dataset to identify potential gaps and improvements.
- Use cross-dataset validation to ensure the algorithm's robustness and generalizability.

6.2.4 Implement Deep Learning for Lip Reading and Compare Results

Objective: Benchmark the algorithm against deep learning-based approaches to assess its relative performance and efficiency.

Plan:

- Develop a deep learning-based lip-reading model using techniques like convolutional neural networks (CNNs), recurrent neural networks (RNNs), or transformers.
- Train the deep learning model on the same datasets and evaluate it under identical conditions as the custom algorithm.
- Compare metrics such as accuracy, computational efficiency, and scalability to highlight the advantages and limitations of each approach.

6.2.5 Address the Homophone Challenge

Objective: Solve the problem of recognizing words with similar lip movements (e.g., *bat* and *pat*).

Plan:

- Investigate auxiliary features like tongue movement (if visible) or contextual cues from sentences to differentiate homophones.
- Explore multimodal approaches, incorporating audio signals or visual clues beyond the lips to disambiguate similar words.
- Develop a probabilistic model or classifier that considers sequential dependencies or context for better predictions in ambiguous cases.

6.2.6 Compare Results with State-of-the-Art Systems

Objective: Evaluate the algorithm's performance against state-of-the-art lip-reading systems to establish its effectiveness and identify areas for improvement.

Plan:

- Compare the results with cutting-edge methods such as LipNet, AVSR (Audio-Visual Speech Recognition), and other established techniques.
- Analyze performance in terms of accuracy, computational efficiency, robustness to noise, and generalization across datasets.
- Highlight scenarios where the proposed algorithm performs better and identify areas where state-of-the-art systems excel.

References

[1] Y. M. Assael, B. Shillingford, S. Whiteson, and N. De Freitas, "Lipnet: End-to-end sentence-level lipreading," *arXiv preprint arXiv:1611.01599*, 2016. DOI: 10.48550/arXiv.1611.01599.

[2] J. S. C., A. S., O. V., and A. Z., "Lip reading sentences in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6447–6456, 2017. DOI:10.1109/CVPR.2017.367.

[3] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 8717–8727, 2018. DOI:10.1109/TPAMI.2018.2889052.

[4] E. M. Hilliard, "The müller-wallee: Method of lip-reading for the deaf," 1915. DOI: https://doi.org/10.1288/00005537-191511000-00011.

[5] E. Petajan, B. Bischoff, D. Bodoff, and N. M. Brooke, "An improved automatic lipreading system to enhance speech recognition," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 19–25, 1988. DOI: 10.1145/57167.57170.

[6] A. Fernandez-Lopez and F. M. Sukno, "Survey on automatic lip-reading in the era of deep learning," *Image and Vision Computing*, vol. 78, pp. 53–72, 2018. DOI: 10.1016/j.imavis.2018.07.002.

[7] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 198–213, 2002. DOI: 10.1109/34.982900.

[8] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003. DOI: 10.1109/JPROC.2003.817150.

[9] Y. Kumar, D. Sahrawat, S. Maheshwari, D. Mahata, A. Stent, Y. Yin, R. R. Shah, and R. Zimmermann, "Harnessing gans for zero-shot learning of new classes in visual speech recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 2645–2652, 2020. DOI: 10.1609/aaai.v34i03.5649.

[10] S. Fenghour, D. Chen, K. Guo, and P. Xiao, "Lip reading sentences using deep learning with only visual cues," *IEEE Access*, vol. 8, pp. 215516–215530, 2020. DOI: 10.1109/access.2020.3040906.

[11] H. Wang, P. Guo, W. Chen, P. Zhou, and L. Xie, "The npu-aslp-liauto system description for visual speech recognition in cnvsrc 2023," *arXiv preprint arXiv:2401.06788*, 2024. DOI: 10.48550/arXiv.2401.06788.

[12] P. Ma, S. Petridis, and M. Pantic, "Visual speech recognition for multiple languages in the wild," *Nature Machine Intelligence*, vol. 4, no. 11, pp. 930–939, 2022. DOI: 10.1038/s42256-022-00550-z.

[13] S. Fenghour, D. Chen, K. Guo, and P. Xiao, "Disentangling homophemes in lip reading using perplexity analysis," *arXiv preprint arXiv:2012.07528*, 2020. DOI: 10.48550/arXiv.2012.07528.

[14] P. Ma, Y. Wang, S. Petridis, J. Shen, and M. Pantic, "Training strategies for improved lip-reading," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8472–8476, IEEE, 2022. DOI: 10.48550/arXiv.2209.01383.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017. DOI: 10.48550/arXiv.1706.03762.

[16] Collins Dictionary, "Collins online dictionary – word pronunciation videos," 2025. Accessed: 2024-2025.

CHAPTER 7

Appendix I

Journal under communication

1. **Md. Laraib Ahmad**, Dr. Debbrota Paul Chowdhury, "**Visual Speech Recognition for Seamless Communication With Hearing Impared Persons**".