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
December, 2024

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: ________________

Md. Laraib Ahmad
2322207

Date: ________________

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 80% 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 thanks to 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 and I have learned so much from him. Whatever I know in the field of Lip Reading is just because of him. I do consider myself lucky for working under such 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 have knowledge of the underlying language as well as 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 better understand 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, 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 (Assael et al., 2016), diverse speaking styles (Chung et al., 2017, Martinez et al., 2020), and the need for large and labeled datasets for training (Afouras et al., 2018). 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 serves as a valuable complement to hearing aids and cochlear implants, enhancing their effectiveness in various communication scenarios. While these assistive devices amplify auditory signals or directly stimulate the auditory nerve, they often face challenges in environments with significant background noise, overlapping conversations, or poor audio quality. In such situations, individuals may still struggle to distinguish spoken words despite the technological assistance.

By incorporating lip reading, users can bridge the gap between the auditory and visual components of speech. The visual cues from lip movements, combined with the amplified audio signal, provide additional context to help decipher words and phrases. This dual-channel approach is particularly effective in scenarios where clarity is compromised, such as crowded settings, outdoor environments with wind interference, or during digital communication over low-quality audio connections.

Furthermore, lip reading enhances the user experience for cochlear implant recipients during the initial adjustment period, as they learn to interpret the electrical signals from the implant as meaningful sounds. Visual reinforcement from lip reading accelerates this adaptation process, allowing users to achieve better comprehension and confidence in everyday conversations.

Ultimately, the integration of lip reading alongside hearing aids and cochlear implants empowers individuals to communicate more effectively, fostering inclusivity and independence across a wide range of social and professional settings.

1.2 Laryngeal cancer

FIGURE 1.4: Laryngeal cancer

Laryngeal cancer, which affects the tissues of the larynx or voice box, can significantly impact an individual's ability to speak. This condition, particularly when it involves the vocal cords, often leads to changes in voice quality, such as hoarseness, reduced volume, or a complete loss of vocal ability. In more severe cases, treatment for laryngeal cancer may necessitate the surgical removal of all or part of the larynx, a procedure known as a laryngectomy. This can permanently eliminate the patient's ability to produce natural speech.

After such interventions, individuals are left with the challenge of finding alternative means of communication. In many cases, they rely on lip movements as a primary mode of expressing themselves. By forming words with their lips silently, they enable others to interpret their intended speech through lip reading. This method becomes essential for interactions, especially for individuals who do not use advanced speech prosthetics or assistive technologies like electronic voice devices.

Lip reading is not only practical but also empowers individuals who have undergone these life-altering treatments to maintain a sense of autonomy and engage in conversations with family, friends, and colleagues. Moreover, visual speech recognition systems, which can interpret lip movements and translate them into text or synthetic voice, offer a promising avenue to further improve communication for those affected by laryngeal cancer. These technologies provide hope for enhanced accessibility, supporting individuals in overcoming the barriers imposed by the loss of natural speech.3.2.

1.3 Reasearch Advancement

1.3.1 Early 20th Century

The early 20th century marked a pivotal period in the formalization of lip reading as a skill, particularly for the benefit of the deaf and hard-of-hearing communities. During this time, significant strides were made in understanding and teaching the principles of lip reading, driven by the broader movement to integrate individuals with hearing impairments into society. Educational institutions dedicated to the deaf began incorporating structured lip-reading curricula into their programs, recognizing its potential as a vital communication tool.

These schools aimed to empower students by teaching them to interpret speech visually through the movement of a speaker's lips, facial expressions, and other non-verbal cues. The approach was often complemented by speech training, fostering a dual skill set that enhanced both receptive and expressive communication abilities. Additionally, advancements in pedagogy during this era laid the groundwork for the development of specialized training materials and methodologies, many of which are still influential in modern deaf education.

This period also saw increased awareness of the need for accessible communication, prompting researchers and educators to refine their understanding of the cognitive and physiological aspects of lip reading. These efforts not only improved the quality of instruction but also contributed to the broader societal recognition of the capabilities and rights of individuals with hearing impairments.

1.3.2 1970s

The 1970s marked the inception of automated lip reading as a research field, driven by advances in computer science and a growing interest in human-computer interaction. Researchers began exploring the visual aspects of speech, aiming to understand how the movements of the lips, mouth, and surrounding facial regions could be systematically captured, processed, and analyzed by computers. This period saw the development of foundational theories and methods for visual speech recognition, inspired by the human ability to lip-read.

Initial efforts were primarily theoretical, focusing on modeling the mechanics of speech articulation and devising ways to extract meaningful features from video data. Researchers utilized basic image processing techniques to isolate and track lip movements, often employing hand-annotated datasets due to the lack of sophisticated automation tools. These studies aimed to identify patterns in the shapes, positions, and motions of the lips that corresponded to specific phonemes or spoken sounds.

The challenges of the era were significant, given the limited computational power and the nascent state of machine learning algorithms. However, these pioneering efforts laid the groundwork for future innovations by demonstrating the feasibility of using visual data to supplement or even replace auditory signals in speech recognition systems. The research also underscored the potential applications of automated lip reading, from aiding the hearing-impaired to enhancing speech recognition technologies in noisy environments.

1.3.3 1980s and 1990s

The 1980s and 1990s were transformative for automated lip reading, as advances in computer vision, pattern recognition, and artificial intelligence expanded the field's scope. Building on the foundational work of the 1970s, researchers developed systems to analyze visual speech elements with greater accuracy by leveraging emerging computational power and novel algorithms.

In the 1980s, efforts focused on segmenting and tracking lip movements, with milestones including feature extraction techniques like edge detection, contour modeling, and motion analysis. These methods aimed to map lip shapes and movements to specific phonemes and were often implemented on specialized hardware due to limited general-purpose computing power.

By the 1990s, statistical models such as Hidden Markov Models (HMMs) became central for recognizing temporal patterns in speech, supported by larger datasets and improved machine learning techniques. Multimodal approaches also gained traction, integrating visual and audio data for robust performance in noisy environments.

Practical applications, including aiding the hearing impaired, enhancing human-computer interaction, and improving surveillance, became a focus during this period. These decades firmly established automated lip reading as a key research domain, paving the way for deep learning breakthroughs in the 21st century.

1.3.4 2000s

The 2000s marked significant progress in automated lip reading, driven by advancements in AI and deep learning. Researchers built upon statistical models like Hidden Markov Models (HMMs) to analyze the temporal dynamics of visual speech, often using features such as lip contours, optical flow, and shape-based descriptors. These methods were bolstered by the availability of larger datasets and improvements in computational hardware, enabling more effective training.

The latter half of the decade saw the emergence of deep learning as a transformative approach. Convolutional Neural Networks (CNNs) were adapted for visual speech recognition, allowing automated feature extraction from raw video frames. Combined with Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, these models captured both spatial and temporal patterns in lip movements.

Multimodal approaches also gained traction, integrating audio and visual inputs to improve recognition in challenging conditions. By the end of the decade, these advancements had significantly enhanced the accuracy and practicality of automated lip reading systems, setting the stage for future breakthroughs.

1.3.5 2010s

The 2010s marked a transformative era for automated lip reading, driven by deep learning techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). CNNs excelled at extracting spatial features from video frames, capturing intricate lip movements, while RNNs, including Long Short-Term Memory (LSTM) networks, modeled temporal dynamics in speech sequences. Together, these formed end-to-end architectures that significantly improved lip-reading accuracy.

The availability of large datasets, such as GRID, TCD-TIMIT, and LRW, played a crucial role in training deep models. These datasets offered diverse speakers, accents, and environments, enhancing system generalizability. Increased computational power from GPUs and cloud computing further accelerated progress.

Applications expanded to assistive technologies for the hearing impaired, silent speech interfaces, and security systems. By the decade's end, deep learning had enabled real-time, robust lip-reading systems capable of performing in diverse real-world conditions.

1.3.6 2016

In 2016, Google DeepMind introduced LipNet, a groundbreaking model that showcased exceptional accuracy in automated lip reading by harnessing deep learning techniques. LipNet represented a significant departure from traditional approaches, as it was the first end-to-end deep learning model specifically designed for sentence-level lip reading, rather than focusing on isolated words or phonemes.

LipNet combined Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) into a unified architecture. CNNs were used to extract spatial features from sequences of video frames, capturing the intricate movements of the lips. These features were then passed to a Gated Recurrent Unit (GRU)-based RNN, which modeled the temporal dynamics and relationships between frames, effectively learning how lip movements unfolded over time.

One of LipNet's key innovations was its use of the Connectionist Temporal Classification (CTC) loss function, which allowed the model to align lip movements directly with textual sequences without requiring frame-by-frame annotations. This made training more efficient and eliminated the dependency on labor-intensive manual labeling.

Trained on datasets like GRID, LipNet achieved unprecedented accuracy in sentence-level lip reading, outperforming both traditional statistical models and earlier machine learning approaches. The model demonstrated the ability to interpret continuous speech, even in challenging conditions, marking a significant milestone in the field.

LipNet's success highlighted the potential of end-to-end deep learning for lip reading and inspired further research into more advanced architectures and applications, setting a new standard for visual speech recognition systems.

1.3.7 2020s

Research in the 2020s has led to significant advancements in lip reading technologies, further improving their accuracy, efficiency, and real-world applicability. The integration of advanced deep learning techniques, including Transformer models and multimodal approaches, has enhanced the robustness of visual speech recognition systems, making them more adaptable to diverse environments and challenging conditions.

Lip reading technologies are increasingly used in assistive technology, providing real-time transcription for the deaf and hard-of-hearing, and improving communication in settings such as education, healthcare, and public services. In the security domain, lip reading is deployed for silent surveillance and forensic analysis, enabling monitoring in environments where audio capture is impractical or undesirable.

In human-computer interaction (HCI), lip reading systems are integrated into voice-controlled applications, allowing users to interact with devices in noisy or crowded environments. These systems are also improving privacy, as they enable speech recognition without sound, particularly useful in public or confidential spaces.

Furthermore, research is focusing on enhancing the performance of lip reading models in real-world scenarios, including training on more diverse datasets to improve multilingual and cross-accent recognition. The push towards real-time lip reading and greater adaptability to varied lighting, backgrounds, and speaker characteristics continues to drive the field forward. As a result, lip reading technologies are finding wider applications across industries, with significant potential for improving accessibility, security, and communication.

1.4 Motivation

1.4.1 Accessibility for Hearing-Impaired Individuals

Lip reading is a crucial tool for those with hearing impairments, enabling them to comprehend spoken language by observing the speaker's lip movements. This research aims to improve the accuracy and efficiency of automated lip reading systems, facilitating better communication in a variety of settings, such as classrooms, workplaces, and social environments. Enhanced lip reading technologies can offer a more accessible world for those who rely on visual cues to communicate.

1.4.2 Speech Enhancement in Noisy Environments:

In environments with significant background noise (e.g., crowded public spaces, factories, or airports), traditional audio-based speech recognition can struggle to accurately capture speech. Lip reading offers a potential solution by using visual cues to supplement or enhance the audio signal. By integrating lip reading with existing audio technologies, we can create more robust speech enhancement systems, improving communication and speech recognition in challenging environments.

1.4.3 Multimodal Systems:

Lip reading can be combined with other modes of communication, such as audio, gestures, and facial expressions, to create multimodal systems. These systems can leverage multiple sources of information to improve understanding and interaction in human-computer interfaces. By integrating lip reading into multimodal systems, we can enhance the accuracy, reliability, and user experience of applications in areas like virtual assistants, augmented reality, and human-robot interaction, especially in noisy or dynamic environments.

1.4.4 Advances in AI and Machine Learning:

The rapid development of AI and machine learning techniques, particularly deep learning, has led to significant breakthroughs in automated lip reading. Advanced algorithms can now analyze visual speech data with greater precision, enabling real-time lip reading systems that can function effectively in complex, unconstrained environments. These advancements make it possible to apply lip reading to a wider range of applications, from assistive technologies to security and surveillance systems, expanding its potential impact.

1.5 Problem Statement

Despite significant advancements in visual speech recognition, several challenges remain that hinder its widespread application and effectiveness. These challenges include:

1.5.1 Variability in Lip Movements

Lip movements can vary significantly between individuals due to differences in facial structure, speech patterns, and articulation. Even the same person may exhibit variations in lip movements based on factors like speech rate, emotion, and context. This variability makes it difficult for lip-reading models to consistently interpret speech across different speakers and situations. Developing systems that can generalize across diverse lip shapes and movements remains a major challenge.

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[2] **"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 is able to 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 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 [3]"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 [4]”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 [5] "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[6].

2.7 combination of vision-level and language-level adaptation

The methodology used in the paper[7] combines vision-level adaptation (using padding prompts and LoRA for lip appearances, movements, and speaking speed) and language-level adaptation (using input prompt tuning to learn speaker-specific linguistic patterns) to adapt a pre-trained lip-reading model to target speakers effectively.

The paper uses the VoxLRS-SA dataset, which is derived from VoxCeleb2 and LRS3 datasets. The VoxLRS-SA dataset is specifically designed to address the limitations of previous datasets, offering a vocabulary of approximately 100K words, diverse pose variations, and enabling the validation of adaptation methods for sentence-level lip reading in real-world scenarios.

The paper finds that the proposed speaker-adaptive lip reading method, which integrates both vision and language-level adaptations, significantly improves sentence-level lip reading performance in real-world scenarios with minimal adaptation data, outperforming previous methods.

The paper presents a novel speaker-adaptive lip reading method that adapts a pretrained model to unseen speakers at both vision and language levels. The proposed method uses padding prompts and LoRA to adapt the visual encoder to target speaker-specific lip appearances, movements, and speaking speeds, while input prompt tuning is applied to adapt the language model to the speaker's linguistic patterns. A new dataset, VoxLRS-SA, is introduced, which contains a large vocabulary and diverse speaker poses, enabling robust validation in real-world scenarios. The results show that the method improves performance in sentence-level lip reading with minimal adaptation data, outperforming previous speaker-adaptive techniques.

2.8 Training Strategies for Improved Lip-Reading

The paper [8] "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.

2.9 Support Vector Machine algorithm to recognize speech.

The [9]paper explores the application of **lip reading** for visual speech recognition (VSR) using deep learning techniques. The primary focus is on recognizing spoken words by analyzing lip, face, and tongue movements when audio is unavailable or unreliable. This can significantly benefit individuals with hearing impairments by offering a way to interpret speech visually.

1. Deep Learning Approach: The study utilizes deep learning methods to improve lip-reading accuracy, surpassing traditional machine learning approaches. The authors suggest that deep learning networks can extract higher-level features like shapes and boundaries to interpret speech from visual inputs.

2. Challenges in Lip Reading: Lip reading is inherently difficult due to variability in speaking patterns, accents, and lighting conditions. The system addresses these challenges by relying heavily on visual features of lip movements.

3. Proposed System: The system includes a process for converting video to frames, detecting and segmenting the face, and identifying the region of interest (ROI), which is the lips. The lip movement patterns are extracted and classified using the SVM (Support Vector Machine)** algorithm to recognize speech.

4. Dataset and Results: The authors trained their model on a custom dataset, achieving a recognition accuracy of 75%. The system detects human lip expressions and translates them into words or sentences, although it operates on pre-recorded videos rather than real-time input.

5. Future Work: The paper proposes extending the system to function in real-time, where live video from a camera can be processed to provide real-time lip reading output.

6. In conclusion, while the paper demonstrates notable progress in VSR, the authors recognize that lip reading is still a challenging problem, especially in ”in-the-wild” data. Future advancements aim to make the technology more robust and applicable in real-time scenarios.

CHAPTER 3

Feature Extraction

3.1 Lip Feature Extraction Using dlib

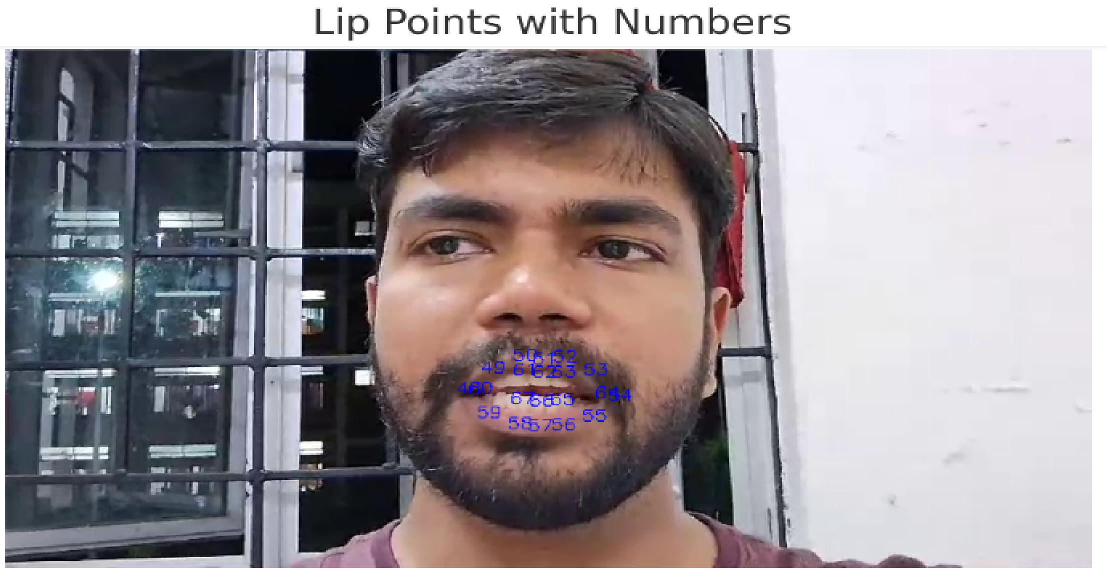


Figure 3.1: lip points with numbers

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

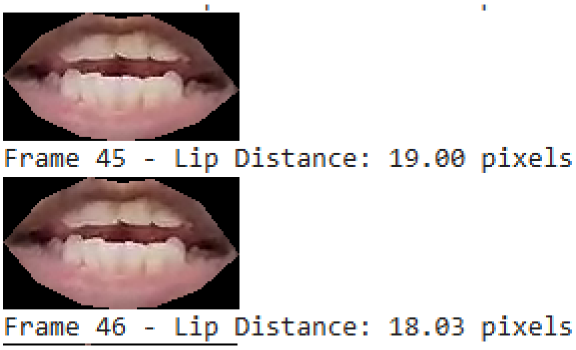
   - Frontal Face Detector: dlib's Histogram of Oriented Gradients (HOG)-based face detector identifies the bounding box of the face in each frame.

   (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.



```
Frame 45 - Lip Distance: 19.00 pixels
Frame 46 - Lip Distance: 18.03 pixels
```

Figure 3.2: ROI

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

- 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`:

```python
import cv2
import dlib

# Load dlib’s face detector and shape predictor
detector = dlib.get_frontal_face_detector()
predictor = dlib.shape_predictor("shape_predictor_68_face_landmarks.dat")

# Initialize video capture
video_path = "input_video.mp4"
cap = cv2.VideoCapture(video_path)

lip_features = []

while cap.isOpened():
    ret, frame = cap.read()
    if not ret:
        break

    # Convert to grayscale
    gray = cv2.cvtColor(frame, cv2.COLOR_BGR2GRAY)

    # Detect faces
    faces = detector(gray)
    for face in faces:
        # Detect landmarks
        landmarks = predictor(gray, face)

        # Extract lip points (48-67)
        lip_points = [(landmarks.part(i).x, landmarks.part(i).y) for i in range(48, 68)]

        # Normalize points relative to point 48
        origin = lip_points[0]
        normalized_points = [(x - origin[0], y - origin[1]) for x, y in lip_points]

        lip_features.append(normalized_points)

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

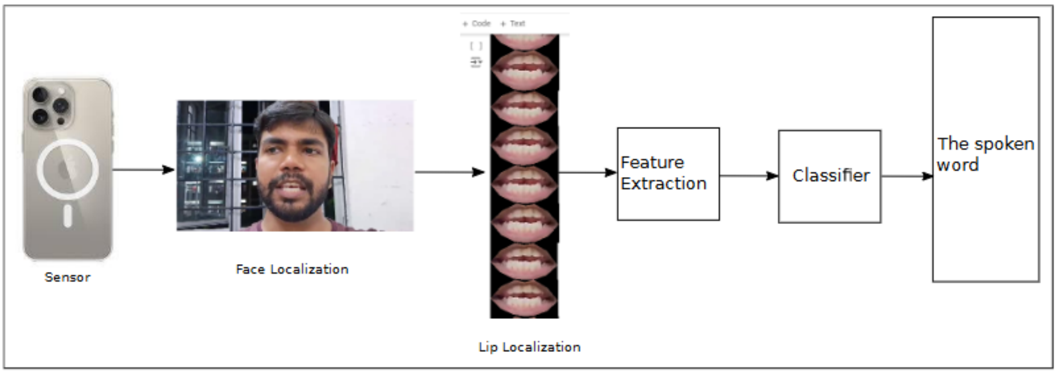


Figure 1: Generalized block diagram of visual speech recognition system

FIGURE 4.1: generalised block diagram

The block diagram in Figure 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
        $totalDistance \leftarrow totalDistance + \text{euclideanDistance}(lipPoints1[i], lipPoints2[i])$
    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
```

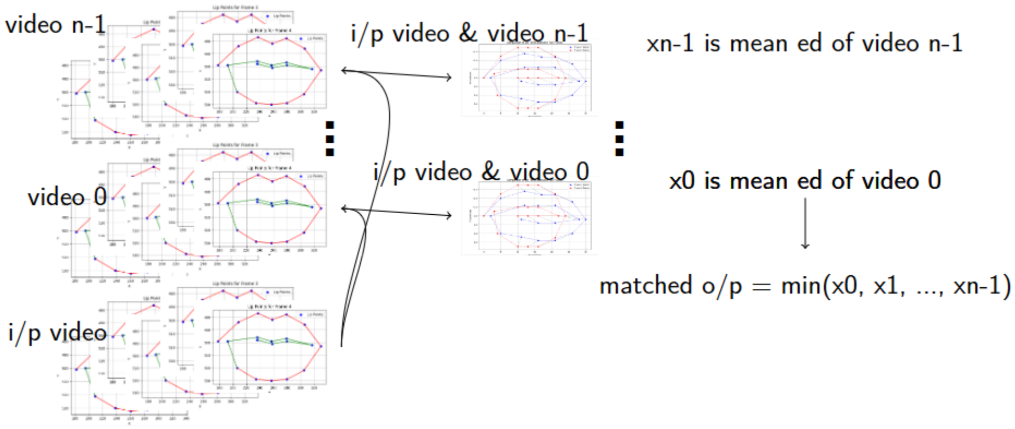


FIGURE 4.2: visual representation of methodology

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: 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:

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

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 3: Frame-Wise Average Euclidean Distance

For each frame, the distances between all corresponding lip points (48–67) are averaged to compute a single frame-wise average Euclidean distance. This step summarizes the similarity for a given frame.

This calculation is repeated for every frame in the input video and reference videos.

Step 4: 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:

$$X_i = \frac{\sum_{j=1}^{N} d_{ij}}{N}, \tag{4.2}$$

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 5: 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:

$$\text{Matched Word} = \min(X_0, X_1, \ldots, X_{n-1}), \tag{4.3}$$

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.

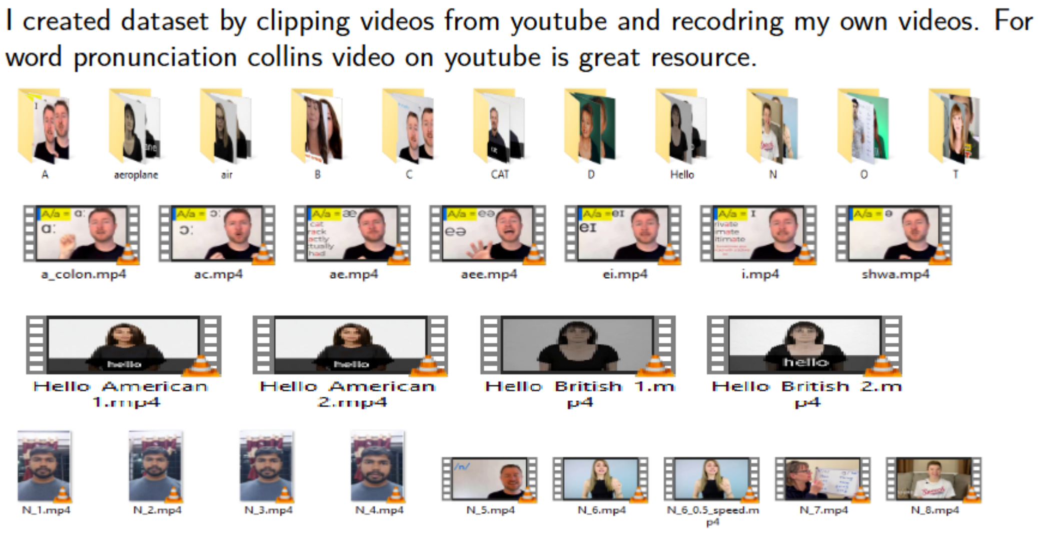
4.3.2 Application Context

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.

CHAPTER 5

Experimental Results and Discussions

5.1 Dataset Preparation for Visual Speech Recognition System



I created dataset by clipping videos from youtube and recodring my own videos. For word pronunciation collins video on youtube is great resource.

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

YouTube Clips:

- Video clips from YouTube were used as a primary resource to gather lip movement data for word pronunciations.
- A notable resource was Collins' 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 were used to ensure 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.

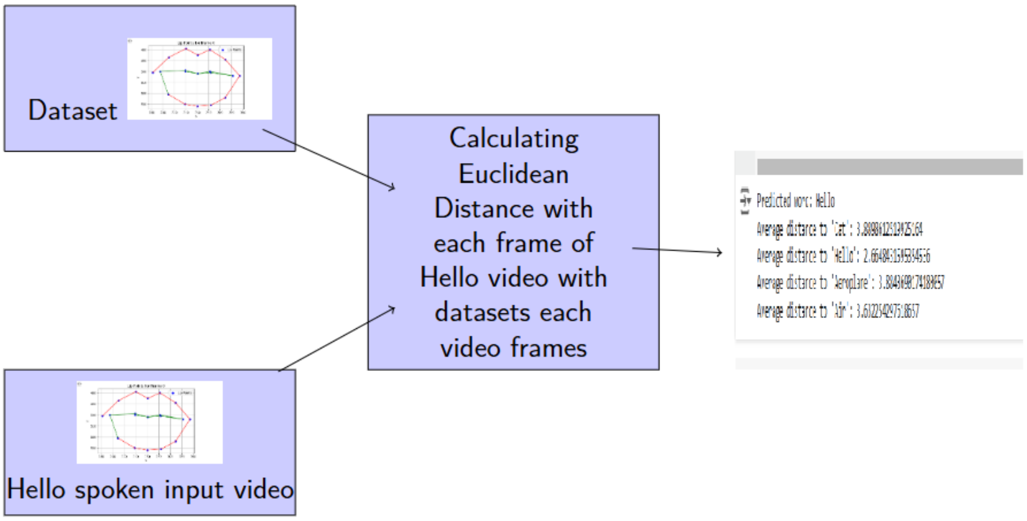
This combination of resources ensures that the dataset is diverse, well-organized, and suitable for lip reading research and applications.

5.2 Result

This diagram 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).


Dataset
Calculating Euclidean Distance with each frame of Hello video with datasets each video frames
Hello spoken input video

```
Predicted word: Hello
Average distance to 'Cat': 3.8898612513925164
Average distance to 'Hello': 2.6648451595534556
Average distance to 'Aeroplane': 3.8843698174189657
Average distance to 'Air': 3.632254297518657
```

Figure 5.2: result

- These extracted lip features are stored for each frame to create a reference for word recognition.

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’: 3.808
Predicted word: Hello.
```

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," *Department of Computer Science, University of Oxford, Oxford, UK; Google DeepMind, London, UK; CIFAR, Canada*, December 2016.

[2] 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," *Cornell University*, January 2020.

[3] S. Fenghour, D. Chen, K. Guo, and P. Xiao, "Lip reading sentences using deep learning with only visual cues," *IEEE Access*, November 2020.

[4] 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*, February 2024.

[5] P. Ma, S. Petridis, and M. Pantic, "Visual speech recognition for multiple languages in the wild," *Imperial College London, Meta AI*, September 2022. Published on September 13, 2022.

[6] S. Fenghour, D. Chen, K. Guo, and P. Xiao, "Disentangling homophemes in lip reading using perplexity analysis," *arXiv*, December 2020. Published on December 15, 2020.

[7] J. H. Yeo, C. W. Kim, H. Kim, H. Rha, S. Han, W.-H. Cheng, and Y. M. Ro, "Personalized lip reading: Adapting to your unique lip movements with vision and language," *arXiv*, September 2024. Published on September 2, 2024.

[8] P. Ma, Y. Wang, S. Petridis, J. Shen, and M. Pantic, "Training strategies for improved lip-reading by koohestani and hadfield," *arXiv*, September 2022. Published on September 29, 2022.

[9] K. Patil, S. Patel, H. Rathod, and A. Siddiqui, "Lip reading: Visual speech recognition using lip reading," *International Research Journal of Engineering and Technology (IRJET)*, April 2022. Published in April 2022.

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**".