TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING

Khwopa College Of Engineering
Libali, Bhaktapur
Department of Computer Engineering

A PROPOSAL ON

LIP READING USING CONVOLUTIONAL NEURAL NETWORKS

Submitted in partial fulfillment of the requirements for the degree

BACHELOR OF COMPUTER ENGINEERING

Submitted by

Chirag Khatiwada	KCE077BCT001
Bishesh Pokharel	KCE077BCT014
Mahim Rawal	KCE077BCT019
Rowel Maharjan	KCE077BCT027

Under the Supervision of
Er.Mukesh Kumar Pokharel
Department Of Computer Engineering

Khwopa College Of Engineering
Libali, Bhaktapur

2023-24

Abstract

Lip reading is the task of decoding text by interpreting the movements, shape and other spatiotemporal facial features from a recorded video clip of a speaker. It invloves mainly studying the movements and configurations in and around the lip area of a speaker. It can be especially useful for people with speech and hearing disabilities so that they can better convey their message to a listener. Besides that, it can be used as a means of comprehending or captioning spoken media from videos recorded in situtations where sound may be difficult to perceive, such as in noisy environments or when the speaker is at a distance.In recent years, technological advancements, particularly in the field of computer vision and machine learning, have led to the development of automated lip reading systems. These systems use algorithms and models to analyze lip movements and convert them into text, providing potential applications in areas such as assistive technologies, human-computer interaction, and surveillance. This proposal outlines a comprehensive research project aimed at advancing the field of lip reading through the integration of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. The proposed hybrid architecture leverages the strengths of both CNNs and LSTMs to enhance the accuracy and efficiency of lip reading systems, addressing the inherent challenges in visual speech recognition.The proposed hybrid architecture will be trained on a comprehensive dataset, including diverse speakers, languages, and environmental conditions, to ensure robustness and generalization. Fine-tuning mechanisms will be implemented to optimize model parameters and improve its adaptability to various lip reading scenarios.We researched on usage of both audio-visual data as input but for implementation of such models would increase the complexity of our project so we have proposed the usage of only visual data as input for timely completion of our project.

Contents

List of Figures

List of Symbols and Abbreviation

ALR	Automatic Lipreading
ASR	Automatic Speech Recognition
Bi-GRU	Bi directional Gated recurrent unit
CNN	Convolutional Neural Networks
SAT	Speaker Adaptive Training
RNN	Recurrent Neural Network

Chapter 1

Introduction

1.1 Background

People often communicate through hearing and vision, that is, through voice signals and visual signals. Speech signals often contain more information than visual signals, so many studies have focused on Automatic Speech Recognition (ASR). Although automatic speech recognition (ASR) technology is mature, there are still some unsolved problems, such as how to accurately identify what the speaker is saying in a noisy environment. Lipreading is a visual speech recognition technology that recognizes the speech content based on the motion characteristics of the speaker's lips without speech signals. Therefore, lipreading can detect the speaker's content in a noisy environment, even without a voice signal. Machine learning methods have a great impact on social progress in recent years, which promoted the rapid development of artificial intelligence technology and solved many practical problems. Automatic lip-reading technology is one of the important components of human–computer interaction technology and virtual reality (VR) technology. It plays a vital role in human language communication and visual perception. This project investigates the task of speech recognition from video without audio. The input data to our algorithm is sequences of still images taken from frames of video. We use models to output one of 10 words that are spoken by a face in the input images. We explore and combine a number of different models including CNNs, RNNs and existing publicly available pretrained models to assist in mouth recognition.

1.2 Problem Statement

At present, the ASR can reach a very high recognition rate without severely damaging the speech signal and also can be used in many practical fields. Visual speech recognition is a technology that recognizes the speech content by lip movement characteristics on no speech signal. The information received by the voice channel is two dimensional. Compared with the one-dimensional voice information received by the voice channel, the visual information often contains more redundant information. So visual speech recognition has always been a difficult problem to solve. Visual speech technology is also known as Automatic Lipreading (ALR), which infers the speech content according to the movement of lips in the process of speaking. In real world, there are people with hearing impairment. They communicate through sign language or observing through people's lip movements. But gesture language has problems such as being difficult to learn and understand, and inadequate expression skills. Therefore, ALR technology can help people with hearing impairment communicate with others better to some extent. Also in noisy environments, the speech signal is easily interfered with by the surrounding noise, resulting in the reduction of recognition rate. However, the visual information needed for ALR will not be affected, so ALR can improve the recognition effect of speech recognition in noisy environments. In the field of security, first of all, with the popularity of face recognition technology, there are many attacks against face recognition system, such as photos, video playback, and 3D modeling, etc. Adding lip features can further improve the security and stability of the security system. In the field of vision synthesis, traditional speech synthesis can only synthesize a single voice, and lipreading technology can generate high-resolution speech scene video of specific people. Besides, in sign language recognition, lip movements are also combined to better understand the content of sign language or improve the accuracy of sign language recognition.

1.3 Objectives

The main aim of this project is:

- To help hearing impaired people.
- To improve the accuracy and stability audiovisual applications.
- To improve data extraction from noisy environments.

Chapter 2

Literature Review

The introduction of Artificial Intelligence has greatly enhanced the interaction capabilities of people with hearing and speech related disabilities and impairments. With there being millions of people suffering from these disabilities, the use of suitable lip reading applications and models can allow them to engage in conversations, thus making them be connected to the real world.However, developing such a model is challenging for both designers and researchers. These models should be well designed, perfected, and integrated into smart devices to be widely available to all people in need of speech understanding assistance.

Lip reading can be conducted on the letter,word,sentence,digit or phrase level. It can also be based on video, voice, video with voice or video without voice as input. There have been studies focused on speaker-independent lip reading by adapting a system using the Speaker Adaptive Training (SAT), which was initially used in the speech recognition field. [2]. Research has also been done towards developing an audio-visual speech enhancement framework that operates at two levels: a novel deep-learning based lip-reading regression model and an enhanced, visually-derived Wiener filter for estimating the clean audio power spectrum. [1] The paper [8] uses CNN and Bi-GRU (Bi directional Gated recurrent unit). According to this algorithm, the system is decomposed into two blocks. The first block consist of lip segmentation. The mouth region is extracted using Haar Cascade classifier. Then hybrid active contours model with an improved of the edge by a designed filter is proposed. The second block consists to classify word lip-reading. First, deep convolutional neural network (CNN) is applied to extract frame features from videos who take the results of first block as inputs. Second, the Bi-GRU with two hidden layers is followed by a global average pooling layer. Finally, the word classification results are obtained by Softmax layer. Using segmented lip inputs can yield stronger features, and vastly improve recognition performance. The paper [3] proposes a novel lip-reading driven deep learning approach for speech enhancement that leverages the strengths of deep learning and analytical acoustic modeling. The proposed audio-visual speech enhancement framework operates at two levels: a novel deep learning based lip-reading regression model and an enhanced, visually-derived Wiener filter for estimating the clean audio power spectrum. This discusses the challenges of lipreading and presents LipNet, a model that can map a sequence of video frames to text, trained entirely end-to-end. On the GRID corpus dataset, LipNet achieves 95.2% accuracy in sentence-level, overlapped speaker split tasks.

The [5] uses the MIRAVL-VC1 dataset which outperforms previous datasets in various aspects. It uses modified form of residual network architecture and uses various techniques in data processing, augmentation and visualization to overcome the scarcity of data and improve the performance. Possible insight into possible improvements and future work in expanding the scale and generalization of the model.The paper [4] attempts to use phonemes as a classification schema for lip-reading sentences to explore an alternative schema and to enhance system performance. In the paper [7], they try to improve the accuracy of speech recognition in noisy environments by improving the lip reading performance and the cross-modal fusion effect.The experimental results show that their method could achieve a significant improvement over speech recognition models in different noise environments.The paper [9] makes GhostNet better by creating Efficient-GhostNet. It improves performance with fewer parameters using a new method for communication within the network, making it more efficient.The improved Efficient-GhostNet is used to perform lip spatial feature extraction, and then the extracted features are inputted to the GRU network to obtain the temporal features of the lip sequences, and finally for prediction.The article [6] looks closely at different deep learning methods for lipreading, discussing how they are structured. It also lists various lipreading databases, providing details about them and the techniques used. The paper ends by talking about the challenges in current lipreading methods and suggesting possible future research directions.Lastly we studied about usage of extraction of audio as well as visual features from a video for predicting the spoken sentence/word in [1].The usage of both the features definitely increases the accuracy of the result but that also increases the complexity of the project as models to extract both audio and visual features are to be trained and the database to be used requires both audio and video essence which is more complex to collect and process than the database containing only visual features.

Chapter 3

Feasibility Study

3.1 Technical Feasibility

The technical feasibility of a lip-reading application falls on usage of advanced image and video processing techniques in order to capture and process and analyze clear lip movement. The system should be seamlessly running with speech-to-text capability. And usage of natural language processing is a must in order to increase the transcription ability. Real-time processing capability and consideration of hardware requirements is a must for bringing the system into practical use. Through the use of large and varied dataset using an effective model and also keeping the functioning of system in different lighting conditions in mind a consistent user experience is expected.

3.2 Economic Feasibility

The economic feasibility of a project verifies project's financial viability by examining a project's costs, benefits, and risks to determine whether it is financially viable and worthwhile to pursue. For this system, the economic feasibility would involve the cost of training and fine-tuning multiple image and video processing models, implementation of the project its software development keeping required hardware in mind along with its maintenance cost.

3.3 Schedule Feasibility

The scheduling feasibility of a project is an assessment of whether the project and be completed within a specified time frame maintaining quality standard. There are several factors which could impact the schedule, including the availability of the resource materials the project is estimated to take a little over then 3 months. There is expectations to finish the documentation and testing of the system in the specified time frame.

Chapter 4

Project Methodology

4.1 Software Development Model

The Agile model is an adaptable and iterative software development process that puts the needs of the client and flexibility first. It breaks the project up into manageable chunks known as sprints or iterations, enabling regular review and modification. Close collaboration between cross-functional teams results in functional software at the conclusion of each iteration. This cycle of iteration guarantees prompt reaction to evolving needs, promoting ongoing enhancement and contentment for the client. The Agile Manifesto's concepts of agile development include a strong emphasis on people and their relationships, functional software, customer collaboration, and adapting to change. In dynamic development contexts, the Agile approach has gained widespread adoption as a framework that encourages efficiency and reactivity.

Figure 4.1: Agile Model

4.2 Description of working flow of proposed system

1. **Input (Video Frames):**
   The input for a lip reading system typically consists of sequences of video frames, where each frame captures the movement and shape of the speaker's lips over time. The primary input is a video recording of a person speaking. Generally,the video contains a sequence of frames, each showing the speaker's lips in different positions.The video is processed to extract individual frames. The number of frames per second (fps) is a crucial parameter that influences the temporal resolution of the input data. Higher fps values can capture more detailed lip movements. The input video for this project has 25fps (frames per second).

2. **Pre-processing:**
   Preprocessing in the context of a lip reading system involves a series of steps applied to the input video frames to enhance their quality, reduce noise, and ensure consistency. Typically, these steps aim to create a standardized and more informative input for the neural network.Normalization and enhancement techniques may be applied to the frames. This could include color normalization, resizing, or other transformations to ensure consistency and improve the model's robustness to variations in lighting and camera settings.The purpose of preprocessing is to prepare frames for optimal model input.Overall, preprocessing plays a crucial role in optimizing the input data, enabling the lip reading model to better learn and interpret the temporal and spatial features of lip motion.

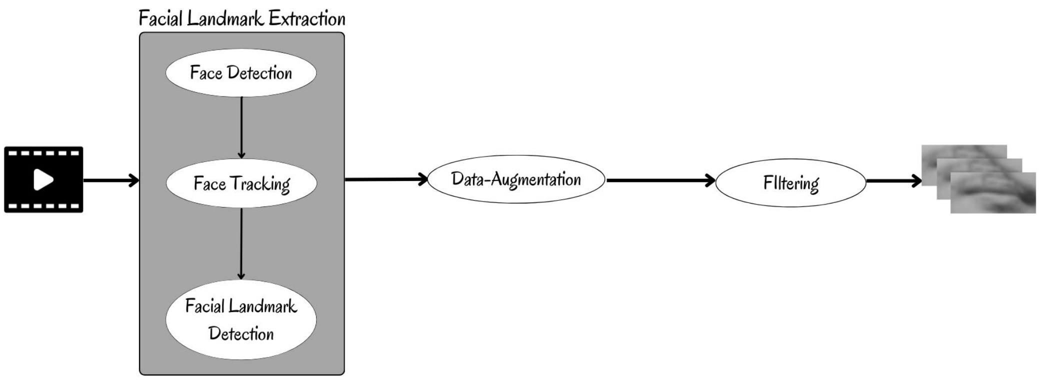

Figure 4.2: Pre-Processing

3. **Convolutional Neural Network (CNN):**
   In the lip reading project, Convolutional Neural Networks (CNNs) are utilized for spatial feature extraction from video frames capturing lip movements. CNNs excel in recognizing patterns within image data by employing convolutional layers and filters. The initial layers detect low-level features like edges and shapes, while deeper layers abstract higher-level representations related to the lips' spatial structures.The convolutional process involves sliding filters across the input image, enabling the network to identify spatial hierarchies and intricate patterns within the lip region. Pooling layers follow, reducing spatial dimensions and retaining essential information. These spatial features are crucial as they represent the distinctive visual cues of lip shapes and movements. By integrating CNNs into the lip reading system, the model becomes adept at discerning spatial intricacies, laying the foundation for subsequent processing stages that capture the temporal aspects of lip gestures using recurrent layers.

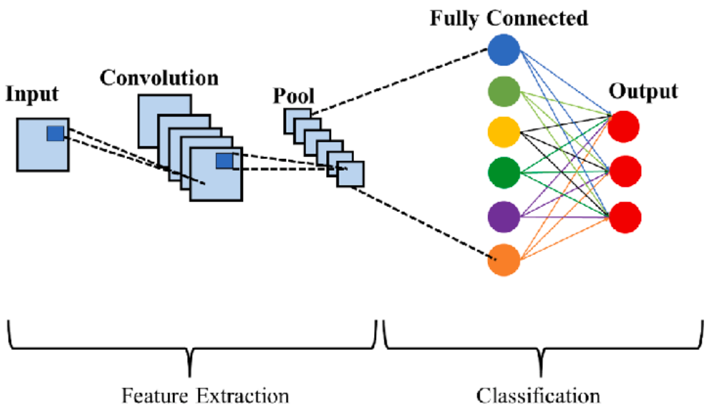

Figure 4.3: Convolutional Neural Network

4. **Feature Extraction:**
   Feature extraction is a crucial step in the lip reading system, encompassing the identification and representation of relevant information from the spatial features obtained by the Convolutional Neural Network (CNN). This process aims to distill essential characteristics that best capture the discriminative aspects of lip movements for subsequent analysis. In lip reading, features might include the shape, contour, and texture of the lips across frames. Feature extraction techniques often involve pooling layers that downsample spatial dimensions while retaining significant information. These features serve as a compact representation of the input data, emphasizing the most salient aspects relevant to lip dynamics. Moreover, temporal feature extraction follows, where recurrent layers like Long Short-Term Memory (LSTM) networks process the sequential nature of the features. This captures the dynamic evolution of lip movements over time. Efficient feature extraction is pivotal, as it enhances the model's ability to discern nuanced patterns in lip articulation, ultimately contributing to the accurate interpretation of spoken language from visual cues.

5. **Recurrent Neural Network (RNN):**
In the lip reading project, Recurrent Neural Networks (RNNs) are employed for temporal feature extraction, recognizing the sequential nature of lip movements captured in video frames. Unlike traditional neural networks, RNNs are equipped to maintain a memory of past inputs, making them well-suited for tasks where temporal dependencies play a crucial role. For lip reading, RNNs, particularly Long Short-Term Memory (LSTM) networks, are adept at learning and representing temporal patterns in the sequence of features obtained from the preceding Convolutional Neural Network (CNN) layers. LSTMs have memory cells that can selectively store and retrieve information over time, allowing the network to capture long-term dependencies in lip gestures.The RNN processes the sequence of spatial features, extracting temporal nuances such as the duration and timing of specific lip movements. By doing so, it enables the model to understand the dynamic evolution of lip articulation and enhances its capacity to decode spoken language accurately from visual cues. Temporal feature extraction with RNNs is vital for comprehending the temporal context inherent in lip reading tasks.

6. **Long Short-Term Memory (LSTM) Cells**
Long Short-Term Memory (LSTM) cells are a specialized type of recurrent neural network (RNN) architecture designed to address the challenges of capturing and learning long-range dependencies in sequential data. LSTMs play a crucial role in temporal feature extraction for tasks like lip reading.In the context of lip reading, LSTMs excel at modeling the temporal dynamics of lip movements over a sequence of video frames. Each LSTM cell contains a memory cell, input gate, forget gate, and output gate. The memory cell serves as a storage unit that can selectively retain or discard information over time. The input gate regulates the inflow of new information, while the forget gate manages the removal of unnecessary information from the memory cell. Finally, the output gate controls the information that is passed to the next time step or to the subsequent layers of the neural network. LSTM cells are particularly effective at mitigating the vanishing and exploding gradient problems often encountered in traditional RNNs, enabling them to capture and remember patterns in sequential data over extended time periods. In lip reading, LSTMs contribute to understanding the nuanced temporal relationships in lip gestures, facilitating accurate interpretation of spoken language from visual cues.

7. **Temporal Feature Extraction:**
Temporal feature extraction is a critical phase in the lip reading system, involving the capture and representation of dynamic patterns in sequential data, such as the temporal evolution of lip movements over video frames. In the context of lip reading, this process occurs after the spatial features are extracted using Convolutional Neural Networks (CNNs) and before the final classification stage. Temporal feature extraction is often facilitated by recurrent neural networks (RNNs), and specifically, Long Short-Term Memory (LSTM) cells. These specialized cells are designed to model and capture dependencies in sequential data, addressing issues like vanishing gradients and enabling the network to retain information over extended time

intervals. During temporal feature extraction, the LSTM cells process the sequence of spatial features obtained from the lip regions across multiple frames. The memory cells within LSTMs maintain contextual information, allowing the model to discern the timing, duration, and patterns of specific lip movements. This step is crucial for understanding the temporal context of spoken language, providing the neural network with the ability to decode and interpret the phonetic information conveyed by the speaker's lips accurately.

8. **Integration Layer:**
   The integration layer in a lip reading system serves as the point where spatial and temporal features are combined to create a comprehensive representation of lip movements. After Convolutional Neural Networks (CNNs) extract spatial features and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) cells, capture temporal dynamics, the integration layer merges these features. This layer harmonizes the spatial and temporal aspects, providing a unified and enriched representation of the lip gestures over time. The integrated features are then forwarded to subsequent layers for final processing and classification. The role of the integration layer is pivotal in ensuring that the model effectively combines both spatial and temporal information, optimizing its ability to interpret spoken language from visual cues accurately.

9. **Fully Connected Layers:**
   Fully Connected Layers, also known as dense layers, are a crucial component in neural networks, including those used for lip reading. Following feature extraction and integration, fully connected layers process the concatenated or flattened features to make predictions or classifications. Each neuron in a fully connected layer is connected to every neuron in the previous layer, allowing the network to learn complex relationships and patterns. In the context of lip reading, these layers play a vital role in mapping the integrated spatial and temporal features to the output classes, facilitating the recognition of spoken words or phonemes. The weights and biases in these layers are adjusted during training to optimize the model's ability to accurately predict lip movements and, consequently, spoken language.

10. **Output (Text Prediction):**
    The output in a lip reading system represents the final result or prediction made by the neural network based on the processed input data. In the context of lip reading, the output typically consists of transcriptions, phonemes, or textual representations of the spoken words or phrases. The output layer of the neural network produces a probability distribution over the possible classes or labels, and the final prediction is often determined by selecting the class with the highest probability. The goal is to accurately interpret the visual cues from lip movements and convert them into meaningful linguistic information. The output is then compared to the ground truth during training, and the model is optimized to minimize the difference between predicted and actual outputs using a suitable loss function.

Chapter 5

Implementation Plan

5.1 Schedule (Gantt Chart)

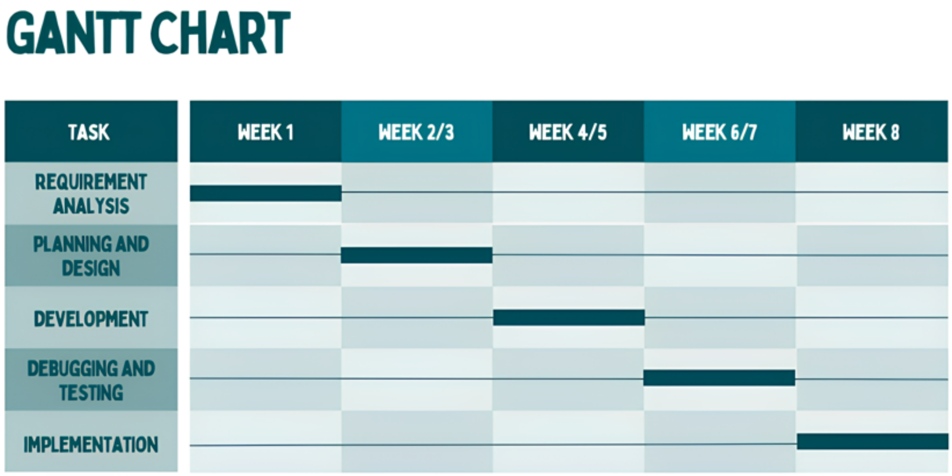

Figure 5.1: Gantt Chart

5.2 Hardware and Software Requirements

5.2.1 Software Requirements

5.2.1.1 Python and Deep Learning

We will be using Python as our programming language for this project. Python is a high-level general-purpose computer programming language often used to build websites and software, automate tasks, and conduct data analysis. It is simple, free, easy to use and highly compatible language consisting of a lot of libraries as well as built-in data structures. Having better library ecosystem, better visualization options, platform independence,and it is well known simplicity, consistency and flexibility, Python has proven itself to be one of the best picks for Artificial Intelligence and Machine Learning. Machine learning is a branch of Artificial Intelligence, where we start with an image and extract it's salient features. Then we create a model that describes or predicts the object on the basis of those features. On the other hand, for Deep Learning, we skip the manual step of extracting the features from the object and directly feed the images into a Deep Learning Algorithm, which then predicts the object. Deep Learning can be used to eliminate the limitations of Machine Learning since it makes it easier to handle complex problems as well as helps us predict through huge amount of data with ease too. Thus, Deep learning is a subset of machine learning which provides the ability to machine to perform human-like tasks without human involvement. It provides the ability to an AI agent to mimic the human brain. Deep learning can use both supervised and unsupervised learning to train an AI agent. Here we will try to utilize technique of Deep Learning and concepts of computerized neural networks using Python for the completion for this project. It serves as the primary programming language for lip reading project, providing a flexible and easy-to-read syntax.

5.2.1.2 OpenCV

OpenCV is an open-source computer vision library that provides a wide range of tools and functions for image and video processing. In our lip-reading project, we use OpenCV to capture and process video frames, apply image preprocessing techniques (such as resizing, filtering, and normalization), and extract relevant features from lip movements such as color or shape information. It also converts the preprocessed frames into a format suitable for input to a PyTorch model.

5.2.1.3 PyTorch

PyTorch is a deep learning framework that is widely used for building and training neural networks. In the context of lip reading, PyTorch can be employed to create and train deep learning models, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs). These models can learn to recognize patterns in lip movements and extract meaningful information for lip reading.

5.2.1.4 Matplotlib

Matplotlib is a plotting library for Python that allows to create a variety of static, animated, and interactive visualizations. In lip reading project, Matplotlib can be used for visualizing different aspects of data and results. For example, we use it to plot training/validation curves, display video frames with overlaid predictions, or create graphs to illustrate the performance of the lip reading model. The Matplotlib is used to visualize the training/validation loss curves during model training.By combining these tools, we can create a comprehensive lip reading system that leverage computer vision, deep learning and visualization to understand and interpret lip movements from video data.

5.3 Functional and Non-Functional Requirements

5.3.1 Functional Requirements

1. **Pre-processing:**
   - Identify and track the face in the video sequence
   - Extract the region of interest (ROI) containing the mouth.
   - Normalize the ROI with respect to size and orientation.
2. **Feature Extraction:**
   - Extract relevant features from the mouth region, such as lip shape, contour, and movement dynamics.
   - Utilize deep learning models (e.g., Convolutional Neural Networks) to achieve accurate feature extraction.
3. **Phoneme Recognition**
   - Based on the extracted features, classify the visual information into corresponding phonemes.
   - Employ deep learning models trained on large lip-to-phoneme datasets.
4. **Sentence Formation**
   - Combine the recognized phonemes into complete words and sentences using language models.
   - Consider contextual information to resolve ambiguities and improve accuracy.

5.3.2 Non-Functional Requirements

1. **Accuracy**

   - The system should achieve a high level of accuracy in translating lip movements to phonemes and subsequently to words and sentences.
   - Specify a target accuracy percentage based on existing benchmarks or project goals.

2. **Real-time performance**

   - The system should process and translate visual information with minimal latency, ideally in real-time.
   - Define an acceptable delay threshold for lip-to-text conversion.

3. **Robustness**

   - The system should perform well under varying conditions, including different lighting, facial expressions, and speakers.
   - Specify the range of scenarios you want the system to handle efficiently.

4. **User Interface**

   - The system should have a user-friendly interface for capturing video, displaying results, and interacting with the system.
   - The system should have a user-friendly interface for capturing video, displaying results, and interacting with the system.

5. **Resource Efficiency**

   - The system should be able to run efficiently on available hardware resources, without excessive memory or processing power requirements.
   - Optimize the model and algorithms to minimize resource utilization without compromising accuracy.

Chapter 6

Expected Outcomes

The Lipreading algorithm is expected to recognize speech better by usage of visual clue given by lip movement especially in noisy places. Anticipated outcomes involve achieving enhanced accuracy through the integration of cutting-edge deep learning architectures and ensemble learning techniques. Furthermore, the objective is to help people with hearing impairments and improve the accuracy in audiovisual applications.

Bibliography

[1] Ahsan Adeel, Mandar Gogate, Amir Hussain, and William M Whitmer. Lip-reading driven deep learning approach for speech enhancement. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(3):481–490, 2019.

[2] Ibrahim Almajai, Stephen Cox, Richard Harvey, and Yuxuan Lan. Improved speaker independent lip reading using speaker adaptive training and deep neural networks. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2722–2726. IEEE, 2016.

[3] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. Lipnet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599*, 2016.

[4] Randa El-Bialy, Daqing Chen, Souheil Fenghour, Walid Hussein, Perry Xiao, Omar H Karam, and Bo Li. Developing phoneme-based lip-reading sentences system for silent speech recognition. *CAAI Transactions on Intelligence Technology*, 8(1):129–138, 2023.

[5] Abiel Gutierrez and Z Robert. Lip reading word classification. *Comput Vision-ACCV*, 2017.

[6] Mingfeng Hao, Mutallip Mamut, Nurbiya Yadikar, Alimjan Aysa, and Kurban Ubul. A survey of research on lipreading technology. *IEEE Access*, 8:204518–204544, 2020.

[7] Dengshi Li, Yu Gao, Chenyi Zhu, Qianrui Wang, and Ruoxi Wang. Improving speech recognition performance in noisy environments by enhancing lip reading accuracy. *Sensors*, 23(4):2053, 2023.

[8] Malek Miled, Mohammed Anouar Ben Messaoud, and Aicha Bouzid. Lip reading of words with lip segmentation and deep learning. *Multimedia Tools and Applications*, 82(1):551–571, 2023.

[9] Gaoyan Zhang and Yuanyao Lu. Research on a lip reading algorithm based on efficient-ghostnet. *Electronics*, 12(5):1151, 2023.