UNIVERSITY OF PETROLEUM AND ENERGY STUDIES

TITLE: LIPNET AI

SOFTWARE REQUIREMENTS SPECIFICATION (SRS)

(MINOR PROJECT - 1)

SEMESTER V

S. No	Students Name	Roll No	Sap Id
1.	Eklavya Gupta	R2142210299	500093960
2.	Aaryak Bhargava	R2142210010	500093996
3.	Aryaman Jain	R2142210158	500093622

BACHELOR OF TECHNOLOGY, COMPUTER SCIENCE

Under the guidance of

Dr. Alok Agarwal

School Computer Science (SOCS), UPES

Bidholi Campus, Energy Acres, Dehradun – 248007

Table of Contents

1. INTRODUCTION

The McGurk effect highlights the significance of lipreading in human communication and speech understanding. It involves the perception of a third phoneme from watching a video of someone speaking a separate phoneme. Lipreading is a challenging ability, especially without context. Most lipreading actuations are latent and difficult to distinguish without

context. Fisher (1968) highlights the confusion of initial consonant phonemes in five categories of visual phonemes (visemes) when watching a speaker's mouth.

Lipreading, a fascinating and demanding area at the interface of computer vision and speech processing, has sparked considerable attention in recent years due to its potential applications in human-computer interaction and accessibility. This Software Requirements

Specification (SRS) specifies the creation of a lipreading model based on pioneering research undertaken at Oxford University in 2016.

The primary goal of this project is to build and enhance the lipreading model provided in the cited research study using Convolutional 3D (Conv3D) neural networks and Bidirectional Long Short-Term Memory (BiLSTM) networks. The suggested methodology seeks to decode visual information from lip movements, making it easier to decipher spoken language using only video input.

The accompanying code snippet outlines the model's architecture, which includes Conv3D layers to capture spatial and temporal information from lip movement data, followed by Bidirectional LSTM layers for successful sequence modeling. The model's design integrates original research principles while also introducing refinements and optimizations for increased efficiency.

This SRS acts as a thorough guide for the development team, outlining the lipreading model's functional and non-functional requirements, system architecture, and performance indicators. We will delve into the specifications, design considerations, and validation methodologies required for the effective implementation of the lipreading system in the following sections.

1.1 Purpose of the Project

The major goal of this project is to enhance the field of visual speech recognition by developing and testing a cutting-edge lipreading model based on influential Oxford University research from 2016. Visual voice recognition, the act of understanding spoken words from face movements, is critical in supplementing standard audio-based speech recognition systems, especially in noisy or difficult acoustic conditions.

This project aims at improving the effectiveness of visual speech recognition through the development and improvement of the lipreading model. The research article will serve as the model's inspiration. It will undergo methodical testing and evaluation to assess the model's accuracy, robustness, and generalizability over a range of datasets and real-world scenarios.

1.2 Target Beneficiary

Researchers and Academics: Researchers and academics working in the domains of computer vision, deep learning, and voice processing will find great use for the developed model and its supporting documentation.

Developers and Practitioners: The model can be used by developers and practitioners working on applications including visual speech recognition, human-computer interaction, and accessibility.

AI Community: Contributions to the broader artificial intelligence community by improving understanding and capabilities of visual speech recognition.

Accessibility Advocates: The improved visual speech recognition model could be used in technologies that improve accessibility for people with hearing impairments.

1.3 Project Scope

The project scope includes the development, enhancement, and systematic testing of a visual speech recognition model based on Oxford University research from 2016. The model, which is built as a Convolutional 3D (Conv3D) and Bidirectional Long Short-Term Memory (BiLSTM) neural network, focuses on decoding spoken language from visual cues captured through lip movements. The scope includes the adaptation and optimization of the existing model architecture to improve the effectiveness of visual speech recognition.

1.4 References

[1] Lipnet AI: End to End Sentence Level Lipreading [1611.01599] LipNet: End-to-End Sentence-level Lipreading (arxiv.org)

[2] Lipytorch C++ Library PyTorch C++ API — PyTorch main documentation

[3] OpenCV OpenCV - Open Computer Vision Library

2. PROJECT DESCRIPTION

2.1 Characteristics of Data

1. Video Input:
   Format: The data consists of video sequences capturing lip movements during speech.

Resolution: Variability in video resolution to simulate real-world conditions.

Frame Rate: Videos may have different frame rates, reflecting natural speaking patterns.

2. Lip Movement Annotations:

Temporal Annotations: Corresponding timestamps indicating the duration of specific lip movements.

Phonetic Labels: Annotations associating lip movements with phonetic representations.

2.2 SWOT Analysis

-Strengths:
1 Based on Proven Research: The model architecture is based on a 2016 Oxford University study, which provides a solid theoretical foundation.
2 Combination of Conv3D and BiLSTM: The model can capture both spatial and temporal features thanks to the integration of Conv3D and Bidirectional LSTM networks, which improves its ability to recognize visual speech patterns.
3 Data Diversity in Training: The inclusion of a wide range of speakers, languages, and environmental conditions in the training data adds to the model's potential for broad applicability.
4 Techniques for Data Augmentation: Data augmentation strategies such as scale, rotation, and synthetic conditions all contribute to the model's robustness.

- Weaknesses:
1 Model Complexity: The complexity introduced by Conv3D and Bidirectional LSTM layers may pose computational and training time challenges.
2 Lip Movements Are Limited: The model's reliance on lip movements may limit its usefulness in situations where other visual or auditory cues are required for accurate speech recognition.
3 Considerations for Ethical Behavior: The importance of paying close attention to ethical considerations, including confidentiality
4 Possibilities for Optimization and Fine-Tuning: There is room for further model architecture optimization and hyperparameter fine-tuning to improve performance and efficiency.
5 Integration of Multiple Modes: In future iterations, the model's overall speech recognition capabilities could be improved by incorporating additional modalities such as audio or facial expressions.

- Opportunities:
1 Integration of Multiple Modes: In future iterations, the model's overall speech recognition capabilities could be improved by incorporating additional modalities such as audio or facial expressions.
2 Implementation in Real-Time: Moving the model to real-time applications could allow it to be used in interactive systems, broadening its potential applications.

- Threats:
1 Restrictive Generalization: The model may struggle to generalize across different datasets or in real-world scenarios that are not adequately represented in the training data.
2 Technological Progress: Rapid advancements in speech recognition technology may outpace the model, making it less competitive in the long run.

3. Availability of Resources: Computational and resource requirements for training and testing may be an impediment, especially for organizations with limited computing infrastructure.

2.3 Project Features

1. Lip Reading in Real Time: Use real-time lip reading to examine and interpret spoken language as it happens.

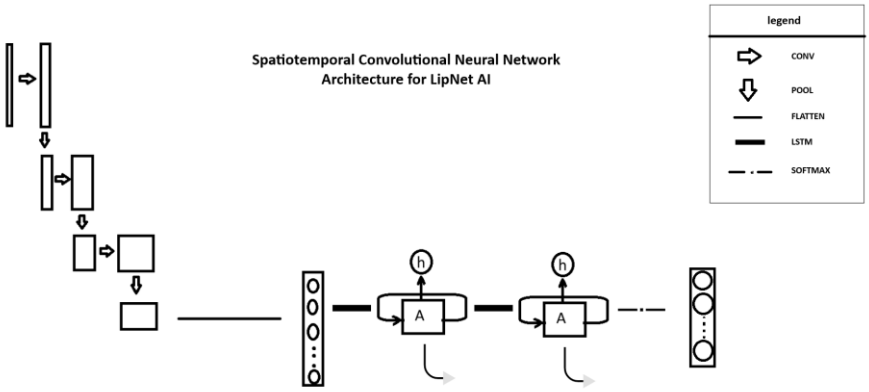
2. Specific Language Customization: Allow users to tailor the model to specific languages or accents in order to increase accuracy in a variety of linguistic circumstances.

3. Documentation and assistance: Provide extensive documentation and support materials to developers and users to aid in understanding, debugging, and integration into multiple applications.

4. Integration of Multiple Modes: Investigate the use of several modalities, such as audio and visual data, to improve the accuracy and robustness of lip reading.

2.4 Design and Implementation Constraints

2.5 Design diagram



3. SYSTEM REQUIREMENTS

3.1 User Interface

For an intuitive experience, the user interface seamlessly integrates video playback and subtitles. Users can easily select a video file that will be displayed within the application. Real-time subtitles overlay the video dynamically, providing a synchronized representation of recognized speech. Controls for play and pause, a

progress bar, and customizable subtitle settings improve user control and personalization. The interface prioritizes accessibility features, and optional features such as language selection and user authentication help to create a comprehensive and user-friendly lipreading application.

3.2 Functional Requirements

: A user-friendly interface should allow users to pick a video file or
.

o: The application must show the selected video and allow users to
needed.

s: Based on the output of the lipreading model, real-time subtitles
nd placed on the video.

Specifications

ystem: Windows 10&11

evelopment Environment: Visual Studio 2019

- Programming Language: C++ 14
- Framework: PYtorch , opencv

3.4 Hardware Specifications

- System Type: 64-bit operating system, x64-based processor
- RAM: 4.00 GB
- Processor: Intel(R) Core(TM) i5-8500H CPU @ 2.50GHz
- SSD: 256 GB

4. NON-FUNCTIONAL REQUIREMENTS

4.1 Performance Requirements

- Scalability: The system should be easily scalable to accommodate a growing user base.

4.2 Security requirements

- Controls of Access: Use access controls to restrict user access to sensitive application functionality and data.
- Video Storage that is Secure: If you save video files, use secure storage techniques to prevent unwanted access or alteration.
- Data Encryption: Ensure that all user data, including video files and any personal information submitted by the user, is encrypted from beginning to finish.
- Communication Security: To safeguard data exchanged between the application and external servers or APIs, use secure communication protocols.

4.3 Software Quality Attributes

- Reliability: The program should create correct subtitles on a consistent basis and handle video playing without frequent faults or crashes.
- Availability: Maintain high availability so that users may use the program and create subtitles without experiencing severe downtime.
- Usability: Create an intuitive and user-friendly user interface that can accommodate users with varied degrees of technical competence.
- Scalability: Make that the lipreading model and application architecture are scalable enough to manage growing user demand and greater datasets.
- User Contentment: Gather customer feedback on a regular basis to measure user happiness and make modifications depending on user feedback.

4.5 Conclusion

It is implementation of oxford university research paper of 2016, main purpose is elvate accuracy and contribute in deep learning community.