ABSTRACT

Our project is to to develop a system to analyse the human behaviour by using deep learning and machine learning technique which can orient itself to read the lip movement from the video and produce an output in the form of text. This system basically does the job of giving the text as an output, when the video as an input is been fed to the system. We here try to get the text as an output in minimum possible time, by maintaining the efficiency of the system as well. The system is developed as per the user inputs and it is tried to be made as economic as possible It is taken care that the ambience of the particular space is not changed.

The system consists of several software modules to serve the purpose. The software modules such as including video processing, training the dataset, feeding the learning model with machine learning algorithms.

The system basically takes video as an input . Video processing on the input video's is been done, after which sequence of data frames are been generated and are given to the learning model. The horizontal and the vertical distance of the lip is calculated and that particular distance of the specific word is been searched and fetched (if found) from the dataset, and it generates the text as an output.

Thus this system intends to help people working in noisy environment to communicate well even in the presence of disturbance.

ACKNOWLEDGEMENTS

This project is the result of the combined efforts of the entire team. However, it would not have been possible without the kind support and help of many individuals and faculty. We would like to extend our sincere thanks to all of them.

We are highly indebted to our guide **Asst. Prof Mahesh S Patil** for his guidance and constant supervision as well as for providing necessary information regarding the project and also for his moral and technical support in completing the project.

We will also like to thank our Head of Department **Dr. Meena S. M** and Vice Chancellor Dr**. Ashok Shettar** for giving the opportunity to do this project.

We would like to express our gratitude towards our parents and member of **KLE TECHNOLOGICAL UNIVERSITY** for their kind cooperation of this project.

SONALI.S.NAIK	**01FE16BCS203**
SUNAINA A.M	**01FE16BCS209**
PRIYANKA NABAPURE	**01FE16BCS243**
SOUMYA KANYAL	**01FE16BCS247**

1. INTRODUCTION

1.1 OVERVIEW OF THE PROJECT

Lip reading also known as lipreading or speechreading, is a technique of understanding speech by visually interpreting the movements of the lips when normal sound is not available. Lipreading can help people who are hearing impaired to cope better with their hearing loss it is not only deaf people who will find it a good skill to have but also anyone who works in a noisy environment finds lip reading an valuable technique to interpret what the speaker is speaking. Lip-reading plays a crucial role in human communication and speech understanding. It is a notoriously difficult task for humans, specially in the absence of context. Hearing impaired people achieve an accuracy of only 12-15% even for a limited subset of words. Hence an important goal is to automate lip-reading by developing an application for lip reading that will help the people working in an noisy environment to communicate with each other even in the constrained environment ,by transforming lip-syncing to text.

1.2 Motivation

1. For millions who can't hear, machine learning can be used to discern speech from silent video clips more effectively than Professional lip readers can.

2. In a noisy environment where voice recognition software tend to underperform , using a approach of lip reading by deep learning technique , increases the efficiency remarkably.

3 .Lip reading using deep learning technique could be implemented in silent dictation , in public spaces.

4. Lip reading using deep learning technique could be used in bus stations, railway stations to communicate directly with the machines for tickets.

1.3 DOMAIN IDENTIFICATION

Computer Vision

Computer vision is concerned with the automatic extraction, analysis and understanding of useful information from a single image or a sequence of images. It involves the development of a theoretical and algorithmic basis to achieve automatic visual understanding.

Deep-Learning Techniques

Deep learning is a machine learning technique that teaches computers to do what comes naturally to humans: learn by example. In deep learning, a computer model learns to perform classification tasks directly from images, text, or sound. Deep learning models can achieve state-of-the-art accuracy, sometimes exceeding human-level performance. Models are trained by using a large set of labeled data and neural network architectures that contain many layers.

1.4 PROBLEM STATEMENT

Develop an application for people working in noisy environment to communicate even under constrained environment ,by transforming lip syncing to the text in hindi language.

1.5 OBJECTIVES

- To develop an application for people working in noisy environment in order to help them in achieving better communication.
- Generation of dataset for regional(Hindi) language.
- To learn Deep-learning techniques and apply it to solve real-world problems.

1.6 LITERATURE SURVEY

1.LipNetSentence-level Lipreading**:**

LipNet introduces the first approach for an end-to-end lip reading algorithm at sentence level. A model that maps a variable-length sequence of video frames to text, making use of spatiotemporal convolutions, a recurrent network, and the connectionist temporal classification loss, trained entirely end-to-end. LipNet is the first end-to-end sentence-level lipreading model that simultaneously learns spatiotemporal visual features and a sequence model. On the GRID corpus, LipNet achieves 95:2% accuracy in sentence-level, overlapped speaker split task, outperforming experienced human lipreaders and the previous86:4% word-levelstate-of-the-art accuracy.

These approaches obtained impressive results (over 70% word accuracy) for tests performed with classifiers trained on the same speaker they were tested on. But performance was heavily damaged when trying to lip read from individuals not included in the training set. Lip detection in males with moustaches was also more difficult and, therefore, the performance on such cases was poor. Hence, the feature engineering approaches, while an improvement, ultimately failed to generalise well.

2. Lip Reading in the Wild:

Author: J. S. Chung, A. Zisserman

Classify temporal sequences with excellent results. On the 333-word test set, accuracy of 65.4% was achieved, which exceeds state-of-the-arton multiple datasets.

3.Lip Reading Word Classification:

Authors: Joon Son Chung, Andrew, Oriol Vinyals, Andrew Zisserman.

Usage of the MIRACL-V1 dataset containing videos of ten people speaking ten words. We pre-process the data by using existing facial recognition software to detect and crop around the subject's face in all frames of the video and then use the sequence of frames as in put to the model. We explore a CNN + LSTM Baseline model, a Deep Layered CNN + LSTM model, an Image Net Pretrained VGG-16 Features + LSTM model, and a Fine-Tuned VGG-16 + LSTM model validation accuracy of 79% and a test accuracy of 59% on our best model.

CHAPTER 2

PROPOSED SYSTEM

2.1 Overview

The process of design and implementation involves continual tradeoffs between cost and performance. Quantifying the performance implications of various alternatives is central to this process. It also is extremely challenging. In the case of existing systems, measurement data is available. In the case of evolving systems, contemplated modifications often are straightforward (e.g. a new CPU within a product line) and limited experimentation may be possible in validating a baseline model. In the case of proposed systems, these advantages do not exist. For this reason, it is tempting to rely on seat-of-the-pants performance projections, which all too often prove to be significantly in error. The consequences can be serious, for performance, like reliability, is best designed in, rather than added on.

2.2 Description of proposed system with simple block diagram

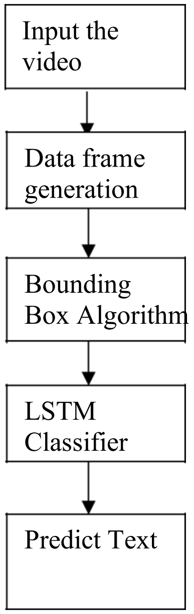


Figure 1 Block Diagram of Lip Reader

Figure 1 represents the block diagram of the system. The system basically takes video as an input. Video processing on the input video is done, after which sequence of data frames are been generated and are given to the learning model i.e LSTM. The vertical distance of the lip is calculated and sequence of heights are trained on the LSTM model and the word is predicted.

2.3 Description of Target users

People working in Noisy environment (where hearing ability is disturbed) so that they can communicate well with the speaker even in the presence of noisy conditions.

It can be also extended as following:

1. People with hearing disorder could easily communicate with normal people as well as people with the same disorder. This Application could be extended in the following fields:
2. In bus stations, railway stations to communicate directly with the machines for tickets.
3. In necessary silent dictation.

2.4 Advantages/Applications of the proposed system

- For millions who can’t hear, machine learning can be used to discern speech from silent video clips more effectively than professional lip readers can.
- In a noisy environment where voice recognition software tend to underperform, using a approach of lip reading by deep learning technique , increases the efficiency remarkably.
- Lip reading using deep learning technique could be implemented in silent dictation ,in public spaces.
- Lip reading using deep learning technique could be used in bus stations, railway stations to communicate directly with the machines for tickets.

CHAPTER 3

SOFTWARE REQUIREMENT SPECIFICATION

3.1 Overview of SRS

A software requirements specification (SRS) is a description of a software system to be developed. It lays out functional and non-functional requirements, and may include a set of use cases that describe user interactions that the software must provide.

It is very important to develop a SRS listing out the requirements and how the requirements are going to be fulfilled. It helps the team to save upon their time as we will be able to comprehend how we are going to go about the project. Doing this also enables the team to find out about the limitations and risks early on.

3.2 Requirements Specifications

3.2.1Functional Requirements

- User shall be able to turn on application.
- User shall have mounted camera, to capture the video of speaker.
- User shall be able to connect his application to the camera device.
- User shall be able to view the text.
- User shall have back up of the text.

3.2.2 Use case Diagrams

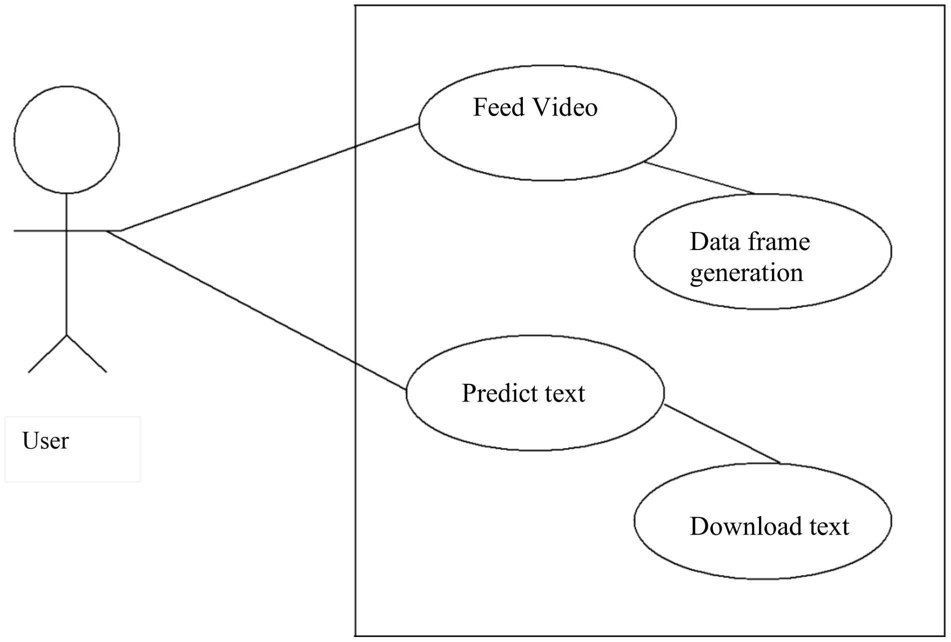


Figure 2. Use Case Diagram for Lip Reader

3.2.3Use Case Description using scenarios

Problem description: Conversion of lip movements of the input video to text.

Pre-Conditions:

- Start the System.
- Start the application.
- The application shall be in the state to ready accept the input video.
- Give the input i.e lip movement from the video.
- The input video quality shall be HD or of other high quality

Basic Flow of Events:

- Input
- Video Processing
- Images
- Calculating the distance
- Input to the Algorithm on database
- Output : Speech in textual form

Post Condition

- Display the speech in the video in the textual form.
- The text must be in English.

Main Success Scenario:

1. Begins when the system is ready to accept the video.
2. User inputs the video.
3. The system analysis the video for the lip movement and displays the status.
4. If the input video has lip movements then the process continues.
5. The vertical and the horizontal distance between the upper lip and the lower lip is calculated.
6. The algorithm analysis the database wrt the above calculated distance.
7. Output is displayed in the text form.
8. The output is in English language.
9. The system provides the option to download and copy the output,i.e text.

Exceptional Scenario:

2.a. If "input, is not video" then display "Invalid input".

3.a.If the video does not have lip movement then, display "Enter valid input".

6.a. If the word in the video does not match the word in the database then, display "Sorry, I couldn't recognise!"

3.2.4 Non-Functional Requirements

- User shall have the following configurations to run the application
  1. Ram-Size: atleast 2 GB.
  2. Storage: atleast 1 GB.
  3. Wi-Fi-Connectivity:2-3m.
  4. BluetoothConnectivity:2-3m.
  5. Camera: atleast 10 mega pixels.
- User shall have the following configurations in wearable device.
  1. Wi-Fi-Connectivity:2-3m.
  2. BluetoothConnectivity:2-3m.
  3. Camera: atleast 2-3m.
- User shall have the response for the given input video within 30 msec.
- User shall be at a maximum distance of 15-20m.

3.3 Software and Hardware Requirement

Specification Software Requirements:

- Video related deep learning libraries
- Software: Python 3.4 or 3.6
- Tool: Anaconda 2.7
- Language**:** Python.
- Platform: Spyder

Hardware Requirements:

- The processor type should be 64 bits or 16 bits and the speed should be minimum of 1.83GHz.
- The size of RAM should be more than 2GB.
- Deep learning capable machine like NVIDIA GPUs with at least 8 GB or more RAM

3.4 Acceptance test plan

3.4.1 Test Cases

User shall feed the video

- The input video shall be of time size max 30 sec and min .1 sec
- The input video quality shall be HD or of other high quality
- The User shall pause the video, resume the video, rewind the video, and forward the video.

Test case Id	Input Description	Expected output	Actual Output
1	Video of time size less than 60 sec and greater than .1 sec	Data Frames	
2	Video of time size more than 60 sec	Invalid Video	
3	Video of time size 60 sec	Data Frames	
4	Video with pixel 144p	Insert a good quality video	
5	Video is paused	Extraction of Data frame is stopped	
6	Video containing more than one lip pair	Data Frames	
7	Forwarding the video until the video size	Data Frames	

User shall View the text

• The output shall be in the form of Text.

• The output text shall be in

Hindi Language.

• The output text shall be displayed after video processing is done.

• The user can download the output, i.e text.

• The user shall copy the output, i.e. text

Test case Id	Input Description	Expected output	Actual Output
1	Video with lip movement	Corresponding text	
2	Video do not consisting face	Unrelated Video	
3	Video with no sound	Corresponding Text	
4	Video with the word not in the dataset	Word not found	
5	Video with more than one lip pair	No proper recognition	
6	Video having lip movement of hindi language	Word not found,hence no text	
7	No Sequence of distance calculated for Data frames of input video	No Text	

CHAPTER 4

SYSTEM DESIGN

4.1 Overview

System design is the process of defining the elements of a system such as the architecture, modules and components, the different interfaces of those components and the data that goes through that system. It is meant to satisfy specific needs and requirements of a business or organization through the engineering of a coherent and well-running system.

Systems design implies a systematic approach to the design of a system. It may take a bottom-up or top-down approach, but either way the process is systematic wherein it takes into account all related variables of the system that needs to be created—from the architecture, to the required hardware and software, right down to the data and how it travels and transforms throughout its travel through the system. Systems design then overlaps with systems analysis, systems engineering and systems architecture.

4.2 Architecture of the system

Batch sequential

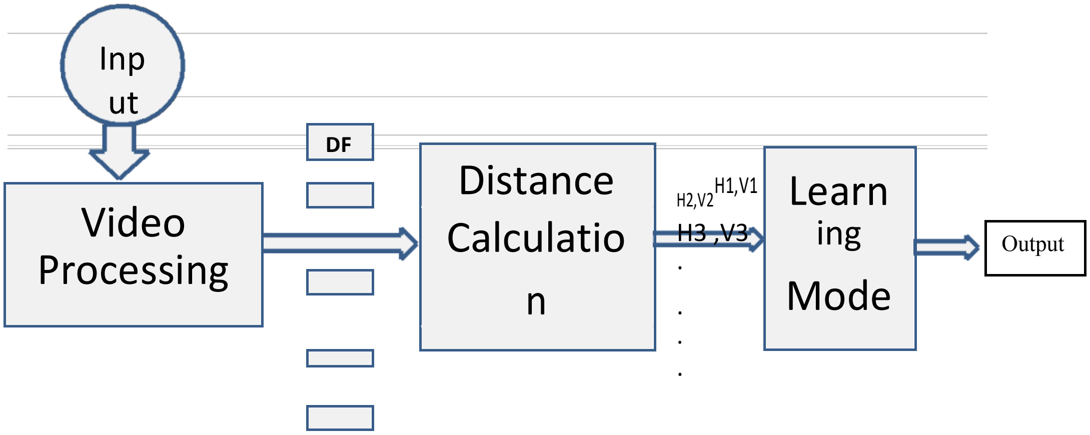


Figure 3. Architecture diagram

Figure 3 represents the of the Architecture diagram (Base Sequential)**.** The system here is divided into sub-systems where output of every subsystem acts as an input data to the next subsystems and every subsystem is independent of every other subsystem.

In Lip Reader the input(video) is fed to the video processing system. The output of this system is the dataframe. The dataframes generated are given to the next system i.e. distance calculation system. Output of this system is {height, width} of the lip region. Lastly, the height and width sequences are trained on the Learning model(LSTM).

4.3 Level 0 DFD

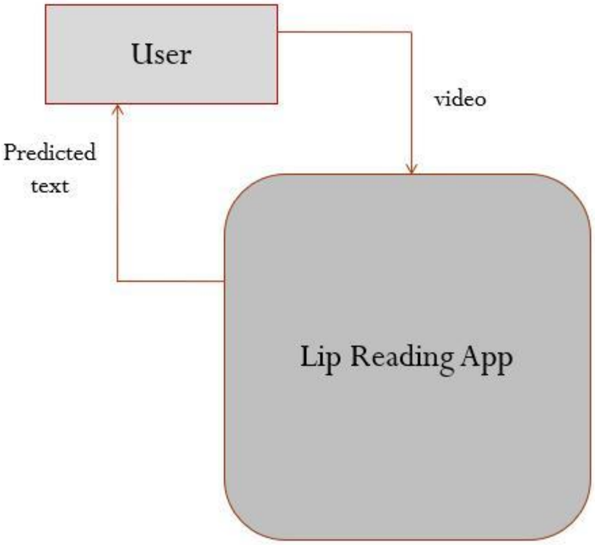


Figure 4. DFD Level 0 for Lip Reader

Figure 4 represents the level 0 Data Flow Diagram. The context diagram here represents all external entities that may interact with a system. Here, the system is at the centre with no details of its interior structure, surrounded by all its interacting systems, environments and activities. Our objective here is to focus on external factors and events that should be considered in developing a complete set of system requirements and constraints.

4.4 Detailed DFD for the proposed system

Level 1 DFD

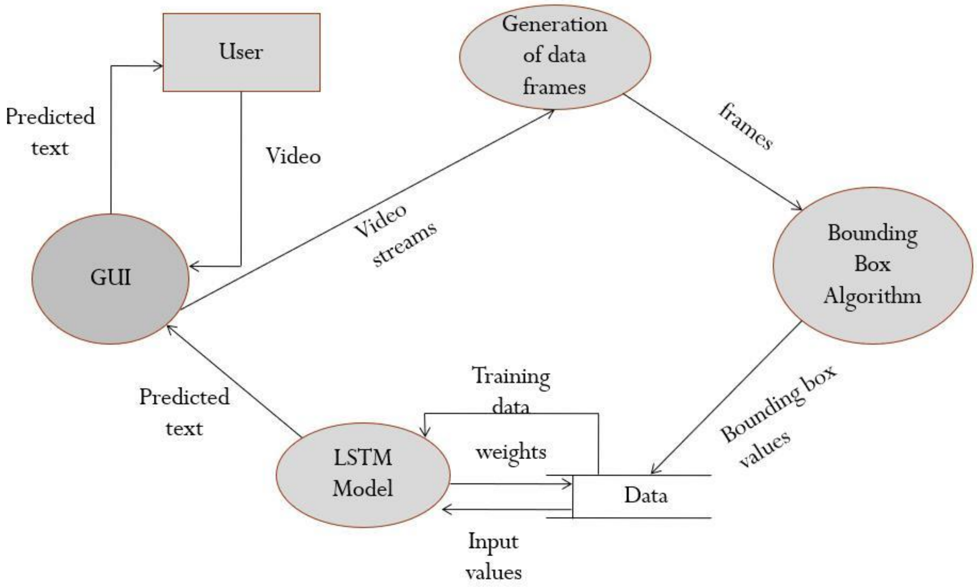


Figure 5. Detailed DFD for Lip Reading System

Figure 5 represents the detailed Data Flow Diagram for Lip Reader. Here it represents the graphical flow of the data through the information system. DFD diagram consists of 5 modules namely User, GUI, Generation of data frames, Bounding Box Algorithm and LSTM Model.

4.6 Activity diagram

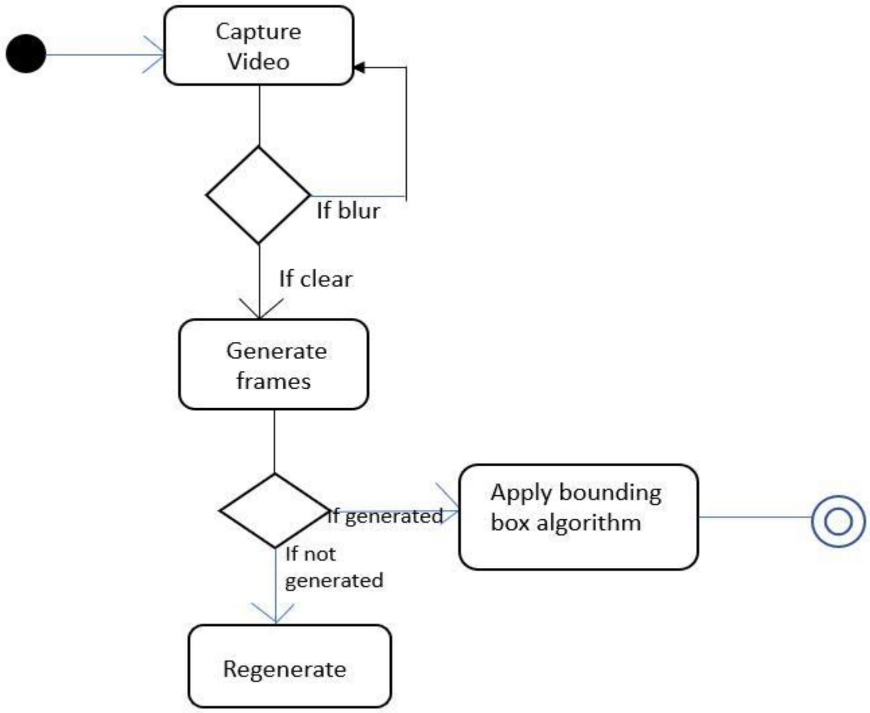


Figure 6. Activity Diagram for Video Processing

Figure 6 represents video processing. The input fed is of the form of video, from which the data frames are generated. Bounding Box algorithm is applied for each frame generated to detect the bounding box.

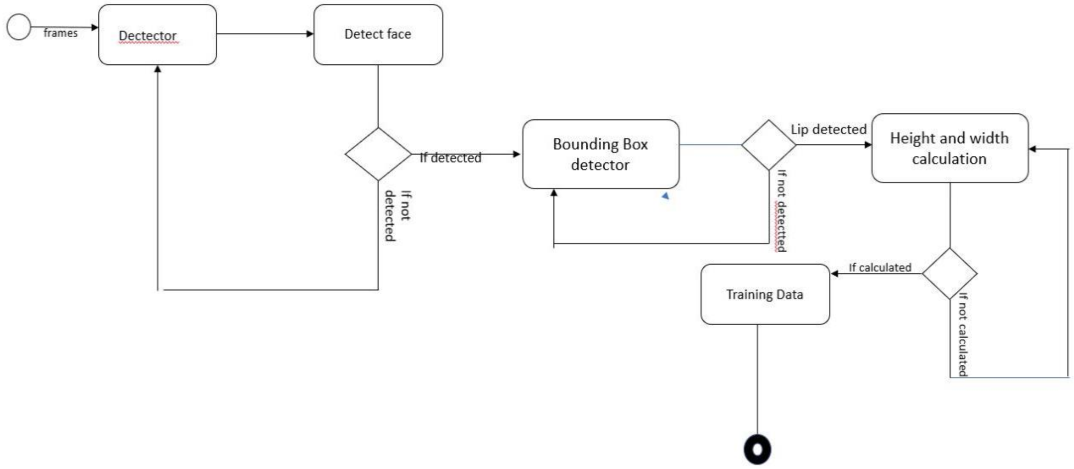


Figure 7. Activity Diagram for Calculation of height and width

Figure 7 represents the activity diagram for the algorithm, for every data frame generated from the previous step,the lip part is detected from the entire face, and the corresponding height and width values of the lip movement is calculated with the help of bounding box algorithm. The height and width values are collected and given for training.

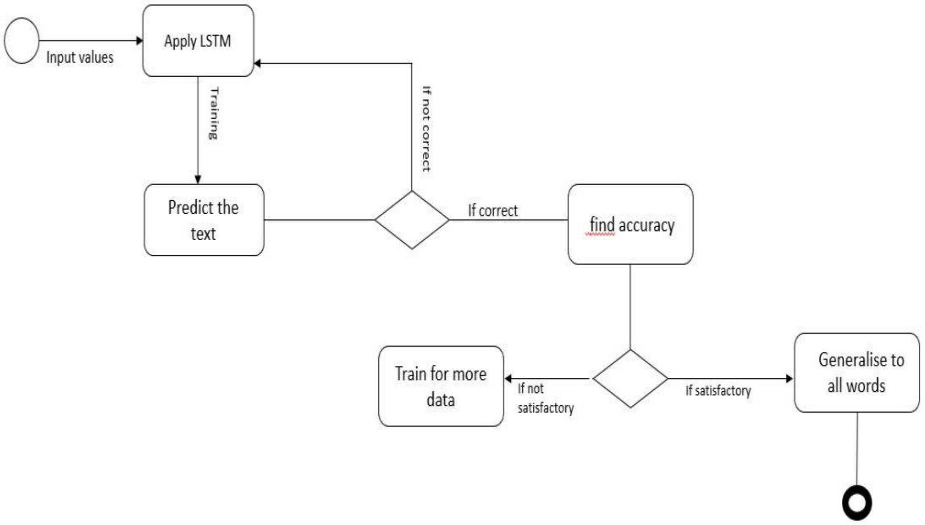


Figure 8. Activity Diagram for LSTM Classifier.

Figure 8 represents that the stored values are further trained using LSTM Classifier and the corresponding text for each sequence is predicted and accordingly trained for many different words.

CHAPTER 5

IMPLEMENTATION

5.1 PROPOSED METHODOLOGY

In order to pronounce a word, series of lip movements are required. So, the video is converted into a series of images(frames). Each frame represents a movement. For each frame generated only the lip part is extracted and the horizontal and vertical distance of the lip is calculated. So, a series of distances are obtained for each word. In this fashion we obtain a dataset consisting of series of height and width for each word. The dataset obtained is further trained using LSTM Model and for the corresponding sequences the word is predicted.

5.2 Modules

1.Data Frame generation

Video processing is a method to perform operations on an video, in order to get sequences of dataframes or the images.Similarly the input to our system is video for the respectives video the frames are generated.For each second 30 frames are generated.

2.Bounding Box detection

The frames obtained from the video processing are used to detect the lip part ,initially the face is detected ,further from each frame the lip part is extracted.

3.Calculation of height and width

The lip part extracted is used to calculate the height and width i.e each word spoken will have various pattern that it follows, for each pattern the corresponding height and width is calculated and these sequence act as dataset.

4.LSTM Classifier

The sequences obtained for each word are trained using LSTM Model. Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections that make it a "general purpose computer". It cannot only process single data points (such as images), but also entire sequences of data (such as speech or video).

CHAPTER 6.

TESTING

6.1 Overview

Software testing is an activity to check whether the actual results match the expected results and to ensure that the software system is detect free. It involves execution of a software component or system component to evaluate one or more properties of interest. Software testing also helps to identify errors, gaps or missing requirements in contrary to the actual requirements

A Test Plan is a document describing software testing scope and activities. It is the basis for formally testing any software/product in a project. It is a document describing the scope, approach, resources and schedule of intended test activities. It identifies amongst others test items, the features to be tested, the testing tasks, who will do each task, degree of tester independence, the test environment, the test design techniques and entry and exit criteria to be used, and the rationale for their choice, and any risks requiring contingency planning. It is a record of the test planning process.

6.2 Test Cases

User shall feed the video

- The input video shall be of time size max 30 sec and min .1 sec
- The input video quality shall be HD or of other high quality
- The User shall pause the video, resume video, rewind the video,and forward video.

Test case Id	Input Description	Expected output	Actual Output
1	Video of time size less than 30 sec and greater than .1 sec	Data Frames	Extracted the dataframes
2	Video of time size more than 30 sec	Invalid Video	Invalid video
3	Video of time size 30 sec	Data Frames	Extracted the dataframes
4	Video with pixel 144p	Insert a good quality video	No dataframes found
5	Video is paused	Extraction of Data frame is stopped	No dataframes found
6	Video containing more than one lip pair	Data Frames	Extracted the dataframes
7	Forwarding the video until the video size	Data Frames	Extracted the dataframes

User shall View the text
- The output shall be in the form of Text.
- The output text shall be in Hindi Language.
- The output text shall be displayed after video processing is done.
- The user can download the output, i.e text.
- The user shall copy the output,i.e text

Test case Id	**Input Description**	**Expected output**
1	Video with lip movement	Corresponding text
2	Video do not consisting face	Unrelated Video
3	Video with no sound	Corresponding Text
4	Video with the word not in the dataset	Word not found
5	Video with more than one lip pair	No proper recognition
6	Video having lip movement of hindi language	Word not found,hence no text
7	No Sequence of distance calculated for Data frames of input video	No Text

CHAPTER 7

RESULTS AND DISCUSSIONS

Results:

Each video is converted into frames.

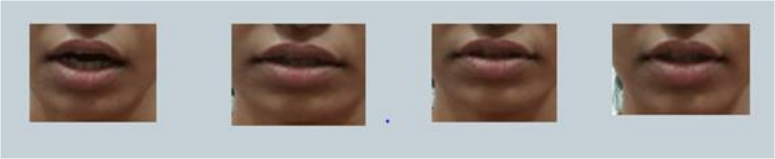


Figure 9. Data Frames generated from video

Figure 9 shows the Frames are generated from the input video. For each second 30 frames are generated.

For each of the frame, firstly face is detected and since our region of interest is lip, and hence the lip part is detected.

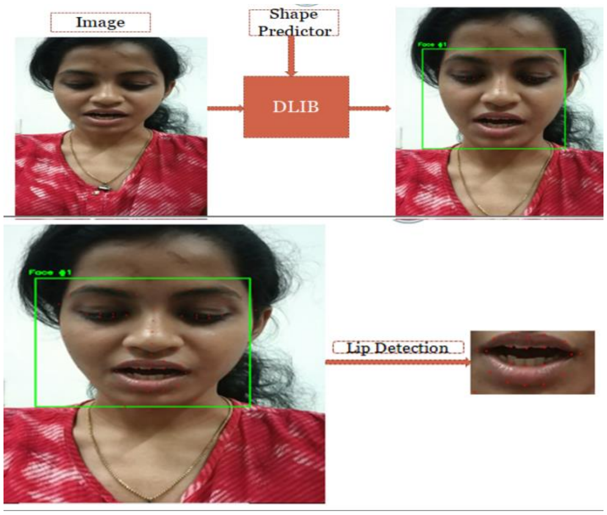


Figure 10. Detection of face and lip

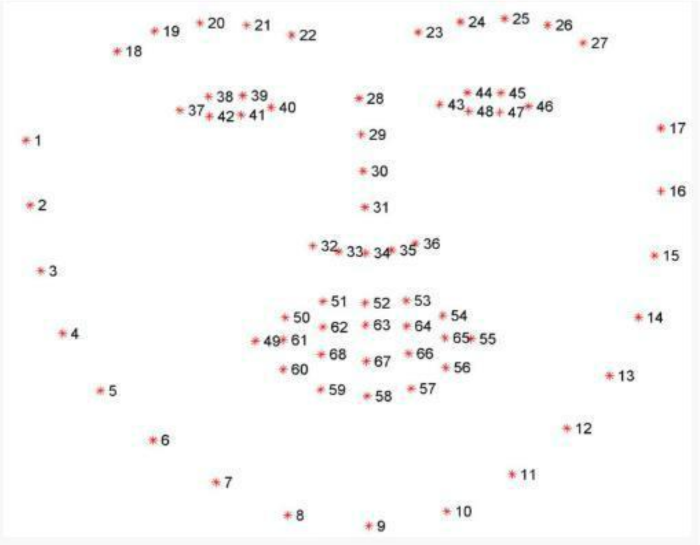
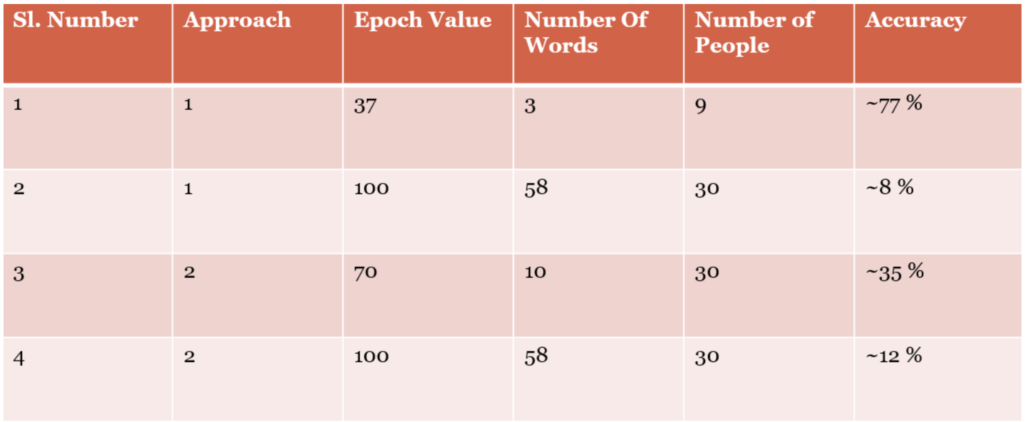


Figure 11. Facial Landmark points

For the word/words spoken by the speaker his/her lip movements are recorded through a video and for every word spoken the corresponding pattern of lip movements are captured and the sequence generated for each word are trained and the word spoken is predicted.

For three words trained using LSTM Classifier 77% accuracy is obtained.



Sl. Number	Approach	Epoch Value	Number Of Words	Number of People	Accuracy
1	1	37	3	9	~77 %
2	1	100	58	30	~8 %
3	2	70	10	30	~35 %
4	2	100	58	30	~12 %

Discussions:

- If the video is blur predicting the word is a challenging task.
- If the video consists of homophones the predicted word may or may not be correct.
- If there is no proper lip movement prediction of the word is difficult.

CHAPTER 8

CONCLUSIONS AND FUTURE SCOPE

In this project an Lip Reading system has been proposed for millions who can't hear, machine learning can be used to discern speech from silent video clips more effectively than Professional lip readers can. The approach is based on deep learning techniques. Overall, the proposed method is adequate for the automatic lip reading for people with hearing disorder. The accuracy assessment (75%) shows that the approach is comparable to techniques, which require manual user interaction videos.

The project can be improved by adding more dataset and by using with high frame rates.

This technique of decoding text from the lip movements can be also used in investigation sector. Lip reading can be further used by people in their preffered language i.e. kannada, telugu etc.

CHAPTER 9

REFERENCES

[1] Assael, Y. (2016). LipNet in Autonomous Vehicles | CES 2017. [Online Video] Yannis Assael.

[2] J. Luettin, N. Thacker, and S. Beet, “Visual speech recognition using active shape models and hidden Markov models,” in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, no. 95, Atlanta, GA, USA, May 1996, pp. 817–820.

[3] Motion, pose (full-frontal view (0◦), angled view (45◦), and side view), Multiple people, video conditions/resolution/lighting, speech methods (accents, styles and rates of speech). *Taken from:* Bear, H.L. (2017, p. 25). Decoding visemes: Improving machine lip-reading. [Online] arXiv: 1710.01288.

[4] Chung, J.S. (2017). Lip Reading Sentences in the Wild (Lip Reading Sentences Dataset), CVPR 2017. [Online Video] Preserve Knowledge.

[5] The M Tank. (2017). A Year in Computer Vision. [Online] TheMTank.com.

[6] Web Sources:

- https://arxiv.org/ftp/arxiv/papers/1802/1802.05521.pdf
- https://medium.com/mlreview/multi-modal-methods-part-one-49361832bc7e
- https://github.com/deepconvolution/LipNet/tree/master/Dataset
- https://regmedia.co.uk/2016/11/08/lipnet.pdf
- https://docs.python.org/3/tutorial/
- https://adventuresinmachinelearning.com/recurrent-neural-networks-lstm-tutorial-tensorflow/
- http://www.themtank.org/a-year-in-computer-vision

CHAPTER 10

APPENDIX

A. Technology used

1. **Machine learning**
   Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

2. **Deep learning**

   Deep learning architectures such as deep neural networks, deep belief networks and recurrent neural networks have been applied to fields including computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics, drug design, medical image analysis, material inspection and board game programs, where they have produced results comparable to and in some cases superior to human experts

3. **Video processing**

   Video processing is a method to perform some operations on an video, in order to **get sequences of dataframs or the images.**

4. **Object detection**

   An image classification or image recognition model simply detect the probability of an object in an image. In contrast to this, object detection refers to identifying the location of an object in the image. An object detection algorithm will output the coordinates of the location of an object with respectthe image. In computer vision, the most popular way to detect an object in an image is to represent its location with the help of bounding boxes.