**Text Recognition From Lip Movement**

**Mentor : Dr.T.Raghuveera**

|  |  |  |  |
| --- | --- | --- | --- |
| **REG NO** | **NAME** | **EMAIL ID** | **PHONE** |
| 2016103017 | Bharath Kumar R | [maxmod98@gmail.com](mailto:maxmod98@gmail.com) | 7358010086 |
| 2016103527 | Hari Krishnan S | sharikrishnan374@gmail.com | 9094242908 |
| 2016103592 | Silambarasan J | [silam.7a@gmail.com](mailto:silam.7a@gmail.com) | 9498375195 |

Contents

1. **PROBLEM STATEMENT ……………………………………………………………………………… 3**
2. **INTRODUCTION ………………………………………………………………………………………… 3**
3. **SUMMARY OF RELATED WORKS ……………………………………………………………….. 5**
4. **HIGH LEVEL BLOCK DIAGRAM ……………………………………………………………………. 7**
5. **DETAILED MODULE DESIGN ………………………………………………………………………. 8**
6. **IMPLEMENTATION DETAILS ……………………………………………………………………… 10**
7. **METRICS OF EVALUATION ………………………………………………………………………… 13**
8. **LIST OF REFERENCES …………………………………………………………………………………. 14**

**1. PROBLEM STATEMENT**

Lipreading is the process of interpreting spoken word by observing lip movement. It plays a vital role in human communication and speech understanding, especially for hearing impaired individuals. Automated lipreading approaches have recently been used in such applications as biometric identification, silent dictation, forensic analysis of surveillance camera capture, and communication with autonomous vehicles. The main objective of our project is to develop a deep neural network that can track human lip movements and predict the sentence spoken.

**2. INTRODUCTION**

A phoneme is the smallest detectable unit of a (spoken) language and is produced by a combination of movements of the lips, teeth and tongue of the speaker. However, some of these phonemes are produced from within the mouth and throat and thus, cannot be detected by just looking at a speaker’s lips. It is for this reason that the number of visually distinctive units or visemes is much smaller than the number of phoneme making lip reading an inherently difficult task.

Lip reading has traditionally been posed as a classification task where words or short phrases from a limited dictionary are classified based on features extracted from lip movements. Some of the early works used a combination of deep learning and hand-crafted features in the first stage followed by a classifier. More recently there has been a surge in end-to-end deep learning approaches for lip reading which focussed on either word level or sentence-level prediction using a combination of convolutional and recurrent networks.

Lipreading plays a crucial role in human communication and speech understanding, as highlighted by the McGurk effect, where one phoneme’s audio dubbed on top of a video of someone speaking a different phoneme results in a third phoneme being perceived. Lipreading is a notoriously difficult task for humans, especially in the absence of context. Most lipreading actuations, besides the lips and sometimes tongue and teeth, are latent and difficult to disambiguate without context. For example, Fisher gives 5 categories of visual phonemes (called *visemes*), out of a list of 23 initial consonant phonemes, that are commonly confused by people when viewing a speaker’s mouth. Many of these were asymmetrically confused, and observations were similar for final consonant phonemes.  
 Consequently, human lipreading performance is poor. Hearing-impaired people achieve an accuracy of only 17*±*12% even for a limited subset of 30 monosyllabic words and 21*±*11% for 30 compound words. An important goal, therefore, is to automate lipreading. Machine lipreaders have enormous practical potential, with applications in improved hearing aids, silent dictation in public spaces, security, speech recognition in noisy environments, biometric identification, and silent-movie processing.  
 Machine lipreading is difficult because it requires extracting spatiotemporal features from the video (since both position and motion are important). Recent deep learning approaches attempt to extract  
those features end-to-end. Most existing work, however, performs only word classification, not sentence-level sequence prediction.

Our model operates at the character-level, using spatiotemporal convolutional neural networks (STCNNs), recurrent neural networks (RNNs), and the connectionist temporal classification loss (CTC).

Milner *et al.* reconstructed audio from video by estimating the spectral envelope using a neural network composed solely of fully connected layers and trained on hand-engineered visual features obtained from mouth region. This approach had the limitation of missing certain speech components such as fundamental frequency and aperiodicity which was then determined artificially thereby compromising quality in order to maximize intelligibility. Ephrat *et al.* modified this technique by using an end-to-end CNN to extract visual features from the entire face while applying a similar approach for modelling audio features using 8th order Linear Predictive Coding (LPC) analysis followed by Line Spectrum Pairs (LSP) decomposition. However, it also suffered from the same missing excitation parameters resulting in an unnatural sounding voice.

**3. SUMMARY OF RELATED WORK**

* **“*Continuous Optical Automatic Speech Recognition by Lipreading*” Alan J. Goldschen, Oscar N. Garciay, Eric Peta jan**

They describe a continuous optical automatic speech recognizer (OASR) that uses optical information from the oral-cavity shadow of a speaker. The system achieves a 25.3 percent recognition on sentences having a perplexity of 150 without using any syntactic, semantic, acoustic, or contextual guides. They introduce 13, mostly dynamic, oral-cavity features used for optical recognition, present phones that appear optical ly similar (visemes) for our speaker, and present the recognition results for our Hidden Markov Models (HMMs) using visemes, trisemes, and generalized trisemes. They conclude that future research is warranted for optical recognition, especially when combined with other input modalities.

* **"*Text Recognition from Silent Lip Movement Video.*" (2018) - Youda Wei, Xiaodong Hu.**

In this paper, he proposed two network architectures lip reading task. First, the visual to audio feature architecture maps a variable-length sequence of video frames to the auditory MFCC features. Second, the audio feature to text architecture distinguishes the text information from the audio feature. For the dataset, a new way of data augmentation is applied in each epoch by randomly insert pictures from the dataset video and adding small Gaussian noise to simulate the situation in real life. The experiments showed that such a combination improves both quality and accuracy in real life situation. The result of the validation accuracy is 92.76%

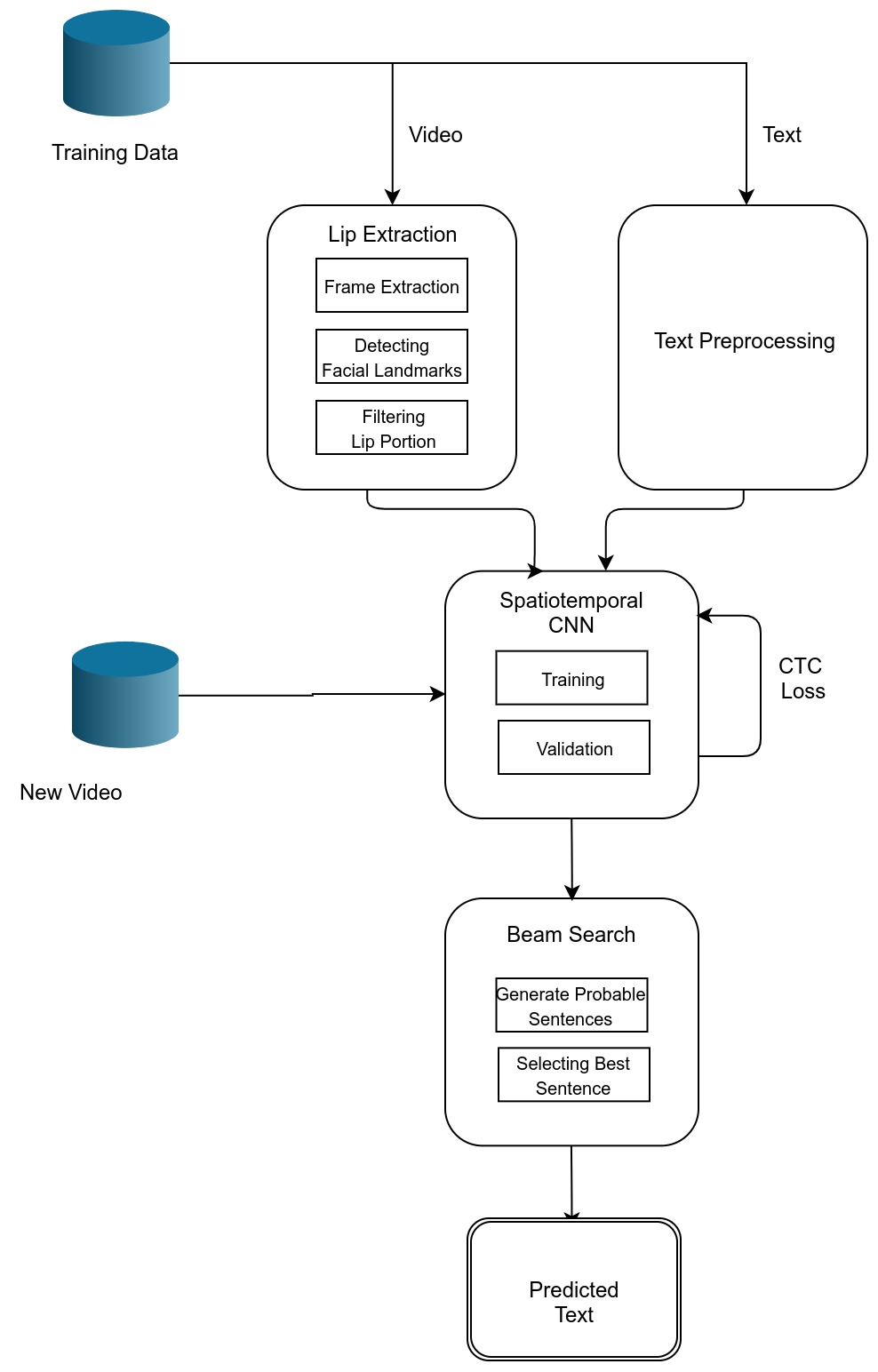
* **"*LIP2AUDSPEC: SPEECH RECONSTRUCTION FROM SILENT LIP MOVEMENTS VIDEO.*" (2018) - Hassan Akbari, Himani Arora, Liangliang Cao, Nima Mesgarani.**

Here they used auditory spectrogram as spectral representation of speech and its corresponding sound generation method resulting in a more natural sounding reconstructed speech. They have proposed a structure consisting of an autoencoder to extract bottleneck features from the auditory spectrogram which is then used as target to our main lip reading network comprising of CNN, LSTM and fully connected layers. They showed that such a combination improves both quality and accuracy of the reconstructed audio. They also conducted different tests for comparing our network with a strong baseline and showed that the proposed structure outperforms the baseline in speech reconstruction. Their future work is to collect more train data, include emotions in reconstructed speech, and to propose an end-to-end structure to directly estimate raw waveform from facial speech-related features.

* **"*A novel approach for lip reading based on neural network*" 2016 - N. Rathee, International Conference on Computational Techniques in Information and Communication Technologies**

In this paper a highly efficient method of speech perception using only visual features is presented. The method is used for identification of 10 Hindi words. The key point extraction method is fully automatic; neither any manual selection nor any inter frame tracking is required. This is an important step in the development of an automatic lip reading system for practical applications. The neural network is used for word identification resulting in better performance due to its inherent capability of large data handling, high speed and false tolerance. Moreover, the LVQ neural network help in minimizing quantization error as it does not require activation function at its output stage.

**4. HIGH LEVEL BLOCK DIAGRAM**



**5. DETAILED MODULE DESIGN**

## **LIP EXTRACTION**

This module first loads the video frame-by-frame. At each frames, the following actions are done. First, the face is detected by a detector that using HOG. Then facial landmarks are detected that estimates co-ordinates of facial region in the image. Here only lip portion is cropped and saved to the folder created for each video.

INPUT: Video

OUPUT: Lip portions in each frame

ALGORITHM:

1. frames = framereader(*path\_to\_video*)
2. for each frame in frames
3. face = face\_detector(frame)
4. landmarks = facial\_landmark\_predictors(face)
5. lip = landmarks[49:68]

* **TEXT PREPROCESSING**

General operations like separating where we split strings of text into smaller pieces and joining tokens are performed. In this module, the “align” dataset is subjected to preprocessing steps like cleaning the data, removing punctuation and special characters and doing a spellcheck.

INPUT: Text Data

Output: Preprocessed Text Data

ALGORITHM:

1. Clean data by removing special characters.
2. Remove fillers
3. Tokenize the text
4. Perform spelling correction
5. Untokenize the text

* **SPATIOTEMPORAL CNN**

The training process of a CNN is done through an iterative algorithm that alternates between feedforward and back propagation passes of the data. The weights of the convolutional filters and fully-connected layers are updated at each iteration of the backpropagation passes. CNN is capable of learning classification features directly from the data. This module is proposed to extract features related to the traces left by different editing operations, and which are utilized to check the authenticity of images.

INPUT: Lips, Text

OUTPUT: Trained model

ALGORITHM:

1. Input of t frames is processed by 3 layers of STCC.
2. Then it is fed into spatial max-pooling layer.
3. Feature extraction is done by 2 Bi-GRUs where each time-step of the GRU output is processed by a linear layer and a softmax.
4. At last CTC loss is computed and proceeded to next batch of training.

* **BEAM SEARCH**

Here we decode the prediction of the neural network by receiving its SoftMax probabilities. **Beam search** in general decides the number of words to keep in-memory at each step to permute the possibilities. Since we perform a non-greedy search, it iteratively creates many possible text candidates (beams). At each step, only the best scoring beams from the previous step are kept. At last, we select the best beam i.e. the probable text sequence.

INPUT: Neural Network prediction

OUTPUT: Predicted Sentence

ALGORITHM:

// NN output matrix *mat,* BW

1. Beams = {}, scores = {}
2. For each iteration do
3. bestBeams = bestBeams(beams,BW) //best of all previous beams
4. for b in bestbeams do
5. score(b) = calcScore(mat,b)
6. for w in words do
7. b’ = b+w
8. scores(b’) = calcScore(mat,b’)
9. beams = beams U b’

**6. Implementation Details and Results/Snapshots**

The GRID corpus consists of 34 subjects, each narrating 1000 sentences. All videos are 3 seconds long with a frame rate of 25fps. The videos were processed with the DLib face detector, and the iBug face landmark predictor with 68. Using these landmarks, we apply an affine transformation to extract a mouth-centred crop of size 100 × 50 pixels per frame.

Facial landmarks are used to localize and represent salient regions of the face, such as:

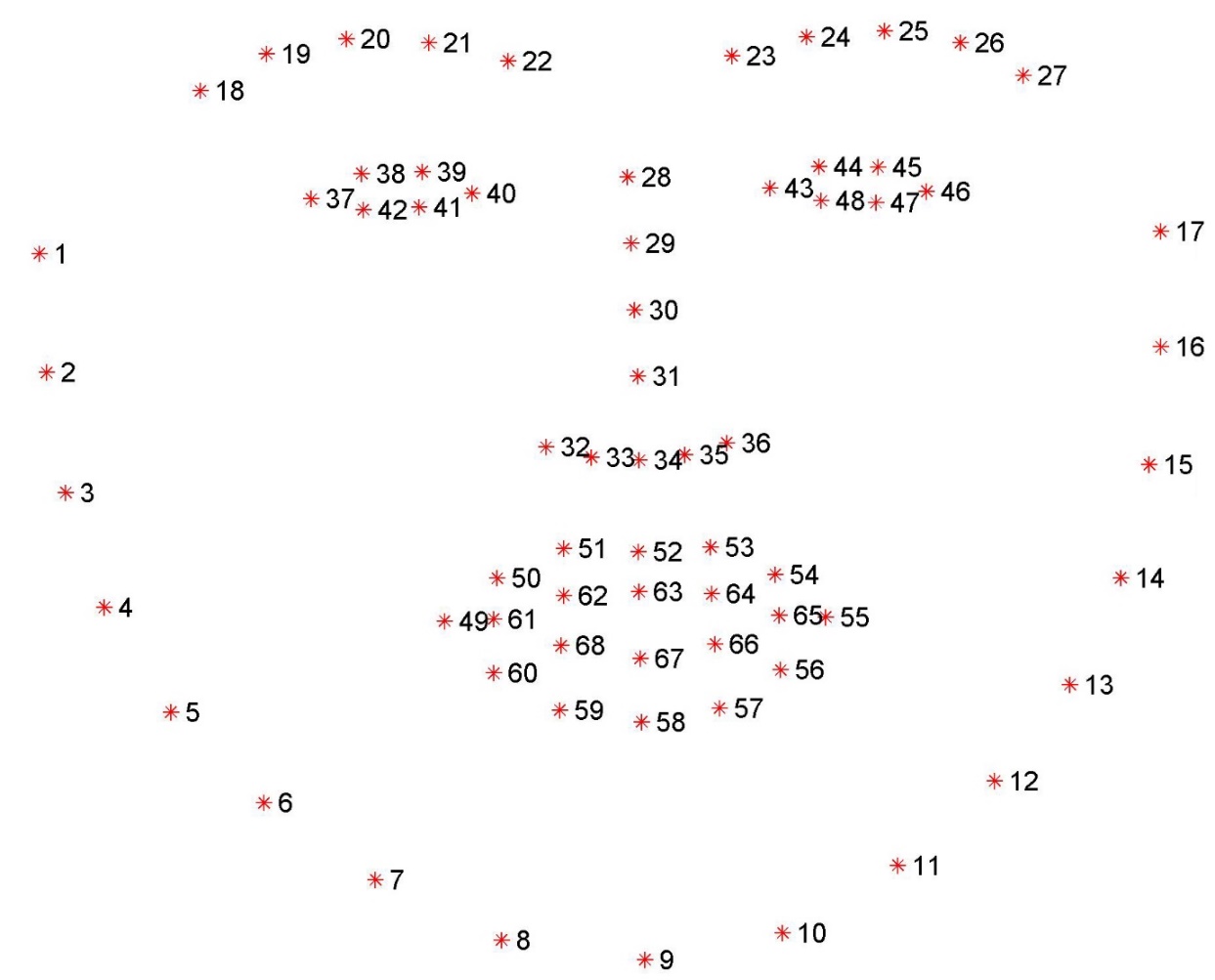
* Eyes
* Eyebrows
* Nose
* Mouth
* Jawline

The pre-trained facial landmark detector inside the dlib library is used to estimate the location of 68 (x, y)-coordinates that map to facial structures on the face.

The indexes of the 68 coordinates can be visualized on the image below:

Detecting facial landmarks in an image is a two step process:

* First we must localize a face(s) in an image. This can be accomplished using a number of different techniques, but normally involve either Haar cascades or HOG + Linear SVM detectors (but any approach that produces a bounding box around the face will suffice).
* Apply the shape predictor, specifically a facial landmark detector, to obtain the (x, y)-coordinates of the face regions in the face ROI.



A feature descriptor is a representation of an image or an image patch that simplifies the image by extracting useful information and throwing away extraneous information.

Typically, a feature descriptor converts an image of size width x height x 3 (channels ) to a feature vector / array of length n. In the case of the HOG feature descriptor, the input image is of size 64 x 128 x 3 and the output feature vector is of length 3780.

In the HOG feature descriptor, the distribution ( histograms ) of directions of gradients ( oriented gradients ) are used as features. Gradients ( x and y derivatives ) of an image are useful because the magnitude of gradients is large around edges and corners ( regions of abrupt intensity changes ) and we know that edges and corners pack in a lot more information about object shape than flat regions.

**Frames**

**Mouth**

**7. Metrics for Evaluation**

To measure the performance of LipNet and the baselines, we compute the word error rate (WER) and the character error rate (CER), standard metrics for the performance of models. We produce approximate maximum-probability predictions from model by performing CTC beam search. WER (or CER) is defined as the minimum number of word (or character) insertions, substitutions, and deletions required to transform the prediction into the ground truth, divided by the number of words (or characters) in the ground truth. Note that WER is usually equal to classification error when the predicted sentence has the same number of words as the ground truth, particularly in our case since almost all errors are substitution errors.

**8. REFERENCES**

[1] Assael, Yannis M, et al. "LipNet: End-to-End Sentence-level Lipreading." (2016).

[2] Goldschen, Alan J., O. N. Garcia, and E. D. Petajan. Continuous Automatic Speech Recognition by Lipreading. George Washington University, 1993.

[3] Youda Wei, Xiaodong Hu. "Text Recognition from Silent Lip Movement Video." (2018).

[4] Hassan Akbari, Himani Arora, Liangliang Cao, Nima Mesgarani. "LIP2AUDSPEC: SPEECH RECONSTRUCTION FROM SILENT LIP MOVEMENTS VIDEO." (2018)

[5] N. Rathee, "A novel approach for lip reading based on neural network," 2016 International Conference on Computational Techniques in Information and Communication Technologies

[6] Ahsan Adeel , Mandar Gogate, Amir Hussain, and William M. Whitmer. "Lip-Reading Driven Deep Learning Approach for Speech Enhancement." (2019)