

24

**PROJECT REPORT ON LIPSYNC-TEXT GENERATION FROM LIP MOVEMENT**

**BY:KUMKUM S**

FROM:

KUMKUM SARANGI

12.5.2024

**Abstract:**

The objective of this literature review is to examine recent developments in machine lipreading technology, including the use of deep learning algorithms, convolutional neural networks, and recurrent neural networks, to analyze lip movements and decode speech. Also, compare and analyze human lipreading performance with the accuracy achieved by machine lipreading systems, considering factors such as visual cues, contextual information, and limitations in both approaches and explore the diverse applications of machine lipreading technology, including improved accessibility for individuals with hearing impairments, speech recognition in noisy environments, biometric identification, and security applications. to investigate the role of deep learning models, such as LipNet, in improving sentence-level lipreading accuracy by capturing spatiotemporal features from video data and integrating them into end-to-end trainable frameworks. On the GRID corpus dataset, LipNet achieves a remarkable 95.2% accuracy in sentence-level lipreading, surpassing both human lipreaders and previous state-of-the-art word-level accuracy.

**Paper Availability:**

**https://www.researchgate.net/publication/309738146\_LipNet\_Sentence-level\_Lipreading**

Framework – arXivLabs

Authors- [Yannis M. Assael](https://dblp.uni-trier.de/search/author?author=Yannis%20M.%20Assael)  
 [Brendan Shillingford](https://dblp.uni-trier.de/search/author?author=Brendan%20Shillingford)  
 [Shimon Whiteson](https://dblp.uni-trier.de/search/author?author=Shimon%20Whiteson)  
 [Nando de Freitas](https://dblp.uni-trier.de/search/author?author=Nando%20de%20Freitas)

Publishing Date- 16 Dec 2016

**Conflicts of Interest:**

The authors declare that they have no conflicts of interest.

Literature review-

**Introduction:**

Lipreading, also known as speechreading, plays a crucial role in human communication and speech understanding. It involves the interpretation of spoken language through the visual cues provided by a speaker's lip movements. The importance of lipreading is underscored by the McGurk effect, as demonstrated by McGurk and MacDonald (1976), where the perception of a phoneme is influenced by the visual information of lip movements. Despite its significance, lipreading is a challenging task for humans, especially in the absence of context.

This page discusses the importance and challenges of lipreading in human communication, highlighting the difficulty humans face in accurately interpreting lip movements without context. It mentions the low accuracy of human lipreading, particularly among hearing-impaired individuals. The paragraph emphasizes the potential of machine lipreading in various applications but notes the complexity due to the need to extract spatiotemporal features from videos.

The author introduces LipNet, an end-to-end sentence-level lipreading model trained using deep learning techniques such as spatiotemporal convolutional neural networks (STCNNs) and recurrent neural networks (RNNs). LipNet achieves a high sentence-level word accuracy of 95.2% on the GRID corpus dataset, outperforming previous methods. Additionally, LipNet demonstrates generalization across unseen speakers and outperforms hearing-impaired individuals in lipreading accuracy.

This literature review will delve into the current state of machine lipreading technology, exploring recent advancements, comparing human and machine performance, and analyzing the applications and challenges in various domains. The review will argue that while machine lipreading holds promise for improving accessibility and speech recognition, significant challenges remain in achieving robustness and accuracy. The subsequent sections will examine these themes in detail, offering insights into the potential of machine lipreading technology and the avenues for future research.

**Summary of studies (virtual bench testing):**

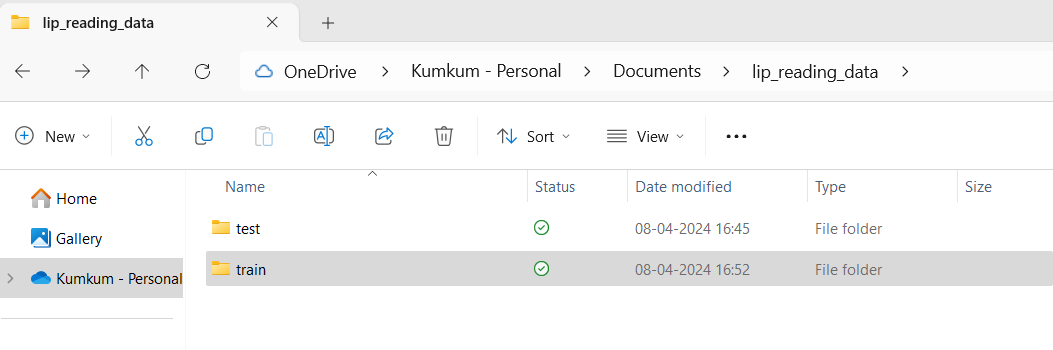
Research has highlighted the difficulty of lipreading, with studies showing that human performance is significantly impacted by factors such as the ambiguity of visual cues and the absence of contextual information (Fisher, 1968; Woodward & Barber, 1960). For instance, Fisher (1968) identified common visemes, or visual phonemes, that are frequently confused by individuals when observing a speaker's mouth. Additionally, Easton and Basala (1982) found that the accuracy of lipreading among hearing-impaired individuals is low, even for a limited subset of words.

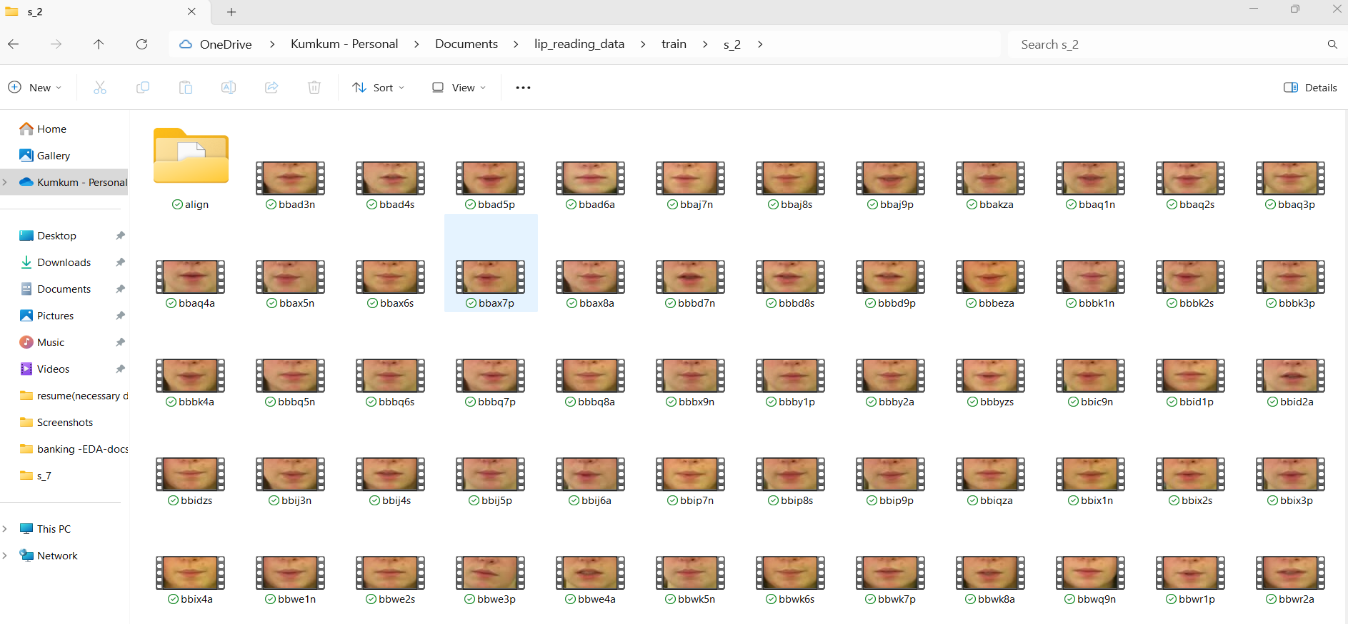
Wand et al. (2016) and Chung & Zisserman (2016a) introduced pioneering deep lipreading approaches that leverage convolutional and recurrent neural networks for end-to-end training. These models represent a departure from traditional methods by simultaneously learning spatiotemporal visual features and sequence models. Despite their advancements, existing work on end-to-end lipreading models has primarily focused on word classification rather than sentence-level sequence prediction. LipNet represents a significant milestone as the first end-to-end model capable of sentence-level sequence prediction for visual speech recognition, trained using connectionist temporal classification loss (CTC). This approach demonstrates the ability to input a sequence of images and output a distribution over sequences of tokens, thereby advancing the state-of-the-art in lipreading technology.

**Datasets:**

[**https://d.docs.live.net/f57499b7b804feff/Documents/lip\_reading\_data**](https://d.docs.live.net/f57499b7b804feff/Documents/lip_reading_data)

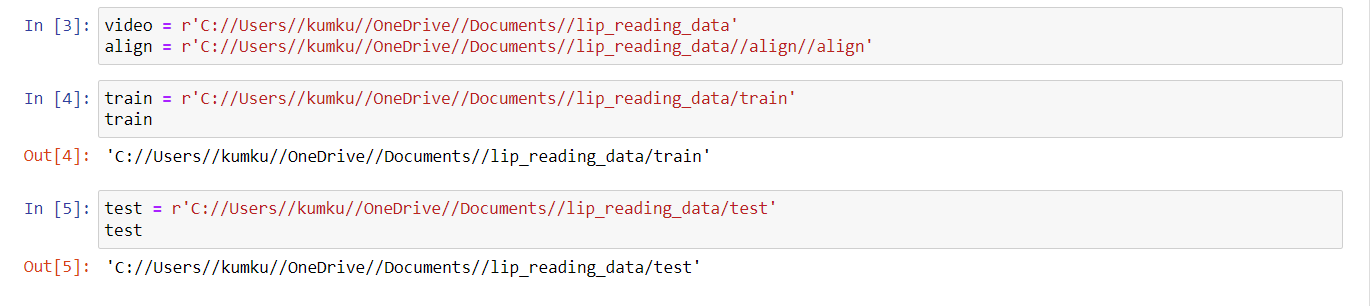
The Grid Corpus dataset consists of 34 subjects, each narrating 1001 sentences consists of high-quality audio and video (facial) recordings comprising of 34 sections along with transcriptions. Each speaker comes with separate folder/sections containing 1001 sentences/words with no of speakers is 1. The dataset contains an align file for each video, the align file contains certain markings and the words being spoken in the respective video. The videos are 3s and 75 frames each.





However, some videos are absent, and are either empty or damaged. To ensure an unbiased evaluation, a split approach is employed, withholding the data of 1 male speaker split it into “Test” and 1 female speaker split into “Train” and “validation” purposes, totalling 30,030 videos. The remaining videos i.e 5,005 are allocated for testing purposes. We standardise the RGB channels

over the whole training set to have zero mean and unit variance.





**Analysis and Synthesis:**

**Primitive Methodologies:**

Traditional approaches focused on word or phoneme classification including learning multimodal ­audio-visual representations, incorporating visual features into traditional speech-style processing pipelines, and combining different methods. Chung & Zisserman (2016a) propose spatial and spatiotemporal convolutional neural networks for word classification, but their models fall short in comparison to spatial architectures and cannot handle variable sequence lengths or attempt sentence-level prediction. Similarly, Wand et al. (2016) introduce LSTM recurrent neural networks for lipreading but do not address sentence-level prediction. Garg et al. (2016) apply pre-trained VGG models to classify words and phrases but achieve limited accuracy due to the small dataset and training approach.

Several datasets, including AVICar, AVLetters, AVLetters2, BBC TV, CUAVE, OuluVS1, and OuluVS2, are mentioned, with references to studies by Zhou et al. (2014) and Chung & Zisserman (2016a). However, it notes that many of these datasets only contain single words or are relatively small in size.

**Advanced Methodology:**

Development of LipNet represents a significant milestone as the first end-to-end model capable of sentence-level sequence prediction for visual speech recognition, trained using connectionist temporal classification loss (CTC). This approach demonstrates the ability to input a sequence of images and output a distribution over sequences of tokens, thereby advancing the state-of-the-art in lipreading technology, offering the potential for improved accuracy and generalization across diverse datasets and speakers. LipNet, introduced by Wand et al. (2016), represents the first end-to-end sentence-level lipreading model, achieving impressive accuracy on the GRID corpus dataset. By addressing the limitations of previous approaches and focusing on sentence-level prediction, LipNet demonstrates the potential of deep learning in advancing the state-of-the-art in lipreading technology.

**Discussions of implications:**

It begins with three spatiotemporal convolutions, each followed by channel-wise dropout and spatial max-pooling. These layers are responsible for extracting features from the input data. Subsequently, the extracted features are passed through two bidirectional gated recurrent units (Bi-GRUs), which play a crucial role in efficiently aggregating the output from the spatiotemporal convolutional neural network (STCNN). Finally, a linear transformation is applied at each time-step, followed by a softmax operation over the vocabulary augmented with the connectionist temporal classification (CTC) blank symbol, and then the CTC loss is computed. All layers in the architecture use rectified linear unit (ReLU) activation functions.

**Data Augmentation:**

For data augmentation purposes, the dataset was augmented using simple transformations to mitigate overfitting. Firstly, training was conducted on both the original and horizontally mirrored image sequences. Secondly, the dataset provided word start and end timings for each sentence video, allowing for augmentation of the sentence-level training data with video clips of individual words as additional training instances. These instances were subject to a decay rate of 0.925. Thirdly, to promote resilience to varying motion speeds, frames were deleted or duplicated with a per-frame probability of 0.05. It's noted that the same augmentation methods were applied consistently across all proposed baselines and models.

The videos underwent processing using the DLib face detector and the iBug face landmark predictor coupled with an online Kalman Filter. This process resulted in the extraction of 68 facial landmarks per frame, which were used to apply an affine transformation to obtain a mouth-centered crop of size 100x50 pixels. Additionally, the RGB channels were standardized across the entire training set to have zero mean and unit variance.

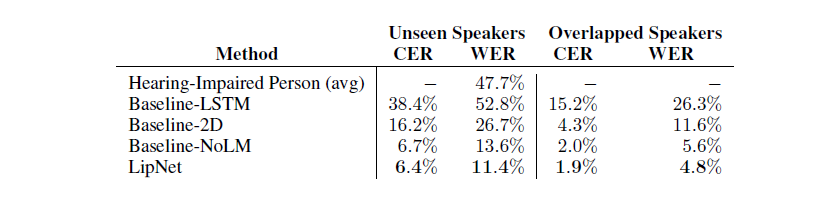
**Baselines:**

The baseline performance of hearing-impaired individuals was assessed by three members of the Oxford Students’ Disability Community. They were introduced to the grammar of the GRID corpus and then observed 10 minutes of annotated videos from the training dataset, followed by annotating 300 random videos from the evaluation dataset. When uncertain, they were asked to select the most probable answer.

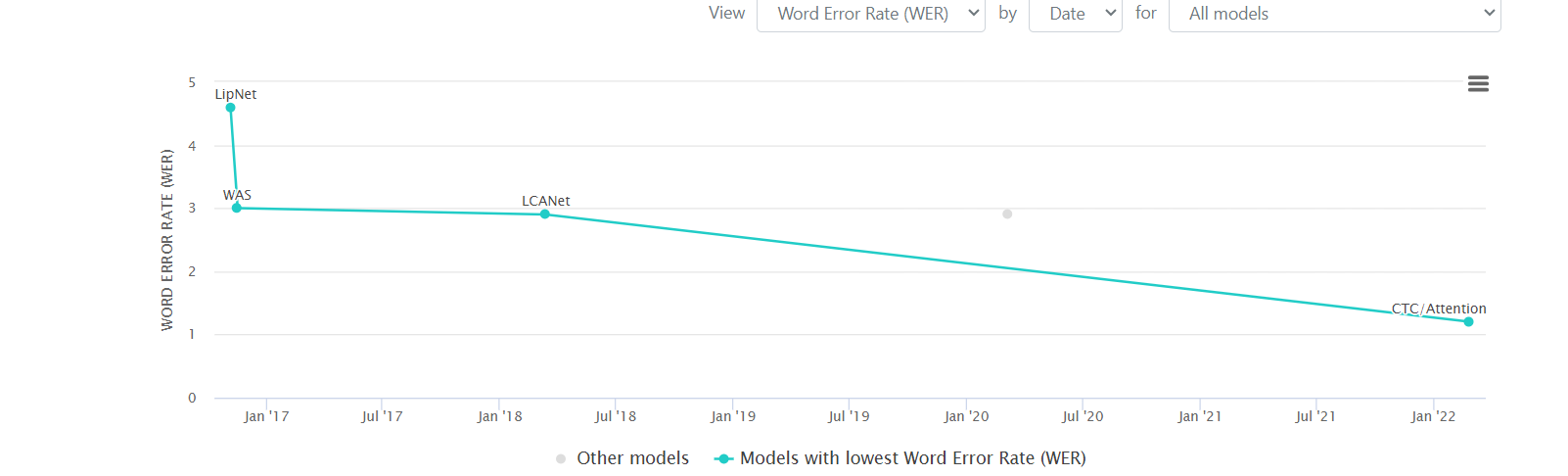
* LSTM replicates the model architecture of the previous deep learning state-of-the-art on the GRID corpus, using the sentence-level training setup of LipNet. More implementation details can be found in Appendix A.
* 2D is based on the LipNet architecture but replaces the spatiotemporal convolutions with spatial-only convolutions similar to those used by Chung & Zisserman (2016a). Notably, contrary to the results observed with LipNet, Chung & Zisserman (2016a) reported 14% and 31% poorer performance of their spatiotemporal convolutional neural networks (STCNNs) compared to the 2D architectures in their two datasets.
* NoLM is identical to LipNet but with the language model used in beam search disabled.

**Findings/Results:**

The performance of LipNet and the baselines is measured using word error rate (WER) and character error rate (CER), which are standard metrics for evaluating automatic speech recognition (ASR) models. WER/CER is calculated as the minimum number of word/character insertions, substitutions, and deletions required to transform the prediction into the ground truth, divided by the number of words/characters in the ground truth.

****

These findings underscore the need for automated lipreading systems to assist individuals with hearing impairments and improve communication in various settings.



**Knowledge Gaps:**

* Lipnet has the potential for improved accuracy and generalization across diverse datasets However, lipreading has its own set of challenges, particularly in extracting spatiotemporal features from video data, enhancing robustness in noisy environments and expanding the range of potential applications.
* Empirical evaluation confirms the importance of spatiotemporal feature extraction and efficient temporal aggregation, surpassing a human lipreading baseline by a significant margin and outperforming the word-level state-of-the-art on the GRID corpus.

**Impact/Key Evaluation insights:**

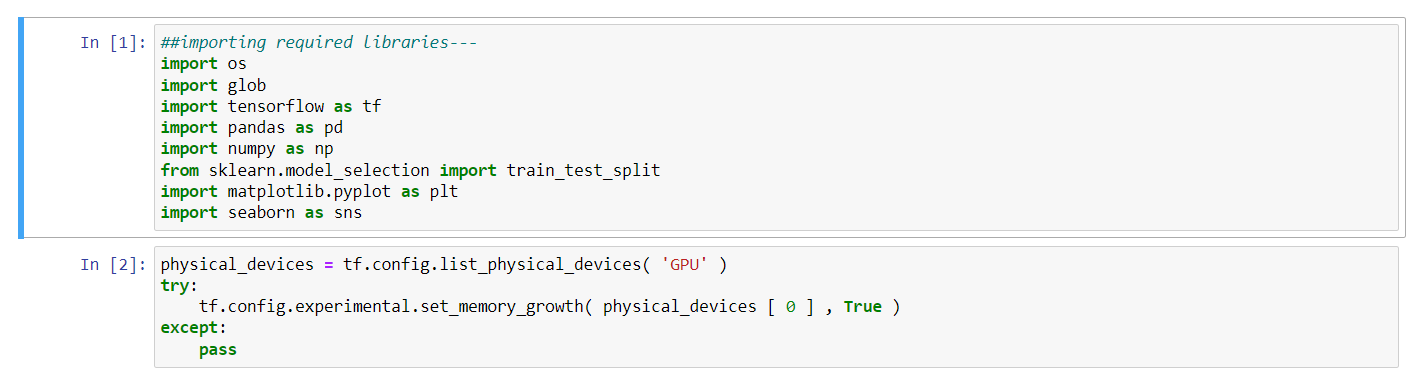
* While LipNet shows promising results, it's mentioned that performance could potentially improve with more data. This shows how the model's performance scales with larger datasets and its ability to generalize to unseen data beyond the GRID corpus.
* While LipNet's primary application is silent dictation, there's an acknowledgment of potential broader applications. Further research could explore the effectiveness of LipNet in various real-world scenarios beyond dictation, such as security, accessibility, or human-computer interaction.

**Recommendations/Future Directions:**

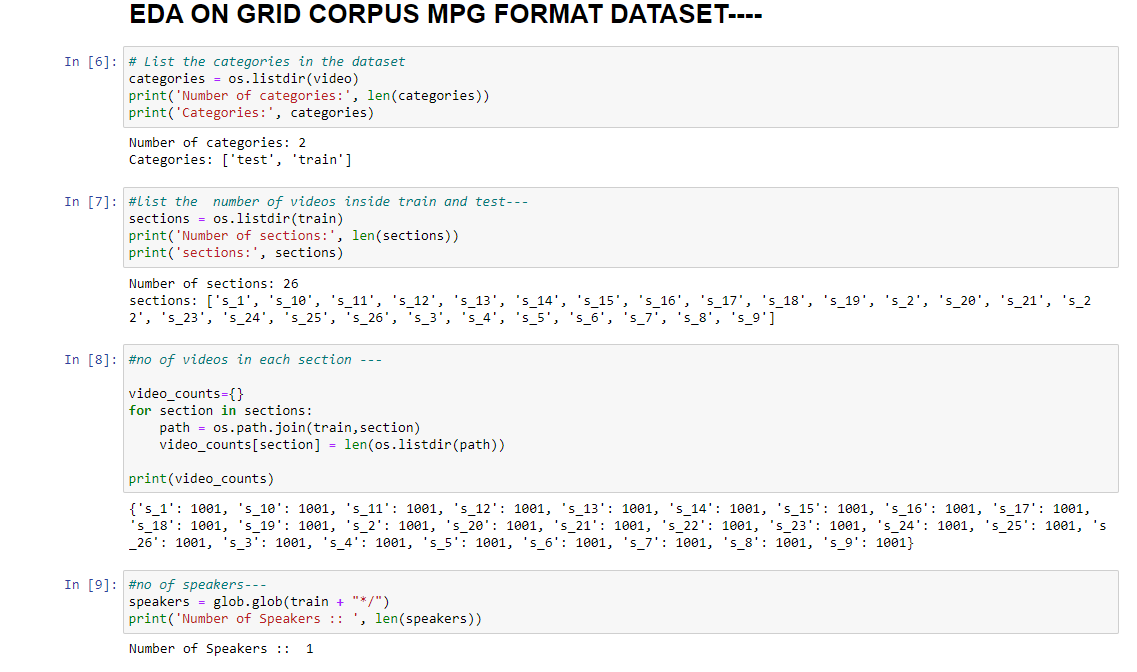
* Extending LipNet to jointly trained audiovisual speech recognition models to enhance robustness in noisy environments. This raises questions about how effectively LipNet can integrate visual input with audio input and whether this approach improves overall speech recognition accuracy.
* There are plans to explore applications beyond silent dictation by extending LipNet to jointly trained audiovisual speech recognition models, enhancing robustness in noisy environments and expanding the range of potential applications.

**Explanatory Data Analysis(EDA):**

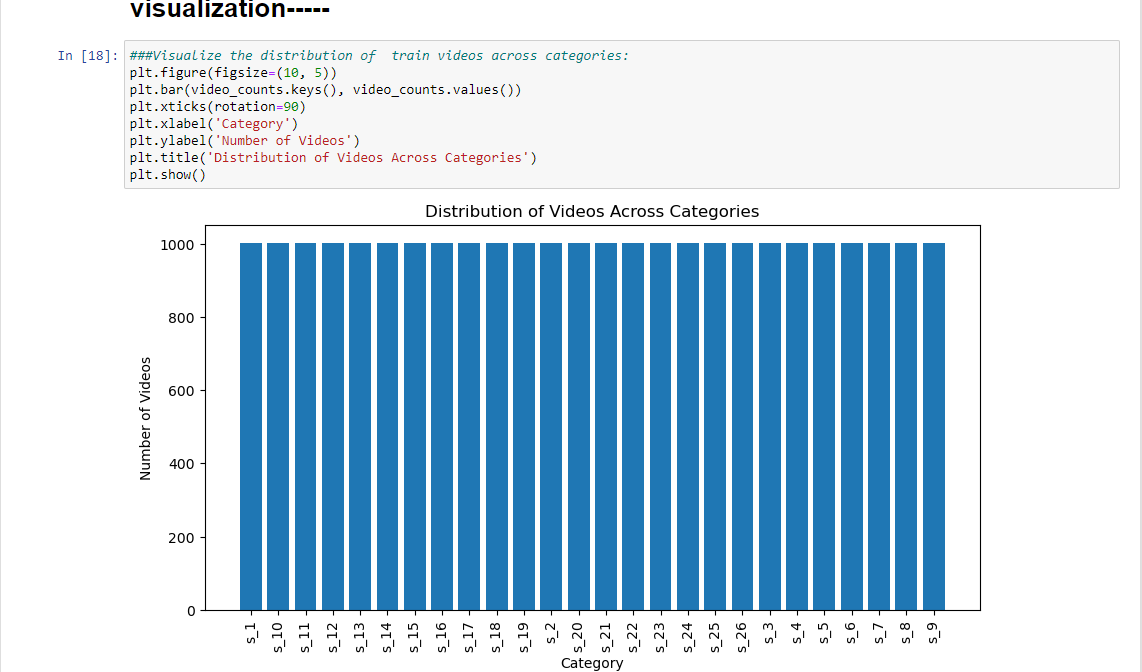
* Importing required libraries

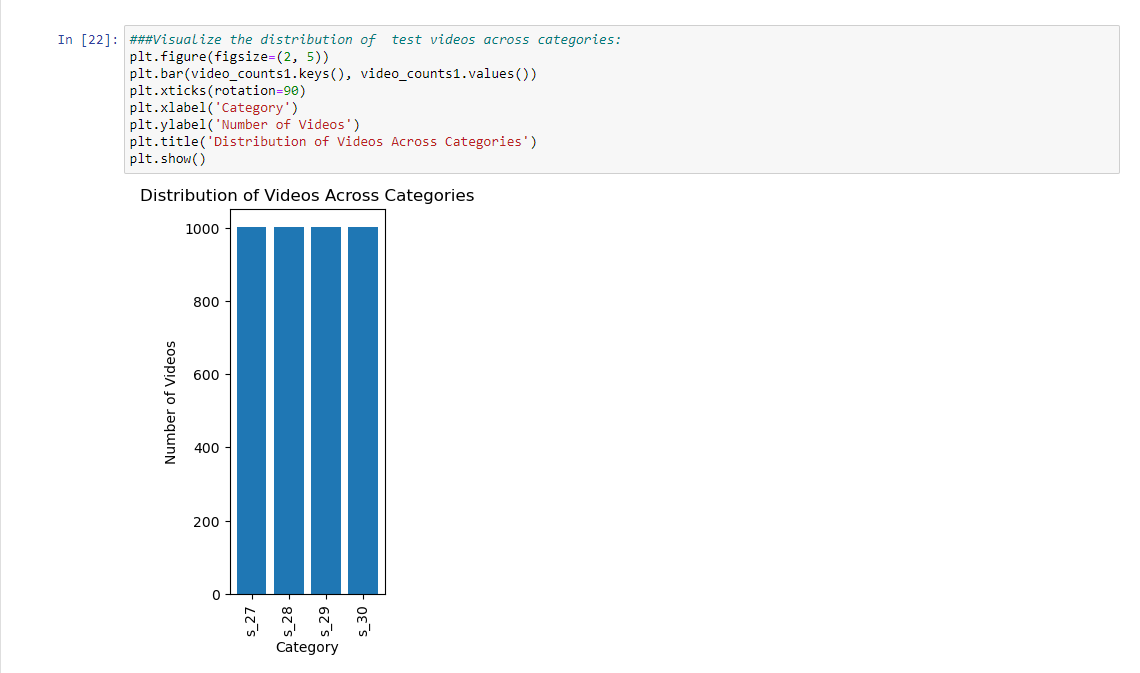
****

* Analysing the video samples data –

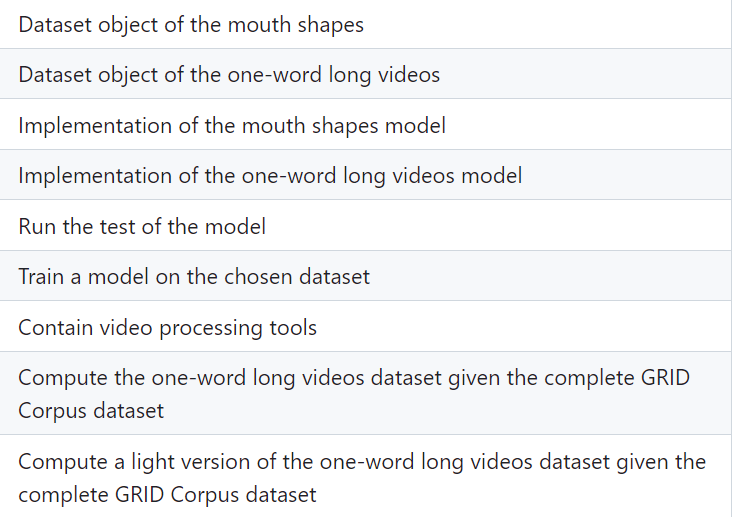


****

****

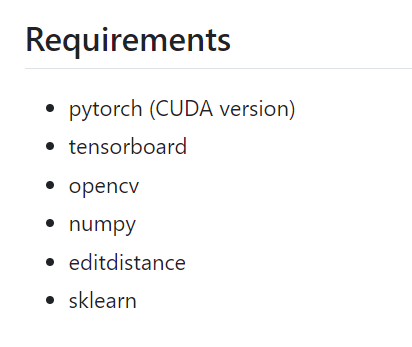


**Methodology used :**

****

**Discussion:**

This project aims to build a deep learning model that can recognize spoken words from a person's lip movements. The model takes in a video of a person speaking and outputs the transcribed text.

****

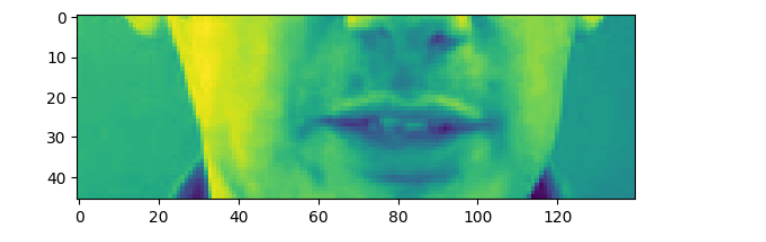
**Dataset**

The dataset used for training and evaluation is the GRID corpus, which is a collection of audiovisual recordings of people speaking short sentences.

**Preprocessing**

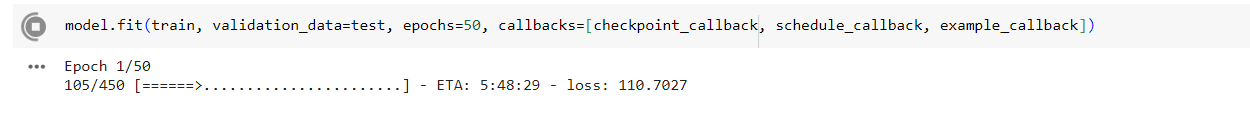
The preprocessing steps include:

1. Extracting the video frames and preprocessing them by converting them to grayscale and cropping them to focus only on the lips.
2. Loading the text alignments for the videos and converting them into numerical values.



## Training

The model is trained using the Adam optimizer and the categorical cross-entropy loss function. The model is trained on 1000 videos and validated on the remaining 50 videos.



#### **Testing**

Change path of weights file and of the test files in test.py

Add the shape\_predictor\_68\_face\_landmarks.dat file in the word\_lipreading folderand run test.py.

**Results:**

Overall, the ultimate goal of lip sync models is to generate realistic lip movements that closely match the input speech or text, enhancing the overall immersive experience in applications such as animation, virtual avatars, dubbing, and video editing. Evaluating and improving the performance of lip sync models is an ongoing area of research in the fields of computer vision, natural language processing, and audio processing.

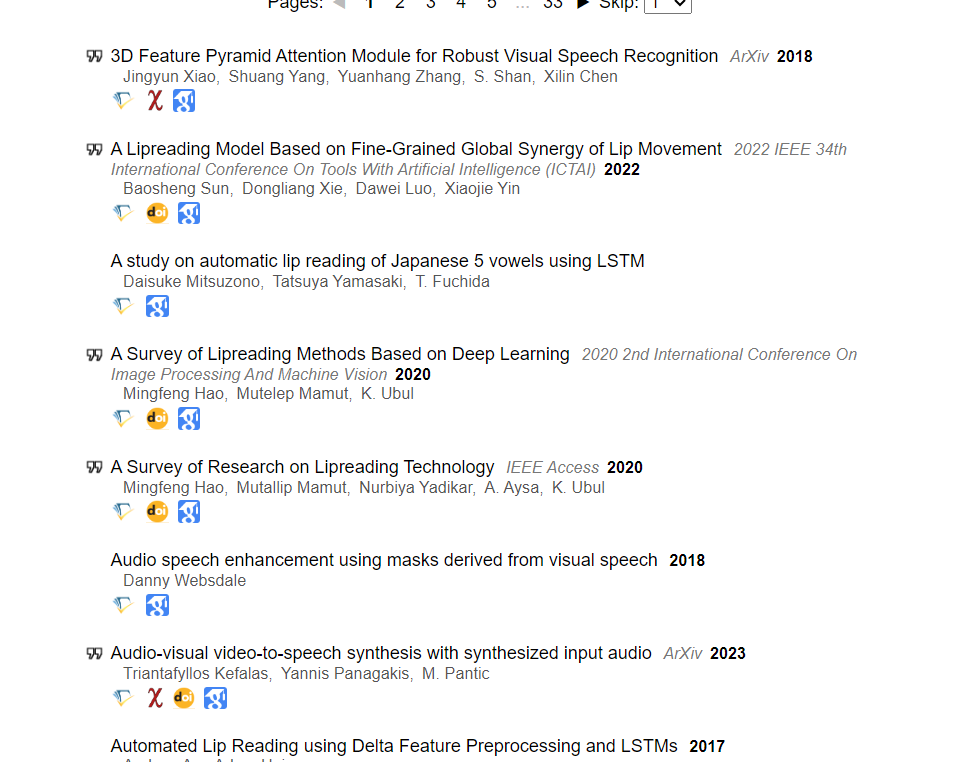
**Conclusion:**

This project demonstrates the feasibility of using deep learning to recognize spoken words from a person's lip movements. The lip reading performance can be improved by using a larger dataset.

**Github Link:**

<https://github.com/Kumkum-1991/kumkum123/tree/main>

**Bibliography/References:**

****