**Project Lip Reading Diverse Testing**

Dr Vikas Tripathi 1, Ayush Saklani 2 and Devanshu Rawat 2

1 Graphic Era University, 2 Graphic Era Hill University

Dehradun, Uttarakhand, India.

**Abstract.** Our aim is to recognise the words being spoken by a talking face, solely by the video not the audio. Previous research has primarily focused on recognizing a few utterances in controlled environments, such as digits and alphabets, partly due to a lack of suitable datasets.

This paper investigates the performance of benchmark model for isolated word lip-reading under different languages and different accent either by using the existing datasets like LRW-1000 and Glips or by making our own dataset.

The current model we are using in this research was introduced in 2022 that gives an accuracy of 93.4% on LRW dataset, which constitute an improvement of 4.6% over previous state-of-the-art model’s performance.

**Keywords** – Lipreading, LRW, LRW-1000, Glips, Machine Learning, Computer Vision, Lip reading in the wild, Visual speech recognition, deep learning, feature extraction, natural language processing.

**1. Introduction**

The ability to comprehend speech solely through visual cues, or lip-reading, is a very appealing field. It is obviously useful for voice transcription in situations when audio is not accessible.

Our inspiration for doing research on this topic is to contribute to the automation of forensic lip reading and effective automation in military reconnaissance.

LRW has been the dataset that is treated as the benchmarks by majority of paper.

Our approach for this diverse testing is evaluating the model (that is tried and tested on English) and testing the accuracy on German dataset.

**2. Literature Review**

The Lip Reading in the Wild (LRW) dataset is a large-scale audio-visual database of 500 words spoken by over 1,000 speakers. It was initially released in 2016. The BBC collected the dataset. You need to ask the BBC for permission to use the dataset. The dataset for LRW is separated into test, validation, and training sets. The target word appears in the centre of each of the 29 frames (1.16 seconds) that make up an utterance. The metadata contains the word duration. The LRW dataset Is one of several audio-visual speech recognition datasets from BBC. The other datasets are LRS2 and LRS3 i.e., lipreading sentence containing whole sentences.

LRW has been the dataset that is treated as the benchmarks by majority of paper.

LRW-1000 is a Mandarin language large-scale benchmark for word-level lipreading. It is video dataset used for lipreading that was firstly released in 2019. It has been renamed as CAS-VSR-W1k. The dataset includes 1,000 classes, 718,018 video samples by more than 2,000 speakers containing more than 1,000,000 Chinese character instances.

The Glips dataset comprises 250,000 publicly accessible videos featuring the faces of speakers from the Hessian Parliament, a German province. The videos were subjected to an autonomous pipeline for word-level lip reading. With each video encoding one word of interest in a context lasting 1.16 seconds, the format is like that of the Lip Reading in the Wild (LRW) dataset, making it compatible for research on transfer learning between the two datasets. Each of the distinct spoken words, which range in length from 4 to 18 characters, has 500 instances and distinct MPEG-4 audio and text metadata files that come from 1018 parliamentary sessions.

The model that we are now utilizing in this study was released in 2022 and has an accuracy of 93.4 percent on the LRW dataset, which is 4.6 percent better than the performance of the previous state-of-the-art model. Our approach for this diverse testing is evaluating the model (that is tried and tested on English) and testing the accuracy on German dataset.

**3. Methodology**

**Phase 1 – Testing if code works on LRW dataset :-**

First we wanted to check that the code we got from the internet was functioning properly or not by running it on LRW dataset. Therefore we got the dataset after requesting to BBC through email, for permission to use the dataset.

After receiving permission and testing the opensource code on LRW dataset we got the same results as stated in the research paper.

**Phase 2 – Converting code to run on Glips Dataset :-**

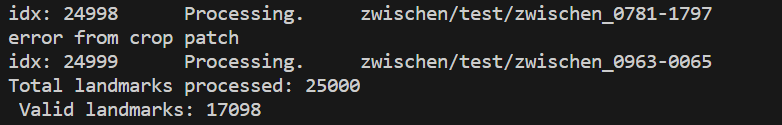
The first problem we encountered is that Glips dataset is the data from a German state parliament so high camera angle was already a reason for low accuracy in landmark marking.

Secondly the member of parliament tends to address many people in the parliaments so there is a lot of body movements further making it difficult to catch the lip area. The dataset compatible with the given model as we have discussed earlier LRW is a highly standardised and refined dataset which acts as ideal learning and testing whereas Glips is more of a practical dataset to work on that gives us a more real life situation to work with that create a lot of problem for this model to tackle during preprocessing due to absent landmarks on many frames of the video.

The code required lanmarks to perform preprocessing so Landmark generation for Glips was done by an opensource code and tweaks were done accordingly.

After landmark generation we had to perform preprocessing using the script crop\_mouth\_from\_video.py to get mouth ROI. Here we encountered the problem that landmarks were required to be in form of numpy array in place of list data structures that we had generated.

After generating landmarks again and discarding the landmarks which had missing frames we were able to extract mouth ROIs.



From above screenshot we can see that from total of 25000 landmarks only 17098 were useful ones as other had missing frames.

**Phase 3- Testing the LRW trained model on Glips dataset’s testing split :-**

After further customization we were able to run the code for testing the model on Glips dataset’s testing split.

As expected we got 0.19% accuracy as the model was trained on English dataset and we were working on German dataset. However, this step helped us in customising the code to run properly for testing accuracy

Learning from this phase will help us in later on in project, after we have produced a trained model on Glips dataset.

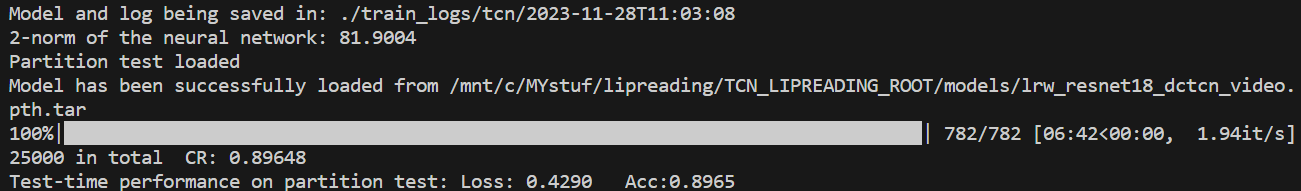
**Phase 4- Training the model on Glips dataset :-**

* On progress

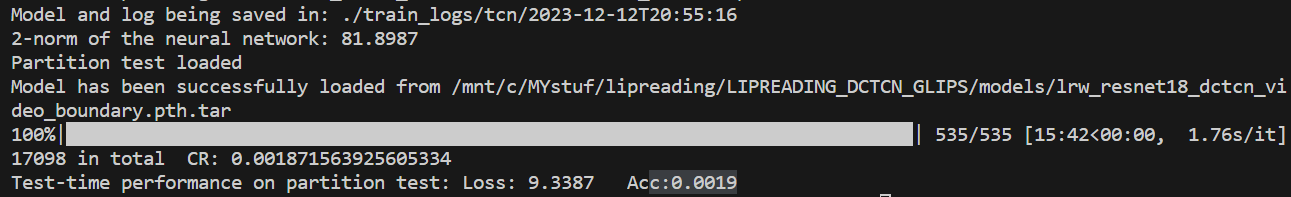
**Phase 5- Testing the new trained model on Glips dataset:-**

* On progress

**4. Result**



Results of testing model on LRW.



Results of testing model(trained in LRW) on Glips.

**5. Future Work**

Creating a dataset which draws inspiration from the LRW. Creating an Indian version of LRW by processing licensed Prasar Bharti DD National News archive news programs to find the accuracy changes on different accents on the current model.

**6. References**

LRW source website:-

<https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrw1.html>

Glips Source:-

<https://www.kaggle.com/datasets/geraldschwiebert/glips-german-lipreading-dataset>

Original open source code:- <https://github.com/mpc001/Lipreading_using_Temporal_Convolutional_Networks>

Modified code for Glips:-

<https://github.com/RawatDevanshu/Lipreading_DCTCN_Glips>

**Idea box or rough page below**

methodology

LRW is highly refined dataset with ambitious standards for preprocessing data.

Glips

The first problem we encountered is that Glips dataset is the data from a German state parliament so high camera angles is already a reason for low accuracy in prediction and landmark marking.

Secondly the member of parliament tends to address many people in the parliaments so there is a lot of body movements further making it difficult to catch the lip area that is compatible with the given model as we have discussed earlier LRW is a highly standardised and refined dataset.

LRW has been the dataset that is treated as the benchmarks by majority of paper.

Our approach for this diverse testing is evaluating the model (that is tried and tested on English) and testing the accuracy on German dataset.

And in future work, we plan to create a dataset like LRW that will be created using news shows from Prasar Bharti DD national English news shows. evaluating the model on Indian accent English to find the accuracy changes on different accents.

The model we are using for testing is the current state of the art model that has 92.3-92.9% accuracy on LRW that was introduced in 2022.