**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[1] and Glips[2] or by making our own dataset.

The current model[3] 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.**

**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)[4] 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[1] 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[2] 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[3] 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 the code on LRW dataset: -**

Initially, we wanted to use the LRW dataset to see if the model we downloaded from the opensource platforms was working correctly or not.

Second, to obtain the LRW dataset, we sent formal emails to the BBC requesting access to the dataset for research purposes.

We obtained the anticipated outcomes as reported in the research paper after gaining access to the LRW and testing the open-source code on a dataset.

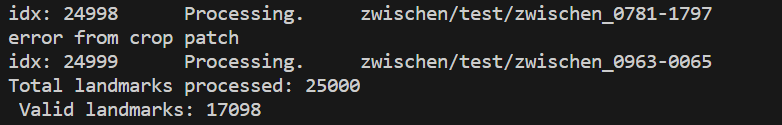
**Phase 2 – Converting code to run on Glips Dataset: -**

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 landmarks to perform preprocessing. Landmark generation for Glips was done by an opensource code and tweaks were done accordingly to make it compatible with Glips dataset.

After landmark generation preprocessing was done using the script crop\_mouth\_from\_video.py to get mouth ROI and it took 22 hours to complete. 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 for the second time correcting the previous errors again took 20-22 hours. we discarded the landmarks which had missing frames during the preprocessing, and we were able to extract mouth ROIs.

|  |
| --- |
| fig. LR\_01 |

From the fig. LR\_01 we can observe that out of 25,000 landmarks in Glips only 17,098 were useful or as standardised as LRW, rest of the 7,902 landmarks had missing frames.

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

After making additional customizations. we were able to test 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 in project, after we have produced a trained model on Glips dataset.

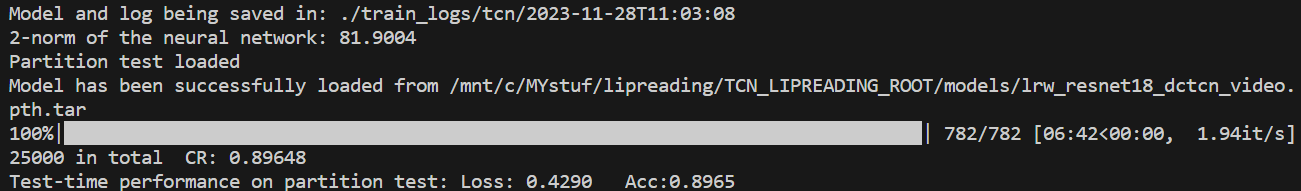
**Phase 4- Training the model on Glips dataset: -**

* On progress

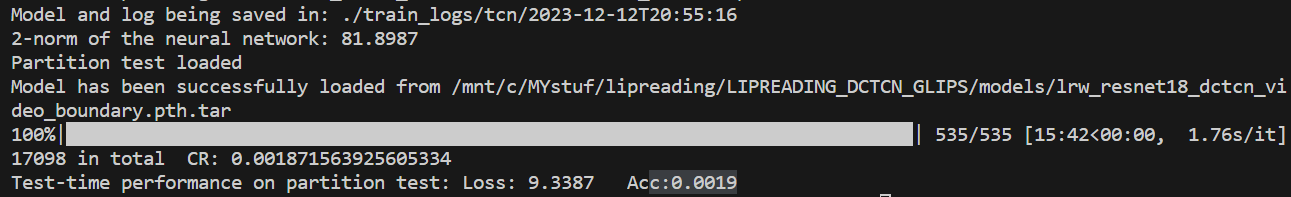
**Phase 5- Testing the new trained model on Glips dataset: -**

* On progress

**4. Result**



|  |
| --- |
| fig. LR\_02 (Accuracy of testing model on LRW.) |



|  |
| --- |
| fig. LR\_03 (Accuracy of testing model (trained in LRW) on Glips.) |

**5. Future Work**

Creating a dataset which draws inspiration from the LRW is our goal. 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**

[1] S. Yang *et al.*, “LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild,” in *Proceedings - 14th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2019*, 2019. doi: 10.1109/FG.2019.8756582.

[2] G. Schwiebert, C. Weber, L. Qu, H. Siqueira, and S. Wermter, “A Multimodal German Dataset for Automatic Lip Reading Systems and Transfer Learning,” in *2022 Language Resources and Evaluation Conference, LREC 2022*, 2022.

[3] P. Ma, Y. Wang, S. Petridis, J. Shen, and M. Pantic, “TRAINING STRATEGIES FOR IMPROVED LIP-READING,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2022. doi: 10.1109/ICASSP43922.2022.9746706.

[4] J. S. Chung and A. Zisserman, “Lip reading in the wild,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017. doi: 10.1007/978-3-319-54184-6\_6.

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

Glips: - <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>

Python Scripts we used: -<https://github.com/RawatDevanshu/Lipreading_DCTCN_Glips/tree/master/python_scripts>