IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 2, February 2023

DOI: 10.17148/IJARCCE.2023.12229

LIP-TO-SPEECH SYNTHESIS USING MACHINE LEARNING

Prof. Padmavathi B[1],

Akash P[2], S Dhanush[3], Kiran Raj R K[4], Krishna Prasad H S[5]

Assistant Professor, Computer Science, East West Institute of Technology, Bangalore, India [1]

Student, Computer Science, East West Institute of Technology, Bangalore, India[2-5]

Abstract: Lip reading technology uses analysis of lip movement to record the speaker's message. It is used widely in many aspects of daily life. The performance of the entire lip-reading system is impacted by the dataset's quality. As a result, this study investigates the dataset for lipreading. Scikit video is used to extract frames from the source video. Idlib is then used to conduct facial detection. Lip cropping is accomplished by processing the feature points to obtain lip pictures. The dataset is then expanded by doing data augmentation. 33 voices are included in the collection, and each speaker's lips are represented by 7,000 images. A technique for creating datasets is suggested. Prior to decomposing the treated films in the Scikit-Video library.

Keywords: Lip reading, Idlib, Scikit video, lip pictures and lip cropping.

I. INTRODUCTION

Natural language processing and machine vision are the foundations of the lip-reading technology, which has several use cases. Lipreading, for instance, can help speech recognition. Because the environment disturbs the audio, speech recognition accuracy suffers when ambient noise levels are too high. Lipreading can help speech recognition become more accurate since it only requires visual information, no audio. A hearing impairment can also be helped by lipreading. In India, roughly there around 23 million people who have language and hearing impairments, and they communicate mostly using sign language. Although, not Lip-reading can compensate for the fact that not everyone is proficient in sign language and improve communication. Compilation of the dataset is a major issue in scanning and affects how accurate future lip feature extraction and scanning identification is. To ensure the accuracy of the lip reading, a good data collection must be used at the outset for the process. The following issues are the key ones with producing lipreading datasets. Faces are absent from certain scenes in the video that has been gathered or they just take up a tiny piece of the image, which, if unprocessed, reduces the recognition accuracy.

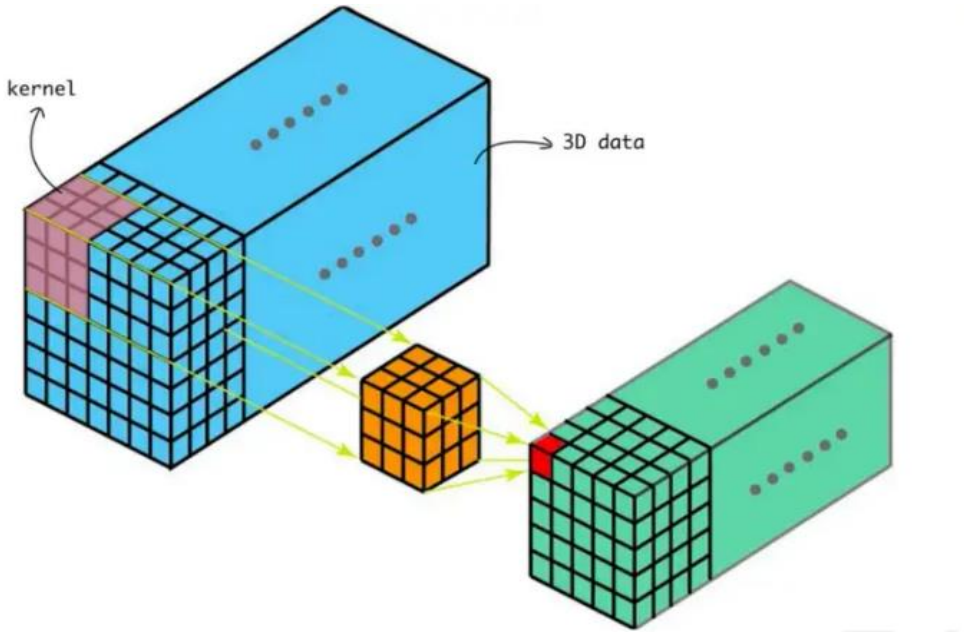
The world's communication is growing more digital and more visual at the same time. There is an increase in the consumption of video content, including YouTube videos, movies, and video calls. Consequently, research on comprehending and enabling talking-face video applications has been busy in recent years. Speech/text-based lip synthesis is one example of a task that has made significant progress. Lip-to-text generation and lip-to-speech generation, which also fall under the category of "lip-reading," are the opposites of these tasks and have shown to be far more difficult. Some ground-breaking efforts have advanced the field of lip-to-text generation with models that are applicable to any speaker in the wild. However, lip-to- speech synthesis, its twin task, has not yet experienced a comparable progress in such unrestricted situations. The lip region's clipping range influences the recognition outcome as well. Lip-to-text generation and lip-to-speech generation, two activities that fall under the category of "lip-reading," have shown to be far more difficult than their opposites. The models that function for every speaker in the world, several amazing efforts have pushed the limits of the lip-to-text generation challenge. However, lip-to-speech synthesis, its brother job, has not yet shown a comparable development in such unrestricted situations. Lip2Speech does not require extra annotations, unlike traditional visual speech recognition tasks that do (i.e., text). It has drawn interest as an additional method of lip reading, though. It is still seen as a difficult subject because, homophones might have identical lip motions and varied vocal qualities depending on it.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

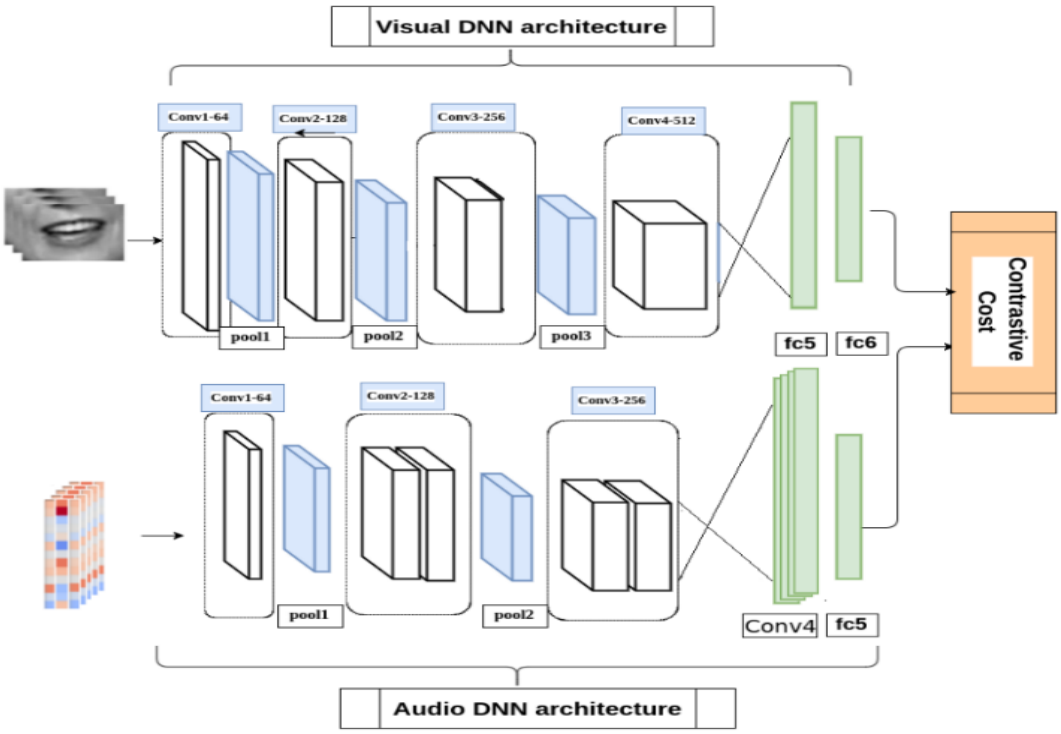
International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 2, February 2023

DOI: 10.17148/IJARCCE.2023.12229



To efficiently determine the connection between temporal information for various modalities, the suggested architecture would combine spatial and temporal information. CNN matches the input video dataset to the pool of speech patterns, and audio is produced as the output. To show that the model pays particular attention to the lip region, we depict the activation maps from the visual encoder



The audio is recovered using the synchronizing audio and video. Using lip motions to decipher audio during a criminal investigation. The development of datasets with improved quality also makes it easier for subsequent lip-reading processes like feature extraction and lip-reading recognition to proceed without difficulty.

- Video which has lost it's audio.
- Entertainment Industry such as Film making, Animated movies.
- CCTV's which records only video

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 2, February 2023

DOI: 10.17148/IJARCCE.2023.12229

II. LITERATURE REVIEW

Literature survey 1:

Ting, Chai, Hongyang, and Tooling (ELSEVIER), 2022, "A Comprehensive Integrated Dataset for a Machine-Learning-based Lip-Reading Algorithm". A video from the publicly accessible English lip-language data collection was utilised in this experiment, which was recorded by 31 persons. Each participant recorded 1000 phrases. Each movie is roughly 3 seconds long and has a 25 frames per second frame rate. For frame splitting and face decoding, it makes use of SK-video and 3D-CNN. has a less training dataset and just 64% accuracy.

Literature survey 2:

Lip2audspec: Speech reconstruction from silent lip movement video, ACM,2020 Lip-to-speech synthesis uses a video of a talking face in silence to simulate an audio voice. Lip2Speech does not require extra annotations, unlike traditional visual speech recognition tasks that do (i.e., text). Only 3000+ words were learned, it only utilizes the Lip2audspec dataset for face decoding and audio encoding, and its accuracy is 62%.

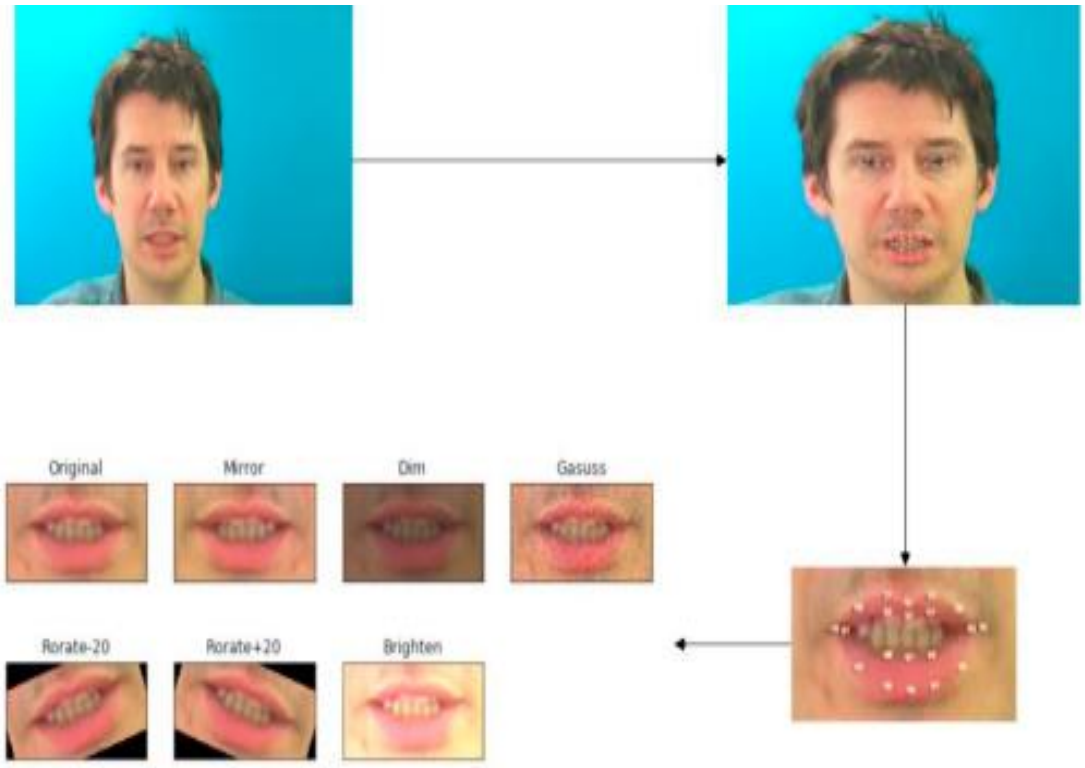
Literature survey 3:

For the first time to tackle this issue of unrestricted lip-to-speech synthesis. It has a 58% accuracy rate, employs 2D-CNN, Vid2Speech datasets for face decoding, and breaks ordering while decoding speech representation.

III. METHODOLOGY

Proposed Method

This experiment used a video from the publicly available English lip-language data set GRID, which was recorded by thirty-four persons (record number 21's video is missing). Each participant recorded 1000 phrases. Each movie is roughly 3 seconds long and has a 25 frames per second frame rate. The first step is to use scikitvideo to extract frames after gathering the videos. The processed frame pictures are then subjected to face detection using the Idlib face detector removing photos with no figures, more than one figure, and a disproportionately low number of figures. The preserved photos were then subjected to lip detection, and the extracted lip images were cropped. In order to increase the data, the retrieved lip picture was improved.

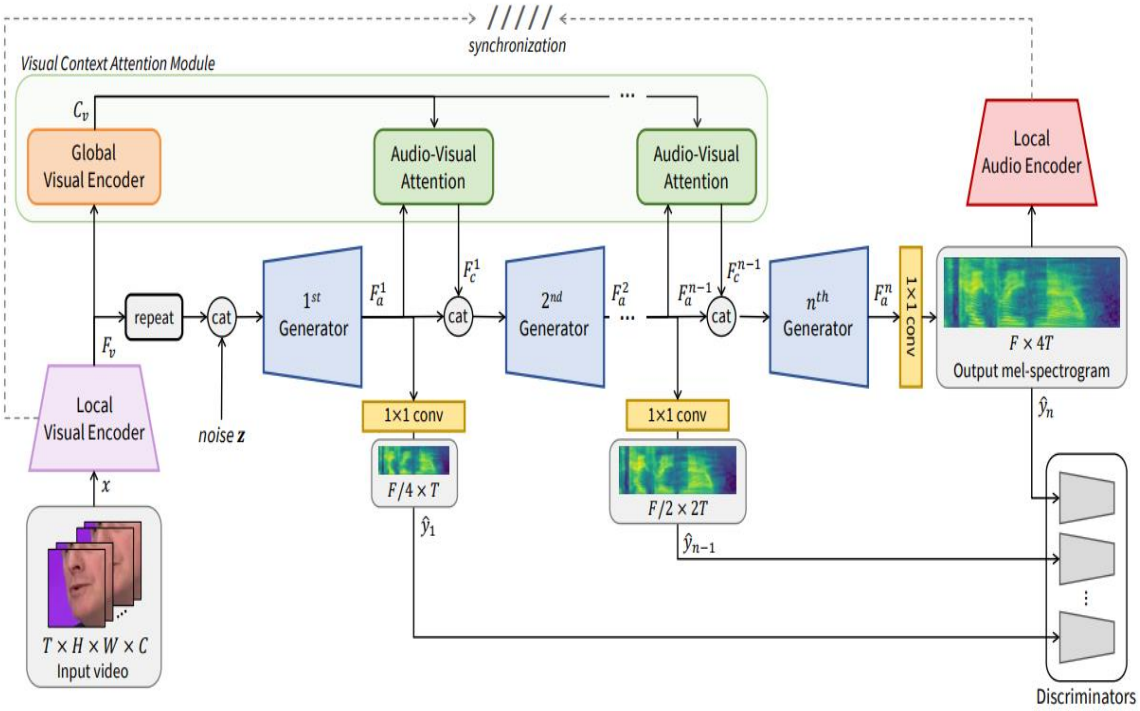


IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

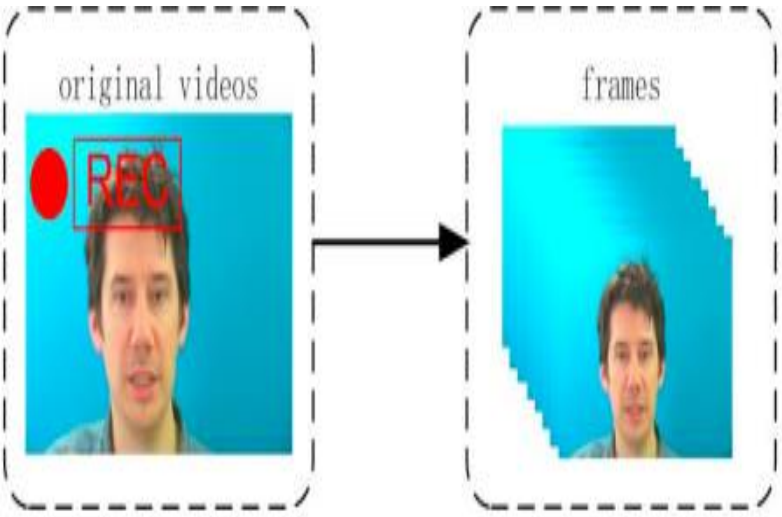
ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 2, February 2023

DOI: 10.17148/IJARCCE.2023.12229



Frame Extraction

Users can call different video processing algorithms using the Python video processing package Skvideo. A video read/write module called Skvideo.io uses FFmpeg/LibAV as its back end. It will parse the video information using the accessible back end and the proper probe (FFprobe, avprobe, or even mediaInfo). The video is changed into a series of pictures using the FFmpegReader function so that it may be processed later. Additionally, it returns the video form in terms of frames, pixels per inch, height, and width.



Face Detection

Idlib is a cross-platform library created in C++ that includes a variety of machine learning techniques. It is quite easy to use. Embedded systems, robotics, high speed computing, and other areas of business and academia are just a few of the areas where Idlib is being employed extensively. The following are the steps for face detection using Idlib: To implement face detection, utilize the method Idlib.get frontal face detector. When a frame has no faces, it is eliminated; however, the frame that contains at least one face and an appropriate amount of faces will be kept.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 2, February 2023
DOI: 10.17148/IJARCCE.2023.12229

Data Augmentation

The experiment demonstrates that the model's influence and generalizability are stronger the more samples there are. Data augmentation can increase the amount and variety of samples by producing more comparable (equally effective) data based on sparse data. Once the necessary lip pictures have been acquired, data improvement may be employed to reduce overfitting and increase the model's resilience. In this study, data augmentation is carried out using the typical picture conversion series, which includes Gaussian noise, horizontal and regular mirroring, brightening and darkening of the original lip image, etc.

IV. CONCLUSION

The main objective of the project described in the paper is to develop a solution to add speech to a silent video of a person speaking. The authors of the paper aim to demonstrate the feasibility of this solution by presenting the different components or modules involved in the project. The initial step of the project is to properly identify the face and mouth region in the video. To achieve this, the authors use computer vision libraries such as Idlib and others to accurately detect these regions. This information is then used to recover the audio from the video. The recovered audio is then synchronized with the video to create a complete representation of the person speaking. The authors also explore the use of lip motions to decipher the audio during a criminal investigation. They extract frames from real videos using SK Video and demonstrate the ability to identify and track the lip movements of a speaker to recover the audio. Hence, the authors present the design concept for creating a dataset to support this project. The goal is to provide a comprehensive set of data that can be used to train machine learning algorithms and improve the accuracy and reliability of the speech recovery solution.

REFERENCES

[1] Ting, Chai, Hongyang, and Toolings (ELSEVIER), 2022, "A Comprehensive Integrated Datasets for a ML based Lip-Reading Algorithm
[2] K Prajwal, C.V Jawahar,Mukhopadhyay, and Vinay P. Namboodiri. Speech to lip generation in the real world. ACM 2020 Proceedings.
[3] Hassan A, Liang Cao, and Nima Mesgarani. Lip2audspec: Speech reconstruction using video of silent lip movements. In 2018, IEEE
[4] Minsu Kim, Joanna Hong, Yong Man Ro,Lip to Speech Synthesis with Visual Context Attentional GAN,2022
[5] Jungil Kong, Kim Jaehyeon, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high-fidelity speech synthesis,2020
[6] https://www.ieee.org/
[7] https://www.wikipedia.org/
[8] https://www.youtube.com/
[9] https://www.researchgate.net/publication/359757377
[10] https://www.geeksforgeeks.org/
[11] https://github.com/

