IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SPEECH RECOGNITION USING LIP READING

[1]Aditya Maheshwari, [2]Sagar Kuvar, [3]Omshree Shetty, [4]Shantanu Ojha, [5]Varunakshi Bhojane

[1]Student, [2]Student, [3]Student, [4]Student, [5]Assistant Professor
[1]Department of Computer Engineering,
[1]Pillai College of Engineering, Navi Mumbai, India

Abstract: There is an audio lag/packet loss in the video streaming/meeting services and video platforms, and sometimes listeners cannot understand what their host is talking about because of network connectivity issues. This generates delays/communication errors in the meeting/streaming and the audience cannot understand the message. Lip-Reading is an art that helps a listener understand, even without hearing it and just observing the Lip-movement, what a speaker tries to say. It's this combination of using a person's lip gestures, facial eloquence, and other nonverbal cues to "read" speech. We describe a recurrent neural network-based system for text-to-speech (TTS) synthesis that can generate speech audio in the voice of different speakers by reading their Lip-movements, including those unseen during training. Our model will demonstrate these capabilities by reading the speaker's face by extracting the Lip features. Next, the model will convert the Lip-movements to sound waves using relevant APIs. Following this, our model will generate artificial speech, synchronous to the speaker.

***Index Terms* - Lip Reading, text-to-speech, Lip2Wav, Wav2txt.**

I. INTRODUCTION

Technology has become an important aspect of human life. It has a great influence on many facets of our day-to-day life and has also helped improve our environment. The introduction of technology in communication with the invention of mobile phones and the internet has caused people to rely on it to improve their way of working and also provide easy ways to use various applications to enhance their standard of living. This establishment of technology in one's life has enhanced not only the way people communicate or trade goods but also in a variety of fields such as medicine, agriculture, home security, etc.

With the increase in demand of online mode for lectures, schools, conferences and meetings during COVID-19, there is an increase in cases where there can be lag in the connection due to poor internet connectivity or due to bandwidth. Also for disabled people to use the online technology speech reproduction is necessary.

Lip reading is the main concept used to reproduce expressions. Lip reading, also known as lipreading or speechreading, is a method of understanding speech that involves visually interpreting gestures of the mouth, ears, and tongue when normal sound is unavailable. It also relies on information provided by the context, language awareness, and any residual hearing. While deaf and hard-of-hearing people use lip reading the most, most people with normal hearing process some speech information from the sight of a moving mouth.

Deep learning (also known as deep structured learning) is part of a broader family of machine learning methods based on artificial neural networks with representation learning. Learning can be supervised, semi-supervised or unsupervised. It is a subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks.

Abbreviations and Acronyms

API: Application Program Interface is a package which can be imported into the respective program being developed.
CUDA: Compute Unified Device Architecture is a parallel computing devised by Nvidia.
TTS: Text to Speech Engine used for converting text into respective speech.

II. LITERATURE SURVEY

A. Deep voice 2: Multi-speaker neural text-to-speech. Advances in neural information processing systems: Andrew Gibiansky, Sercan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. It was published on 24 May 2017.

B. Signal estimation from modified short-time fourier transform: The Authors are Daniel W. Griffin and Jae S. Lim. It was published on 16 April 1983.

C. Tcd-timit: An audio-visual corpus of continuous speech: Naomi Harte and Eoin Gillen. It was published on 26 February 2015.

D. Deep residual learning for image recognition: Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. It was published on 10 Dec 2015.

2.1 Summary of Related Work

Table 2.1: Literature Survey Summary

Date	Title	Author Name	Remarks
Oct 28th 2020	Wav2Lip: Accurately Lip-sync Videos to Any Speech	K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, C V Jawahar	Syncing lip with the video and making and making proper video
Jul 6th 2020	Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis	K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, C V Jawahar	Made complete synthetic voice of the user from the video (without the original audio)
Dec 10th 2018	Deep lip reading: a comparison of models and an online application	Triantafyllos Afouras, Joon Son Chung, Andrew Zisserman	Comparison of new model with previously existing models
Aug 9th 2018	The conversation: Deep audio-visual speech enhancement	Triantafyllos Afouras, Joon Son Chung, Andrew Zisserman	They made voice more clear and speech enhancement

III. PROPOSED WORK

We aim to develop a model which can facilitate alternatives over the air transmission of speech. This model will read lips via video of the speaker and generate text of the speech **"Fig. 1"**. This text will then be transmitted over the air to the receivers end and will be converted to the respective audio format **"Fig. 2"**.

3.1 Proposed System Architecture

3.1.1 Modules on Speaker's End

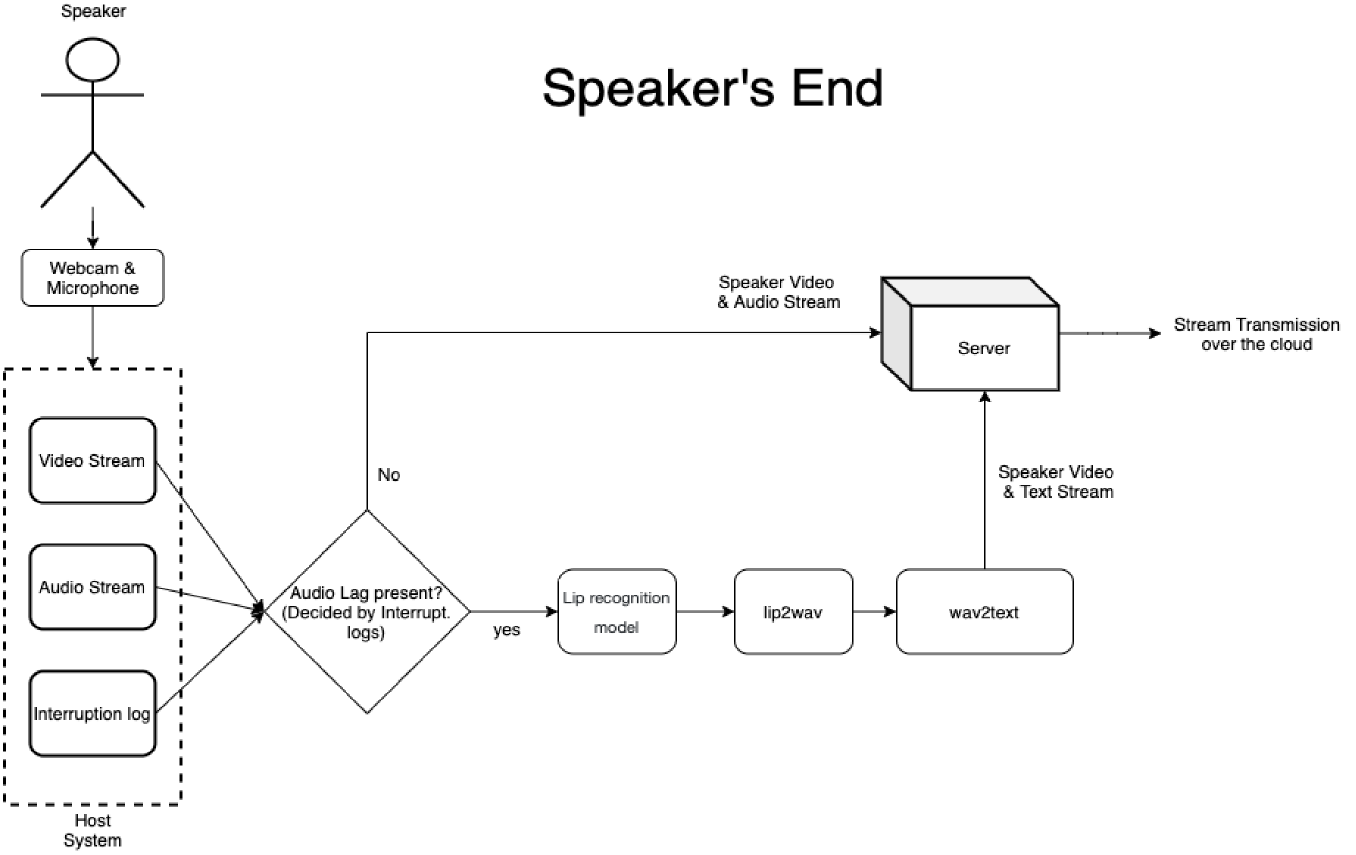

Figure 1: Proposed System on Speaker's End

- **Lip Recognition Model:** Used to determine the exact coordinates of the speaker's lips in the video stream. This method makes use of computer vision. To begin, the model employs a haar-cascade classifier to read face coordinates in the frame. Once a face has been read in the frame a coordinate set is returned. Using this, the model then starts reading a set of predefined coordinates in the box to determine the specific edges of lips. Once a set of coordinates is returned for the lips being read, the model then passes these coordinate parameters to the Lip2wav API.
- **Lip2Wav:** This API utilises physics. Over here the API performs the operation of tracking lip movement angles and generating a respective sound waveform. It is a highly trained convoluted neural network model processing several neuron features over multiple hidden layers. It is specifically configured to optimize its operating time in CUDA enabled devices. CUDA is a proprietary tool by NVIDIA.

- **Wav2Txt:** This API tool will be reading the .WAV file generated previously by the lip2wav API. It will input each frame of audio from the file and then, according to the respective amplitude and frequency of the instances, will produce a grammar parse tree which will help produce syntactically proper words. These words are then stitched together into sentences and then saved into the .TXT file. It is a highly trained recurrent neural network model processing each neuron feature in a given waveform order.

3.1.2 Modules on Receiver's End

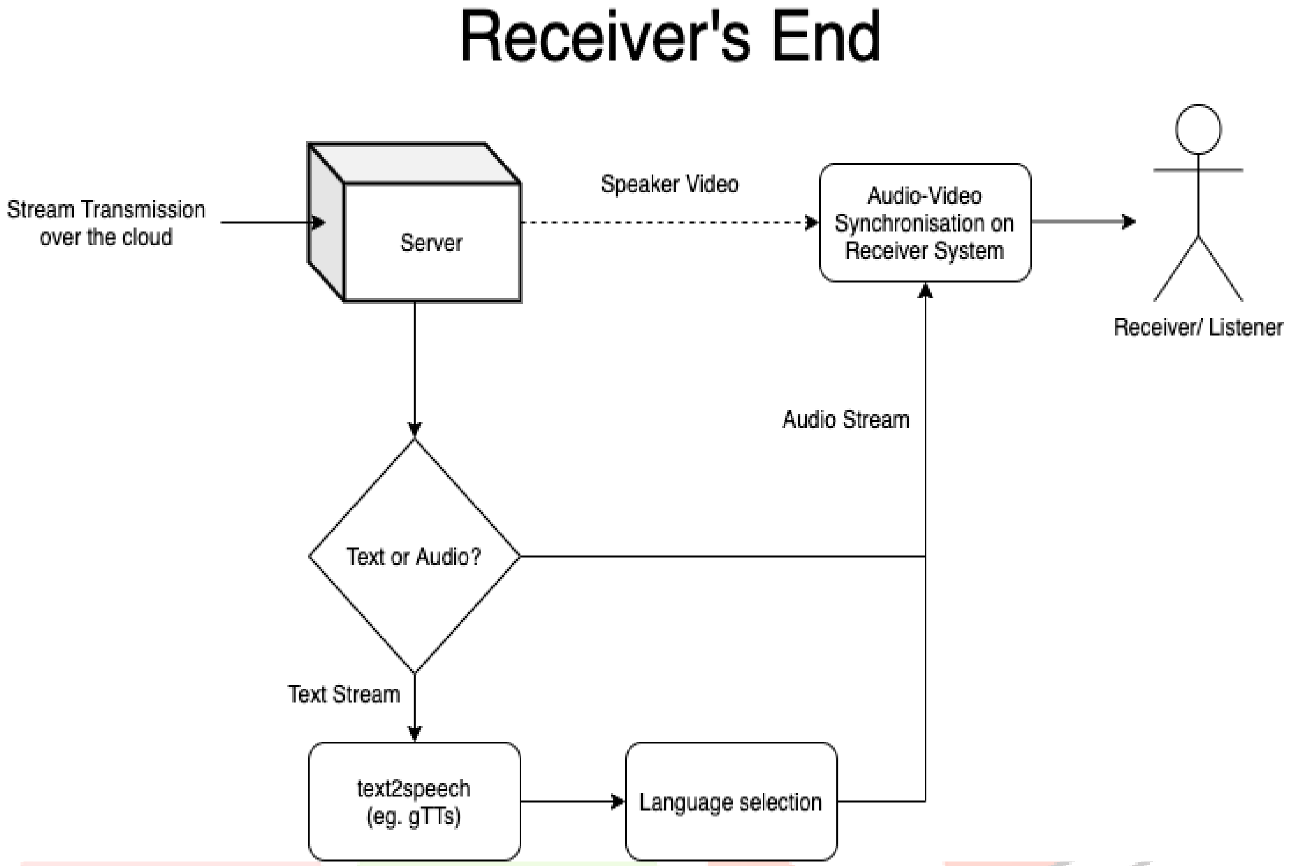

Figure 2: Proposed System on Receiver's End

On the receiver's end, the following constituents process the received stream from the server to present it to the listener. Components:

- **Text2speech (or TTS):** This API processes the input text from the stream and converts it into respective speech waveform
- **Language selection:** This optional step gives the listener an option to convert the respective speech into a language of their choice. The speech will then be played in the language mode selected.

IV. Requirement Analysis

Both the speaker's and receiver's ends are cooperating in a complementary manner. For the livestream to function, both ends must be time synchronized through the internet. This framework is intended to run on any pre-existing platform without the need to install system-specific modules.

This system will require the user to accept several permissions for access of camera, microphone(optional) and screen. Access to the camera is required for recording and transmitting video footage of the speaker. Screen reading is utilised for reading the video stream on the receiver or listener's end (for the event of a delayed stream or pre-recorded video where file cannot be read).

This device specifically employs a camera to record or read the speaker's live video stream. The speaker is granted control over the quality of feed that they choose to transmit; the higher the streaming quality, the more processing power the device requires. If the model is unable to equalize the picture illumination, the device detecting lip movements would also prompt the user to adjust lighting conditions.

The receiver is given the option of choosing the receiving streaming quality, so that receiver is able to manage quality based on resources available on their system. This system requires a CUDA-enabled GPU (preferably by Nvidia Corporation), be it cloud-based or on the user's own system, for the processing of video codecs and merging of the audio-video files.

4.1 Dataset

This application's development would necessitate datasets of pre-recorded video streams containing the faces of people speaking and looking at the camera from different angles on the X and Y axes. Only if our face and lip detection models can detect respective features for at least 70% of the feed time will the dataset be suitable for training.

V. Technical Applications

- **Speech impaired individual:** People who are disabled who cannot communicate with others. They communicate using sign language but the drawback is not everyone understands the signs. This tool can be used to communicate easily.
- **Lectures:** During the online lectures there may be an issue of connectivity either from the student or from the professor's side. At such times only the lip movement of the speaker can be seen. In such cases to prevent the discontinuity of the flow this tool

can be used.

- **Videographer:** Sometimes there can be background noise while capturing the video and it ruins the quality of the audio and video, this tool can be used to recover the audio for the same.
- **Recovering Vintage footage:** It is hectic work to find the audio of some old videos either because of unavailability or the file is damaged. This can be easily retrieved by using this tool.

VI. Acknowledgment

We want to give our heartfelt gratitude to Pillai College of Engineering, New Panvel, for providing us with the resources and support us with our project. This would not have been possible without the support from our college, and also, thank you to everyone who has contributed to this research.

References

[1] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, C V Jawahar, Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis. 2020

[2] Triantafyllos fouras , Joon Son Chung, and Andrew Zisser- man. The conversation: Deep audio-visual speech enhancement. arXiv preprint arXiv:1804.04121, 2018.

[3] Triantafyllos fouras, Joon Son Chung, and Andrew Zisser- man. Deep lip reading: a comparison of models and an online application. In INTERSPEECH, 2018.

[4] Triantafyllos fouras , Joon Son Chung ,and Andrew Zisser- man. Lrs3-ted: a large-scale dataset for visual speech recognition. arXiv preprint arXiv:1809.00496, 2018.

[5] Hassan Akbari, Himani Arora, Liangliang Cao, and Nima Mesgarani. Lip2audspec: Speech reconstruction from silent lip movements video. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2516–2520, 2017.

[6] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3444–3453. IEEE, 2017.