IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

LIP READING USING CNN AND BI-LSTM

[1]DivyaPrabha, [2]Sagar Sonale T V, [3]Manu G, [4]Gokul Chandra Reddy, [5]Manoj M O

[1]Associate Professor, [2]Student, [3]Student, [4]Student, [5]Student
Electronics And Communications Engineering
Sri Siddhartha Institute of Technology, Tumkuru, Karnataka, India, 572105

Abstract: The CTC-CNN-Bidirectional LSTM-based Lip Reading System seeks to increase speech recognition accuracy. To more accurately identify spoken words from lip movements, this system integrates Convolutional 3D Neural Networks with Bidirectional Long Short-Term Memory (LSTM) architecture and Connectionist Temporal Classification (CTC). Through the integration of 3D CNN and Bidirectional LSTM, lip dynamics' temporal and spatial properties are efficiently captured. On the test dataset, the system attains a word error rate of just 16.6%
and a character error rate of 6.2% after 92 epochs. In real-world settings, this technology could greatly improve user interfaces, safety, and communication for the hard of hearing.

Index Terms – Audio-visual Processing, Feature Extraction, Lipnet, Lipreading, Robustness, Deep Learning

I. INTRODUCTION

Communication is fundamental to the existence and survival of humans to express their ideas and feelings to reach a common understanding among the people. Lip Reading plays a supreme role in grasping human speech particularly for the listeners with hearing impairment. This serves as a hearing aid for the hearing impaired particularly interacting with people with no knowledge of sign language. For building this system various techniques and approaches have been adopted by so many researchers. As Artificial Intelligence outperforms well in majority of the domains, this is where deep learning comes into play in the field of automatic lipreading. Deep Learning is a subset of machine learning concerned with algorithms which mimics the structure and function of the human brain called artificial neural networks. Deep learning models are trained by using a large set of labelled data and multi layered neural network architectures which can be visualized as a set of points each of which decides based on the inputs to the node. These algorithms learn gradually about the image as it goes through each neural network layer. Initial layers learn how to detect low-level features and following layers combine features from initial layers into a more holistic representation. Far from traditional methods of machine learning techniques, deep learning classifiers are trained through automatic feature learning. ALR (Automatic Lip Reading) systems are comparatively behind when compared to the Automatic Speech Recognition (ASR). In automatic Lip reading, different approaches have been used to pre process the raw data, extract the features and train the model in such a way speech is converted to text through Visual cues. Just like speech recognition, these lip-reading systems encounter several challenges which may be in the form of background noise, colour of the skin, the intensity of a person's speech and many more which can be dealt with and solved with the help of various deep learning algorithms. Communication is the only key to survival, wanted that to be possible for everyone in the world.

II. PROBLEM STATEMENT

LIP READING, A VALUABLE SKILL FOR INDIVIDUALS WITH HEARING IMPAIRMENTS AND APPLICATIONS IN HUMAN-COMPUTER INTERACTION, FACES SIGNIFICANT CHALLENGES IN ACCURACY AND ADAPTABILITY. TRADITIONAL LIP READING METHODS OFTEN STRUGGLE TO CAPTURE THE NUANCED DYNAMICS OF LIP MOVEMENTS, LEADING TO LIMITATIONS IN REAL-WORLD SCENARIOS. THE EXISTING GAP IN ACHIEVING ROBUST AND PRECISE LIP READING MOTIVATES. DESPITE SIGNIFICANT ADVANCEMENTS IN AUTOMATIC SPEECH

RECOGNITION (ASR), ACCURATELY UNDERSTANDING SPOKEN LANGUAGE SOLELY FROM VISUAL CUES REMAINS A DAUNTING CHALLENGE. LIP READING, OR SPEECHREADING, OFFERS A TANTALIZING GLIMPSE INTO A WORLD WHERE VISUAL INFORMATION UNLOCKS THE SECRETS OF SPOKEN COMMUNICATION.

III. PROPOSED SYSTEM

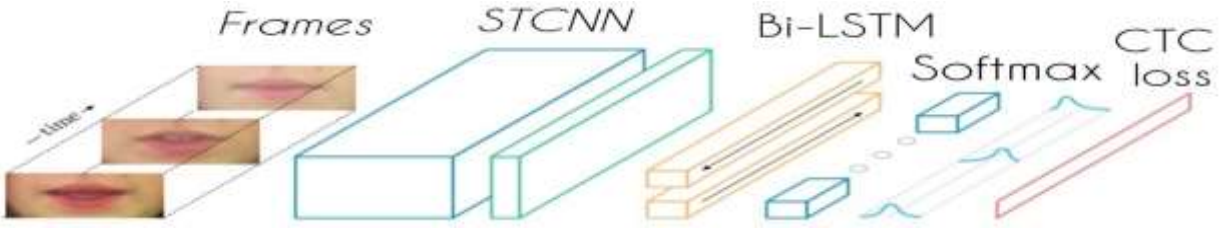

Fig:1 Proposed Block Diagram

A. Dataset Architecture

The architecture of dataset grids for lip reading comprises a structured framework that integrates diverse datasets, each contributing unique linguistic and environmental contexts. These grids are designed to accommodate datasets of varying sizes, demographics and recording conditions, providing researchers with a comprehensive training and evaluation platform. Central to this architecture is the emphasis on diversity, scalability, annotation quality, and standardization. Diverse datasets ensure that deep learning models are exposed to a wide range of linguistic variations and environmental factors, enhancing their robustness and generalization capabilities. Scalability enables researchers to access datasets suitable for their specific research goals and computational resources. High-quality annotations, including accurate phonetic transcriptions and precise temporal alignment of audio and video streams, are essential for effective model training and evaluation. Standardization efforts ensure consistency and comparability across different datasets, facilitating interoperability and reproducibility in lip reading research. Overall, the architecture of dataset grids serves as a foundation for advancing lip reading technologies, providing researchers with the tools and resources needed to develop and evaluate robust and generalizable deep learning models. Generally the Grid audio-visual dataset is the widely used data for audiovisual or visual speech recognition tasks. Each sentence of Grid has a fixed structure with six words structured as: command + colour + preposition + letter + digit + adverb. For example ”set red with m six please”. The dataset has 51 unique words. The 51 words include four commands, four colors, four prepositions, 25 letters, ten digits, and four adverbs. Each sentence is randomly chosen combination of these words. The duration of each utterance is 3 seconds. For overlapped speakers (seen speakers), we randomly select 255 utterances from each speaker to test set and remaining utterances are used for training. For speaker independent case (unseen speakers), we held out two male speakers (1 and 2) and two female speakers (20 and 22) for evaluation and remaining speakers are used for training.In the dataset it contains ~988 videos spoken by an single person with duration of 3 sec. The video frame width is 360 and frame height is 288, frame rate of each video is 25.0 fps. taken 450 videos for training the model and remaining videos for testing. The purpose for compromising with the dataset by choosing only one person video because of lack of computational resources.

B. 3-D Convolutional Neural Networks

3D convolutional neural networks (CNNs) have emerged as powerful tools for spatiotemporal feature extraction in tasks such as action recognition, video analysis, and, notably, lip reading. Unlike traditional 2D CNNs, which operate on spatial information in images, 3D CNNs extend their convolutional operations into the temporal domain, allowing them to capture both spatial and temporal dependencies in video data. By processing consecutive frames over time, 3D CNNs can effectively model motion patterns and temporal dynamics, making them well-suited for tasks involving sequential data like videos.The architecture of a 3D CNN typically consists of multiple convolutional layers followed by pooling layers for spatial and temporal down sampling, as well as fully connected layers for classification or regression tasks. One key advantage of 3D CNNs is their ability to automatically learn spatiotemporal features from raw video data without the need for manual feature engineering. However, training 3D CNNs can be computationally intensive due to the increased complexity of processing volumetric data. Nonetheless, their ability to capture both spatial and temporal information makes 3D CNNs a valuable tool for various applications requiring spatiotemporal understanding of data.

C. BI-LSTM

Bidirectional Long Short-Term Memory networks (Bi-LSTMs) are a variant of recurrent neural networks (RNNs) designed to capture long-range dependencies and temporal dynamics in sequential data. Unlike traditional LSTMs, which process input sequences in only one direction (either forward or backward), Bi-LSTMs employ two LSTM layers running in parallel: one processing the input sequence from the beginning to the end, and the other processing it from the end to the beginning. This bidirectional processing enables the network to incorporate information from both past and future time steps, allowing for a more comprehensive understanding of the input sequence. Bi-LSTMs are particularly effective in tasks where contextual information from both preceding and succeeding elements is crucial for accurate prediction or classification. For instance, in natural language processing, Bi-LSTMs can capture the context of a word based on both its preceding and succeeding words, leading to better performance in tasks such as sentiment analysis, named entity recognition, and machine translation. Additionally, in speech recognition and synthesis, Bi-LSTMs can leverage information from both past and future audio frames to improve accuracy in phoneme recognition and speech generation tasks. Overall, Bi-LSTMs offer a powerful mechanism for modeling sequential data with bidirectional dependencies, making them well-suited for a wide range of applications in fields such as natural language processing, speech recognition, and time series analysis.

D. Connectionist Temporal Classification(CTC) loss:

CTC loss is a key component in sequence-to-sequence learning tasks, particularly in fields like speech recognition, handwriting recognition, and machine translation. It addresses the challenge of aligning variable-length input sequences with variable-length output sequences, where the correspondence between input and output elements may not be one-to-one. CTC loss operates by maximizing the likelihood of correct alignment between input and output sequences, without requiring explicit alignment information during training. It achieves this by summing over all possible alignments and computing the probability of each output sequence given the input sequence. The core idea of CTC loss is to introduce a blank label and allow repeated occurrences of labels in the output sequence, representing multiple input elements being mapped to the same output label. During training, the model learns to predict not only the correct labels but also the necessary repetitions and alignments to match the input sequence. By optimizing the CTC loss, the model learns to produce output sequences that are aligned with the input sequence, enabling it to effectively handle variable-length inputs and outputs. This makes CTC loss a versatile and powerful tool for sequence learning tasks where explicit alignment information may be unavailable or difficult to obtain.

IV. METHODOLOGY

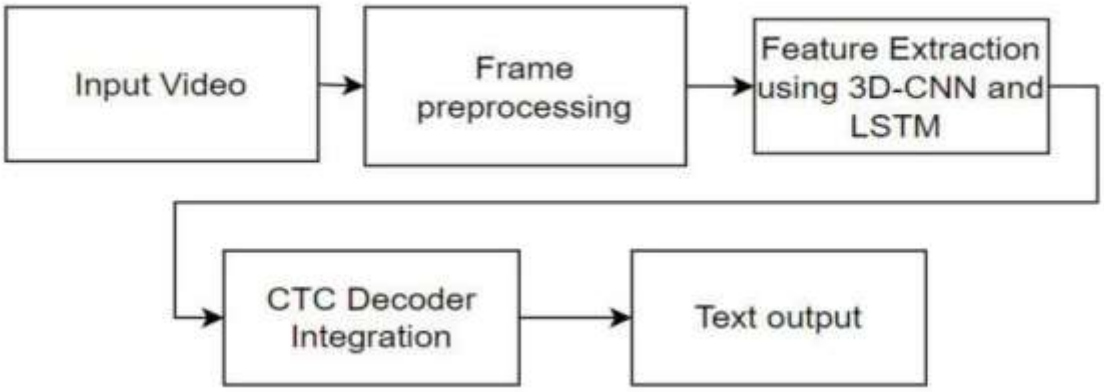

Fig:2 System Architecture

It begins with the GRID dataset, which provides high-quality, synchronized audio-visual data. Each video clip, containing clear frontal views of a speaker's lip movements, was processed to focus on the lip region. The original frames, with a resolution of 366x288 pixels, were resized to a 45x145 pixel region centred around the lips. This preprocessing step isolates the relevant area, reducing computational complexity and improving feature extraction accuracy. Each 3- second video, shot at 25 frames per second (fps), was converted into 75 grayscale frames to simplify the data and enhance processing speed. The pre-processed

video frames were fed into a 3D-CNN designed to extract spatiotemporal features from the lip movements. The 3D-CNN captures the temporal dynamics and spatial features within a sequence of frames. It applies 3D convolutional filters across the spatial and temporal dimensions of the input data, generating feature maps that encapsulate the motion and appearance of the lips over time. The feature maps generated by the 3D-CNN were then input into a Bi-LSTM network. The BiLSTM processes these sequences in both forward and backward directions, effectively capturing the temporal dependencies and contextual information from the entire sequence of frames. This dual-direction processing allows the model to understand the full context of lip movements, which is crucial for accurate speech recognition. The model was trained using the Connectionist Temporal Classification (CTC) loss function, which is well-suited for sequence-to-sequence learning where input and output sequences have different lengths. CTC loss enables the model to learn alignments between the lip movements and the corresponding phonetic labels without requiring explicit frame-level annotations. It maximizes the likelihood of the correct sequence by considering all possible alignments during training.

Lip reading relies only on the visual analysis of videos which is used to interpret the language. It involves computer vision and deep neural networks, for example, are machine learning approaches to track lip movements, extract the various features, train and predict phonemes or words from the lip movements in the videos. This is the methodology developed for lip reading model. The methodology includes various steps towards developing the model. In brief, a deep neural network is developed with CNN as frontend, LSTM (Recurrent Neural Network) as backend and Connectionist Temporal Classification (CTC).After the dataset is preprocessed, the gray scale frames are then sent through the deep neural network. The steps ahead are as follows.

1) Input Video:

The input videos plays a critical role in determining the models performance. The initial resolution of each video 366 x 288 pixels. The video plays at 25fps this frame rate is chosen to ensure smooth motion capture. It provides enough temporal quality to capture the rapid movement of the lips during speech, which is crucial for accurately interpreting phonemes and words .The data set contains one speakers of 988 videos where 450 videos are used for training and remaining videos are used for testing the model. The speaker speaks six words : command + colour + preposition + letter + digit + adverb. For example ”set red with m six please”. Which as 51 unique words.

2) Frame Preprocessing:

The system begins with the crucial step of frame processing, essential for preparing the input video data for subsequent analysis. Each input video frame originally has a resolution of 366x288 pixels. For effective lip reading, the system processes these frames to focus on the lip area. The first step in this process involves converting the 3 second video clip into 75 individual frames. This conversion is based on the video’s frame rate, ensuring that the temporal sequence of lip movements is captured accurately over the entire duration. After breaking down the video into frames, each frame undergoes cropping to isolate the lip region. The cropping operation adjusts the frame size to a width of 45 pixels and a height of 145 pixels, focusing specifically on the area around the lips. This adjustment is crucial as it ensures that the neural network concentrates on the most informative part of the image, where the lip movements that signify speech occur. Following the cropping, the RGB video frames are converted to grayscale but converting to grayscale reduces this to a single channel. This conversion simplifies the processing pipeline while retaining the essential visual information needed for lip reading.

Fig 3: Input Image

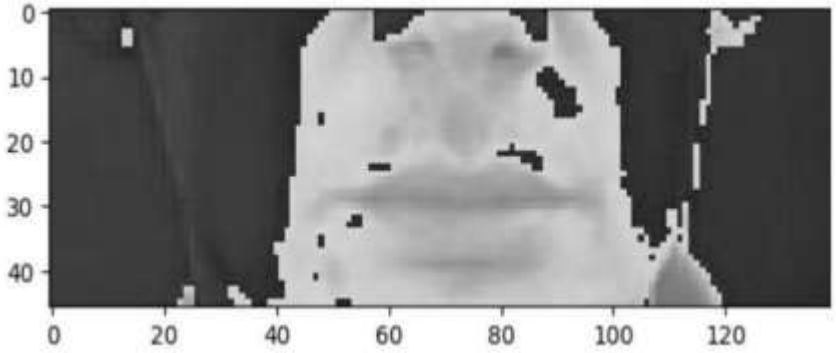

Fig 4: Image After Frame Preprocessing

3) Convolution Neural Network and BI-LSTM Feature Extraction:

Following the frame preprocessing steps, the preprocessed video frames are fed into a Convolutional Neural Network for feature extraction. CNNs are renowned for their ability to extract hierarchical features from visual data, which makes them particularly suitable for tasks such as image classification, object detection, and importantly, lip reading. The architecture of CNN typically includes several convolutional layers followed by pooling layers. The convolutional layers apply convolutional filters to the input frames, performing operations that detect various features such as edges, textures and shapes. Each layer extracts increasingly complex features. Starting from simple edges in the early layers to more sophisticated patterns in the deeper layers. This hierarchical feature extraction is crucial for understanding the detailed movements of the lips. Maxpooling Layers Interspersed between convolutional layers, Maxpooling layers reduce the spatial dimensions of the feature maps. This process, known as down sampling, helps in reducing the computational load and the number of parameters in the network. Maxpooling layers also help in making the detected features more robust to spatial translations. In the context of lip reading, the CNN processes the pre-processed grayscale video frames, systematically extracting relevant visual features that encode information about lip movements and phonetic cues. By focusing on the lip region, the CNN captures the dynamic changes and patterns that correspond to different phonemes and words. While CNNs are adopt at extracting spatial features, lip reading also requires understanding the temporal dynamics of lip movements. This is where Recurrent Neural Networks (RNNs), specifically Long ShortTerm Memory (LSTM) networks, come into play. Bidirectional LSTM To effectively capture the temporal dependencies and sequential patterns in the input video data, our model employs a Bi-LSTM network. Traditional LSTMs process sequences in one direction (either forward or backward), but Bi-LSTMs process sequences in both directions. This bidirectional approach allows the model to have a complete context of the sequence, improving its ability to capture long-range dependencies and contextual information from the input frames. The integration of CNN and Bi-LSTM, when the pre-processed grayscale frames are first passed through the CNN layers. The CNN extracts spatial features form each frame, creating a sequence of features maps that represent the detailed lip movements over time. Then sequence of feature maps produced by the CNN is then fed into the Bi-LSTM network. The Bi-LSTM processes these sequential features in both forward and backward directions, capturing the temporal dynamics and dependencies between frames. This enables the model to understand the progression of lip movements, essential for accurate speech interpretation.

4) Connectionist Temporal Classification (CTC) Loss:

To train the lip reading model effectively, the system employs the Connectionist Temporal Classification (CTC) loss function. CTC loss is specifically designed for sequence-to-sequence learning tasks, which are characterized by variable-length input and output sequences and the absence of a direct, one-to-one correspondence between them. This makes CTC particularly well-suited for tasks like speech recognition and lip reading, where aligning input data with output labels is inherently complex. in sequence-to-sequence tasks, such as lip reading, the input sequence (video frames of lip movements) and the output sequence (phonetic labels or spoken words) often have different lengths. Furthermore, there is no straightforward way to align each frame directly with a specific phoneme or word because the duration of phonemes can vary widely. Traditional loss functions, which assume fixed alignment between input and output, are inadequate for this scenario.

The CTC LOSS Equation given below:

CTC path probabilities denoted as α(t,u)
Which represent the probe of producing label sequence y1,y2, Up to time step t.
$\alpha(t,u)=\sum \alpha(t-1,j).\text{softmax}(ht)j$
$\alpha(t,U)=\alpha(t-1,U).\text{softmax}(ht)\text{blank}$
$\alpha(T,U)=\alpha(T-1,U)$

In this equation:

- Softmax (ht)j represents probability of label j at time step t
- Softmax(ht) blank represents probability at time step t
- T is the length of the input and U is the length of the target label sequence.
- The sum over j in first eq. Represents sum of all possible labels.

Parameters	Count
3-D Convo Layers	3
BI-LSTM	1
CTC LOSS	1
Max-polling	3
SoftMax	1
Kernal(Orthogonal)	2

Table:1 Parameters

Parameter	Size
Learning Rate	0.0001
Optimizer	Adam
Epoch	96

Table: 2 Training Parameters

V. **RESULT ANALYSIS**

Word Error Rate(WER):
WER is a metric that measures the number of errors in spoken speech or lip movement transcriptions. It calculates the minimum number of changes (insertions, deletions, and substitutions) required to convert the model results into reference (ground truth) text. A lower WER means greater accuracy. WER is particularly suited for investigating the accuracy of lip reading models because it reflects the ability to successfully convert lip movements into words. The model achieved a WER of 16.6% which means that, on average, the model made 16.6% of errors when transcribing spoken words from lip movements.

Character Error Rate(CER):
CER works at the character level. Measure the percentage of misspellings in the copy compared to the reference. CER is often used when evaluating the performance of lip-reading models because it can capture errors at a more granular level, including misinterpretation of individual letters or characters. Character Error Rate (CER): The CER, at 4.16%, indicates that the model is reasonably accurate at recognizing individual letters or characters in lip movements. It suggests that the model can capture fine grained details..

Fig:5 Graphical Representation Of Results

Parameters	Percentage
WER	16.6%
CER	4.16%
WA	83.4%
CA	95.84%
Total Accuracy	95%

Table 3: Representation Of Results

The model has achieved a WER of 16.6% and a CER of 4.16% which implies a Word accuracy of 83.4% and character accuracy of 95.84% when it was tested on the 100 videos dataset of the same person on which the data is trained on. This indicates that the model is functioning correctly with the majority of the data when it comes to lip reading. The model achieved WER of 16.6%, CER of 4.16% and for a total of 96 epochs. From this, the model makes mostly correct predictions for the test data.

Correct Prediction

Fig 6: Correct Prediction
Video:swwv8p.mpg
REAL Text: set white with v eight please
Predicted Text: set white with v eight please
WER Score : 0.0
CER Score: 0.0

Wrong Prediction

Fig 7: Wrong Prediction
Video: bbaf3s.mpg
Real Text: bin blue at f three soon
Predicted Text: bin blue at j three soon
WER Score : 0.1666
CER Score : 0.04

VI. **CONCLUSION**

The project has been developed using the MERN stack, with React.js leveraging the virtual DOM concept to track changes in the application efficiently. Significant progress has been made in constructing the CTC-3D CNN-Bidirectional LSTM Lip Reading System. This system offers several advantages, including improved accessibility for the hearing impaired, enhanced feature extraction, and the integration of 3D-CNN with Bidirectional LSTM. The Bidirectional LSTM allows the system to extract features effectively even from videos with variable speeds. However, the system may face challenges when the number of speakers increases or when dealing with different accents. To further enhance the system, future work should focus on making it more resilient to various conditions, scaling it to handle larger datasets, and utilizing more computing power.

REFERENCE

[1] S. Fenghour, D. Chen, K. Guo and P. Xiao, "Lip Reading Sentences Using Deep Learning With Only Visual Cues," in IEEE Access, vol. 8, pp. 215516- 215530, 2020, doi:10.1109/ACCESS.2020.3040906.

[2] S. Fenghour, D. Chen, K. Guo, B. Li and P. Xiao, "Deep Learning-Based Automated LipReading: A Survey," in IEEE Access, vol. 9, pp. 121184- 121205, 2021, doi: 10.1109/ACCESS.2021.3107946.

[3] N. Deshmukh, A. Ahire, S. H. Bhandari, A. Mali and K. Warkari, "Vision based Lip Reading System using Deep Learning," 2021 International Conference on Computing, Communication and Green Engineering (CCGE), Pune, India, 2021, pp. 1-6, doi: 10.1109/CCGE50943.2021.9776430.

[4] K. Neeraja, K. Srinivas Rao and G. Praneeth, "Deep Learning based Lip Movement Technique for Mute," 2021 6th International Conference on Communication and Electronics Systems (ICCES), Coimbatre, India, 2021, pp. 1446- 1450, doi: 10.1109/ICCES51350.2021.9489122.

[5] S. M. M. H. Chowdhury, M. Rahman, M. T. Oyshi and M. A. Hasan, "Text Extraction through Video Lip Reading Using Deep Learning," 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART), Moradabad, India, 2019, pp. 240- 243, doi: 10.1109/SMART46866.2019.9117224.

[6] S. Pujari, S. Sneha, R. Vinusha, P. Bhuvaneshwari and C. Yashaswini, "A Survey on Deep Learning based Lip-Reading Techniques," 2021 Third International Conference on Intelligent Communication

Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, 2021, pp. 1286-1293, doi: 10.1109/ICICV50876.2021.9388569.

[7] T. Shirakata and T. Saitoh, "Lip Reading Experiments for Multiple Databases using Conventional Method," 2019 58th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE), Hiroshima, Japan, 2019, pp. 409- 414, doi: 10.23919/SICE.2019.8859932.

[8] F. S, C. J. S and N. Sripriya, "Convolutional Neural Network Based Lip Reading System for Hearing Impaired People," 2022 8th International Conference on Smart Structures and Systems (ICSSS), Chennai, India, 2022, pp. 1-5, doi: 10.1109/ICSSS54381.2022.9782208.

[9] M. Varshney, R. Yadav, V. P. Namboodiri and R. M. Hegde, "Learning Speaker specific Lip-to-Speech Generation," 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 2022, pp. 491-498, doi: 10.1109/ICPR56361.2022.9956600.

[10] B. Martinez, P. Ma, S. Petridis and M. Pantic, "Lipreading Using Temporal Convolutional Networks," ICASSP 2020 - 2020 IEEE International Conference Lip reading using deep learning Dept. of ECE SSIT,Tumkur Page 16 on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 6319- 6323, doi: 10.1109/ICASSP40776.2020.905384