Lip Reading Using Neural networks and Deep Learning

Akash Chandra
Dept. of CSE
VNRVJIET
Hyderabad, India
akash30kanna@gmail.com

Charitha Paruchuri
Dept.of CSE
VNRVJIET
Hyderabad,India
charithap242@gmail.com

Awalgaonkar Karthika
Dept.of CSE
VNRVJIET
Hydearabad,India
awalgaonkarkarthika@gmail.com

Pallamedi Yogitha
Dept. of CSE
VNRVJIET
Hyderabad, India
pallamediyogitha@gmail.com

***Abstract*—.This review investigates the latest advancements in efficient models utilized in lip reading. Lip reading, a method aimed at comprehending speech solely through visual cues derived from facial movements, particularly those of the mouth and lips, stands as an intriguing area of research, devoid of audio input. Given the vast array of languages spoken globally, coupled with the variations in pronunciation and articulation across different regions, developing a computer program capable of accurately deciphering spoken words based solely on visual lip movements presents a formidable challenge. Traditionally this process been approached in two stages a visual feature design followed by prediction. Recent advancements in deep learning have enabled end-to-end trainable models for lipreading. However, existing end-to-end models primarily focus on word classification rather than sentence-level sequence prediction. Human lipreading performance suggests the significance of capturing temporal context for understanding longer words in an ambiguous communication channel,motivated by this a model is developed that maps variable-length sequences of video frames to text using spatiotemporal convolutions.The trained lip-reading models were evaluated on various datasets and based on their accuracy to predict words the best performing model was implemented for real-time word prediction.**

Keywords— Lip reading, Deep learning, Neural networks, CNN architectures, Word prediction, Real-time application

I. INTRODUCTION

Lip reading, the interpretative process of spoken language through the observation of lip movements, has emerged as a prominent focus of research, largely propelled by the availability of extensive datasets such as the GRID corpus. Deciphering spoken language through lip movements, known as machine lipreading, poses a considerable challenge due to the intricate process of extracting spatiotemporal features from video data. Recent strides in deep learning techniques have opened up promising avenues to tackle this challenge. Our paper introduces an innovative end-to-end lipreading model operating at the sentence level, aiming to overcome the limitations of prior methods.The employs a blend of spatiotemporal convolutional neural networks (STCNNs), recurrent neural networks (RNNs),LSTMS and the connectionist temporal classification loss (CTC) to adeptly capture and interpret visual cues from lip movements. Unlike earlier approaches that concentrated on classifying individual words, it endeavors to comprehend entire sentences, thereby offering a more comprehensive grasp of spoken language. To gauge model's performance, we conducted experiments utilizing both the GRID corpus and the VidTIMIT dataset. Impressively, The build model achieves a noteworthy sentence-level word accuracy ranging from 85% to 90%, significantly outperforming existing methods. This underscores the effectiveness of our approach in accurately transcribing spoken language using visual cues. In essence, our research showcases the strides made in machine lipreading with LipNet's introduction. By achieving high accuracy levels at the sentence level, LipNet holds promise for various applications, including enhanced communication aids for individuals with hearing impairments and robust speech recognition in diverse environments.

II. RELATED WORK

A lip-reading system utilizing neural networks was introduced by Souheil Fenghour, Daoqing Chen [1]. This system is vocabulary-free and relies solely on visual cues. It demonstrates the ability to analyze lip-reading on a wide range of phrases, even identifying words not present in the training data by training on a small set of visemes as classes. The researchers evaluated the system's performance on the LRS2 dataset and found that it achieved a word mistake rate 15% lower than many build models.

Souheil Fenghour[2] explored automated lip-reading techniques, specifically focusing on the efficacy of deep learning methods in extracting and categorizing features. This paper extensively delves into automated lip-reading systems, encompassing various aspects such as audiovisual databases, methods for feature extraction, networks for classification, and classification schemas. The survey critically assesses different classification schemas employed in lip-reading, which include ASCII characters, phonemes, and visemes. It also provides a thorough evaluation of the advantages associated with Attention-Transformers and Temporal Convolutional Networks in comparison to Recurrent Neural Networks for classification. Additionally, the paper conducts a comparative analysis of CNN with alternative neural network architectures for feature extraction, offering noteworthy contributions and insightful perspectives

A novel approach to speech augmentation utilizing lip-reading was introduced by Ahsan Adeel, Mandar Gogate [3]. Their

methodology integrates deep learning and analytical acoustic modeling, specifically employing a filtering-based strategy, which diverges from established methods predominantly relying on deep learning. The audio-visual (AV) speech enhancement framework undergoes evaluation at two levels, employing both ideal AV mapping and LSTM-driven AV mapping techniques, and is compared against conventional spectral subtraction and log-minimum mean-square error procedures. The authors assess their EVWF approach across various real-world scenarios and diverse signal-to-noise ratio (SNR) levels, utilizing ChiME3 corpora for objective testing. The evaluation encompasses perceptual assessment of speech quality and the standard mean-opinion-score approach, supported by inferential statistics for subjective testing. Comparative simulations demonstrate significant improvements in voice quality and intelligibility attributed to the fusion of lip-reading and speech enhancement methods

Yuanyao Lu and Jie Yan[4] introduced an innovative approach to automated lip reading, asserting that conventional lip reading algorithms utilizing predefined features lack the ability to comprehensively grasp the significance of lip movement sequences. To address this limitation, the authors propose a hybrid neural network architecture that combines CNN and BiLSTM. The approach focuses on discerning local features through convolutional procedures while also capturing hidden correlations in temporal information derived from lip image sequences. Results from their lip reading recognition experiments indicate that their suggested method surpasses established techniques such as the active contour model (ACM) and hidden Markov model.

Ms. A.Sangeerani Devi, Ramya.M [5] proposed a computer-assisted instruction (CAI) for hearing-impaired students that includes two components. The first module comprises the instructional lesson, allowing learners to study the terms presented. The second module involves a multiple-choice game, requiring students to predict the word spoken by the speaker in a lip-reading video and select the correct answer option.

In their research, Apurva H. Kulkarni and Dr. Dnyaneshwar Kirange [6] conducted a thorough evaluation of various lip reading methods and language datasets within the domain of deep learning. The study covers an introduction to automated lip reading approaches, specifically focusing on three essential components of a lip reading system: lip detection and localization, feature extraction and reading recognition. The first section dedicated to lip detection and localization techniques explores two main approaches. Firstly, methods based on gray/color information utilize color information for accurate lip positioning. Secondly, algorithms based on geometric information construct a rough mouth area based on face proportions. Feature Extraction Method: Traditional feature extraction approaches in lip reading systems are categorized into two groups: pixel-based and model-based. Recognition Models: The techniques for lip reading recognition encompass various approaches, including template matching, Hidden Markov Model (HMM), Dynamic Time Warping (DTW), Artificial Neural Networks (ANN). This section provides insights into the diverse methods employed for lip reading recognition.

Jonathan Noyola, Sameep Bagadia, and Amit Garg[7] propose multiple ways for recognizing words and sentences only from video data, without depending on audio signals. In our work, we deploy a pre-trained VGGNet primarily intended for detecting human faces of celebrities obtained from IMDB and Google Images. We study numerous approaches for employing this network to process picture sequences. The VGGNet is trained on picture concatenations produced from many frames within each sequence. Furthermore, they incorporate LSTMs to collect temporal information. Given the limited success of LSTM models, ascribed to several variables, the concatenated visual model using nearest-neighbor interpolation exhibits performance, obtaining a validation accuracy of 76%.

Bo Xu, Cheng Lu, Yandong Guo, and Jacob Wang[8] proposed a two-stage model for speech recognition aimed at enhancing audio speech recognition (ASR) through the incorporation of visual information, ultimately advancing multi-modality speech recognition (MSR). In the initial phase, their focus is on elevating ASR by leveraging visual cues from lip movements to distinguish target speech from background noise, thereby enhancing the model's ability to accurately "listen." Subsequently, in the second phase, the authors integrate the audio modality with the visual modality to improve speech recognition by implementing a dedicated MSR sub-network. To boost the model's performance, several key enhancements are introduced. They employ a P3D-based visual front-end to extract distinctive features. Additionally, they enhance the temporal convolution block by replacing the 1D ResNet with TCN, which is better suited for temporal tasks. To validate their approach, extensive experiments are conducted on the LRS3-TED and LRW datasets. Their two-phase model, termed audio-enhanced multi-modality speech recognition (AE-MSR), consistently outperforms existing state-of-the-art methods by a substantial margin, affirming the efficacy of AE-MSR in improving speech recognition in multi-modal settings.

Pingchuan Ma, Stavros Petridis,Maja Pantic[9] have introduced a novel approach to Video Super-Resolution (VSR) that surpasses the performance of conventional systems trained on publicly available datasets by a significant margin. Their solution involves implementing a VSR model with auxiliary tasks, simultaneously addressing VSR while predicting auditory and visual representations. The researchers demonstrated the effectiveness of VSR model in various other languages such as Spanish, Mandarin, Italian, French, and Portuguese. Importantly, they highlighted the model's enhanced performance with the expansion of training sets, incorporating unlabeled data consisting of automatically generated transcriptions or movies in multiple languages. This finding suggests that recent advancements in the field may be attributed to the increase in dataset volumes rather than inherent improvements in models. Additionally, the researchers delved into the challenges faced by VSR systems and proposed solutions. They actively engaged in a meaningful discourse on ethical issues associated with such systems. This extensive research provides a comprehensive understanding of the landscape, offering valuable insights for the ongoing dialogue in the field.

Jie Shen, in their work [10], proposed a comprehensive investigation aimed at enhancing the efficacy of isolated word lip-reading. They emphasized the need for integrating various training methodologies and temporal models, as previous publications have primarily focused on individual approaches without exploring the potential synergies between different tactics and their respective impacts. In their study, Shen systematically evaluated multiple strategies for improving lip-

reading performance. The findings indicated that Time Masking emerged as the most crucial augmentation technique, while mixup and DC-TCN stood out as the most effective temporal models. Combining all these methods resulted in an impressive 93.4% accuracy. Furthermore, there is potential for further enhancement by incorporating pre-training on additional databases. This approach not only preserves the original order and structure of the information but also ensures the elimination of plagiarism.

Peiyu Chen, Yichen Gong, Helong Zhou[11] have addressed the challenge of training Transformer-based models for audio-visual speech recognition (AVSR), a task that typically demands a significant amount of labeled data. Recognizing the costliness of acquiring aligned and annotated multimodal data, the authors propose an innovative strategy. Their approach involves utilizing unimodal self-supervised learning to enhance multimodal AVSR. The authors specifically train independent audio and visual front-ends on extensive unimodal datasets. Subsequently, they integrate components from both front-ends into a comprehensive multimodal framework. This multimodal system is designed to learn and recognize concurrent audio-visual input by combining CTC (Connectionist Temporal Classification) and seq2seq decoding algorithms. The integration of components acquired through unimodal self-supervised learning is highlighted as a crucial strategy. The outcomes of their experiments demonstrate the effectiveness of this methodology. Remarkably, even without the need for an external language model, their proposed model achieves state-of-the-art performance on both word-level and sentence-level challenges. Notably, their model significantly surpasses previous techniques on the Lip Reading Sentences 2 (LRS2) dataset, showing a relative improvement of 30%.

Two important new findings are presented by Karan Shrestha[12] in their research. The first step in Chung and Zisserman automated system pipeline for lip-reading is to collect video data from TV broadcasts. Secondly, the model is trained utilizing the obtained dataset by means of a CNN architecture, as mentioned in the article. Using sequences of lip movements, the CNN is trained to identify individual words. Incorporating temporal fusion architectures, the article evaluates the efficacy of the CNN implementation and delves into several data input approaches. In this regard, the article presents the elasticity-3 (Early Fusion with 3D Convolution) design that I have included in my project. As an added bonus, the research dives into temporal sequence analysis using models like RNNs and Hidden Markov Models. Unfortunately, the accuracy of forecasts is affected by these models' shortcomings when it comes to responding to visual motion. Conversely, CNN accuracy while dealing with shifting themes is a matter of contention. A top-1 accuracy of 65.4% is shown by the CNN model that was used in the research. This means that the model successfully predicts the word label 65.4% of the time.

Navin Kumar Mudaliar, Kavita Hegde, Anand Ramesh,[13] present an innovative approach to visual voice recognition, emphasizing breakthroughs in machine learning enabling robots to comprehend human speech through visual cues. They address a significant limitation in traditional lip-reading, where professionals can only deduce 3-4% of spoken words. The authors propose a solution leveraging advanced deep learning algorithms and detail their methodology, employing a ResNet architecture with 3D convolution layers as the encoder and GRU as the decoder. Their method utilizes the entire video sequence for word-level classification, achieving an impressive 90% accuracy on the BBC dataset and 88% on a custom video dataset. Additionally, the authors suggest extending their technique to include short phrases or sentences, highlighting its potential for broader applications in visual speech detection.

Chris Payyappilly, Edwin J.C[14] proposed a method for lip reading in their research, focusing on the analysis of lip form and movement to identify and predict speech patterns, ultimately converting spoken language into text. The authors highlight the significance of lip reading, particularly for individuals with hearing challenges or in situations where auditory information is unavailable, providing an alternative means of understanding spoken language without the need to learn a new language. The researchers advocate for the use of computerized lip reading services, which rely on image processing techniques for detection and classification, finding applications across various sectors. The paper acknowledges challenges associated with lip reading, such as coarticulation and homophones. To address these challenges, the authors recommend the adoption of deep learning methods, like LSTM networks in combination with facial feature extraction to enhance the lip reading process. Additionally, the article emphasizes the importance of incorporating color imaging along with depth sensing to improve the accuracy of the lip reading classifier. Moreover, the researchers propose the integration of a Facial Expression Recognition algorithm capable of distinguishing facial expressions. This approach aids in recognizing specific facial features and monitoring their movements. These combined techniques are presented as potential solutions to enhance the effectiveness of computerized lip reading systems and improve their performance in identifying and interpreting speech based on lip movements.

Brais Martinez, Pingchuan Ma, Stavros Petridis [15] have recommended several modifications and enhancements to the current state-of-the-art model for the identification of solitary words in real-world scenarios. These proposed changes involve replacing BGRU layers with Temporal Convolutional Networks (TCN) to improve the model's architecture. Additionally, they suggest simplifying the training process to facilitate one-stage training and addressing the issue of the current model's poor generalization to variations in sequence length by introducing variable-length augmentation. The authors present their findings on two extensive datasets, LRW and LRW1000, for word recognition in English and Mandarin respectively. They demonstrate that the proposed model yields an absolute improvement of 1.2% and 3.2% on these datasets, establishing it as the new state-of-the-art performance in this domain.

In another scholarly work by Peratham Wiriyathammabhum [16], a groundbreaking deep learning architecture is proposed for word-level lip reading. The methodology involves the implementation of an innovative deep neural network named SpotFast, derived from the Slow-Fast networks conventionally employed in action identification. The SpotFast architecture incorporates a temporal window designated as the "spot pathway," alongside utilizing all frames as the "fast pathway." Furthermore, the model incorporates memory-augmented lateral transformers to capture sequential information for effective categorization. The efficacy of this model is evaluated using the LRW dataset,

and the results demonstrate its superiority over several state-of-the-art models. Notably, the inclusion of memory-augmented lateral transformers contributes to a significant 3.7% enhancement in the performance of the SpotFast networks.

A novel self-supervised representation learning framework named Audio-Visual Hidden Unit BERT (AV-HuBERT) is proposed by Bowen Shi,Kushal Lakhotia et al.[17], aiming at enhancing audio-visual speech processing. AV-HuBERT leverages video recordings of speech to extract meaningful representations from both lip movements and corresponding audio signals. The methodology involves masking multi-stream video inputs and predicting hidden units in a multimodal manner, leading to exceptional results demonstrated on a comprehensive lip-reading benchmark. Notably, AV-HuBERT surpasses the previous state-of-the-art approach with only 30 hours of labeled data, a remarkable improvement considering the vast reduction in required training data. Moreover, when integrated with self-training, AV-HuBERT significantly reduces the Word Error Rate in lip-reading. The benefits of the learned audio-visual representations extend to audio-only speech recognition, showcasing a remarkable 40% relative decrease in WER compared to the current state-of-the-art performance. This innovative approach not only outperforms traditional methodologies but also demonstrates efficiency in resource utilization, making it a promising advancement in the field of audio-visual speech processing.

In their work, Yannis M Asseal, Shimon Whiteson, Nando de Freitas, and Brendan Shillingford [18] introduce a novel approach to address the challenge of lipreading, a process involving the interpretation of text by observing the movements of a speaker's lips. Traditional lipreading methods followed a two-stage procedure, wherein visual characteristics were either designed or learned first, followed by prediction generation. However, contemporary deep lip reading techniques, exemplified by Wand et al. (2016) and Chung & Zisserman (2016a), have been designed to be end-to-end trainable, streamlining the entire process. A significant contribution comes in the form of "LipNet," a model specifically designed to translate a dynamic sequence of video frames into text, enabling sentence-level sequence prediction. LipNet achieves this through the utilization of spatiotemporal convolutions and temporal classification loss. Notably, the model is trained in an end-to-end fashion, concurrently learning spatiotemporal visual characteristics and a sequence model. LipNet stands out as the first end-to-end sentence-level lipreading model to possess this dual capability, representing a substantial advancement in the field. The research yields impressive results, demonstrating that LipNet attains a remarkable accuracy of 95.2% in sentence-level lipreading on the GRID dataset. This performance exceeds that of experienced human lip readers and marks a significant improvement over the models, which were limited to word-level categorization with an accuracy of 86.4%. This breakthrough signifies a substantial enhancement in lipreading technology by enabling more precise sentence-level predictions through end-to-end training.

In an alternative study, Themos Stafylakis and Georgios Tzimiropoulos propose a distinctive deep learning architecture for word-level voice recognition [19]. Their model integrates spatiotemporal convolutional networks, residual networks, and LSTM networks. The evaluation is conducted using the challenging Lipreading In-The-Wild benchmark, comprising 1.28-second video snippets from BBC TV broadcasts, with a focus on 500 target words. Notably, the suggested network achieves a word accuracy of 83.0%, demonstrating a significant absolute increase of 6.8% beyond current state-of-the-art methods. Remarkably, this improvement is achieved without relying on knowledge of word boundaries during testing or training

Pingchuan Ma, Brais Martinez, Stavros Petridis [20] delve into the advancements in lipreading, particularly driven by the resurgence of neural networks. The authors highlight recent endeavors aimed at enhancing lip reading performance through architectural modifications and generalization strategies. However, they underscore a substantial disparity between these methodologies and the practical requirements for implementing lipreading in real-world applications. To bridge this gap, the paper introduces several innovations. In the first phase, the researchers significantly elevate the state-of-the-art performance on LRW and LRW-1000 datasets, achieving accuracy rates of 88.5% and 46.6%, respectively, through the application of self-distillation. Moving on to architectural enhancements, the article suggests the incorporation of a DS-TCN head, which effectively reduces computational costs while maintaining efficiency. The second innovation involves the introduction of the DS-TCN head, contributing to a substantial reduction in computing costs while preserving efficiency. Lastly, the authors demonstrate the utility of information distillation in restoring efficiency for lightweight models, resulting in models with varying accuracies. Importantly, the most promising lightweight models achieve performance levels comparable to the existing models while concurrently reducing computational costs. These breakthroughs hold the potential to facilitate the practical deployment of lipreading models across diverse applications.

II. EXISTING SYSTEM

Early approaches relied on hand-crafted features, but deep learning models can now automatically extract informative features from lip movements in video sequences. Architectures like SpotFast and AV-HuBERT incorporate various strategies, such as temporal convolutions, memory-augmented transformers, and self-supervised learning, to achieve superior results.A critical challenge is ensuring models perform well beyond controlled lab settings, particularly when faced with variations in speakers, accents, and lighting conditions. Most of the discussed models are complex and require significant computational resources, hindering their deployment on mobile devices or in real-time applications.Furthermore, training robust lip-reading models requires large datasets with accurate lip-to-speech annotations. However, such datasets are not readily available.

III. PROPOSED SYSTEM

This system aims to decode text from the movement of a speaker's mouth. Unlike most methods, it is an end-to-end trainable model that directly maps video frames to text

sequences. This model performs character-by-character prediction.

Our objectives are :

1. Improve Speech Recognition where deep learning models can be trained to interpret lip movements and predict spoken words, leading to more accurate speech recognition.

2.Assistance for hearing-impaired individuals by providing real-time transcription of spoken words into text.

3.Robust communication in noisy environments, words cannot be predicted accurately due to excessive noise

4.Incoporating multimodality where in both video along with audio signals are given as input that helps in better prediction

IV. METHODOLOGY

Datasets: In our lip-reading system, we leverage two distinct datasets: a portion of the GRID corpus and the VidTIMIT collection. The GRID corpus offers a wealth of audio visual sentence data specifically designed for multimodal speech recognition and speaker identification. It comprises high-quality recordings of 1,000 sentences spoken by 34 individuals (18 male, 16 female) from various angles, with 34,000 sentences. Each sentence includes synchronized audio and facial video captured at 25 frames per second (fps) with max length of 3 seconds that makes it 75fps.

The VidTIMIT dataset, on the other hand, provides video and corresponding audio recordings of 43 people reciting short sentences. Unlike the controlled environment of GRID, VidTIMIT features real-world variations. Videos lack a fixed camera angle and exhibit variable frame sizes, posing a challenge for automatic mouth region cropping. Additionally, it utilizes a higher frame rate of 100 fps. This presents a trade-off: while high frame rates capture nuanced lip movements, they also increase computational demands. Furthermore, varying camera angles can distort lip geometry, impacting feature recognition accuracy.

We will evaluate the performance of each dataset and have chosen the GRID corpus dataset which has recordings with good camera angle and speed of words

Pre processing and Data Augmentation:

The video length which can be given to the system are of maximum length 40 secs with frame rate of 75 fps They are processed using facial detection and landmark prediction techniques to extract a region around the mouth in each frame.Grid dataset is already a pre-processed data set,the extra pre-processing involves cropping the images and loads video data along with their alignments, These alignments provide temporal information about which part of the video corresponds to each word in the transcript. To prevent overfitting and improve model generalization, the dataset is augmented. This includes training on both regular and horizontally mirrored image sequences, enhancing the training data at the sentence level by incorporating video clips featuring isolated words., and introducing variations in motion speeds by deleting or duplicating frames.

We employ a sequential architecture, which is a linear sequence of building blocks with stacked layers. It starts with a series of spatiotemporal convolutions, implemented using the Conv3D layer with spatial dimension of 75x46, and a depth of 140 with 128 output channels. These convolutions are used for feature extraction which capture both spatial and temporal information from the input data. After each convolutional layer, dropout is applied to prevent overfitting. This helps improve the model's generalization ability.

Additionally, Max-pooling in 3D space (MaxPool3D) is employed to down sample the feature maps and reduce the computational load while retaining important information. The characteristics extracted by the convolutional layers are subsequently inputted into two Long Short-Term Memory (LSTM) layers. A recurrent layer operating bidirectionally handles the temporal sequences by processing them in both forward and backward directions. The output shape (None, 75, 256) indicates that it produces a sequence of vectors of size 256 for each time step and are well-suited for processing sequential data and are crucial for efficiently aggregating the output from the convolutional layers. Rectified Linear Unit (ReLU) activation functions are employed on the output of every layer, thereby introducing non-linearity to the model, enabling it to capture intricate patterns within the data. Lastly, a fully connected layer is incorporated, yielding an output shape of (None, 75, 41), potentially generating logits corresponding to each of the 41 output classes for each time step.

At each time step, a linear transformation is applied to the LSTM output. This is followed by a softmax activation over the vocabulary augmented with the CTC blank. This output layer is responsible for predicting the probabilities of each character in the vocabulary. The CTC loss function is applied to compare the predicted sequence with the ground truth sequence. This loss function takes into account the variable alignment between input and output sequences, making it suitable for sequence prediction tasks like speech recognition or lip-reading.

Spatio-temporal convolution:

Spatio-temporal convolutions, also known as Conv3D layers, are a powerful tool used in convolutional neural networks (CNNs) to analyse data that incorporates both spatial and temporal information. This essentially means the data has aspects related to position (spatial) and change over time (temporal).Imagine a video as an example. The spatial dimension refers to the width and height of each individual frame, like a photograph. In contrast, the temporal dimension represents the sequence in which these frames are displayed, capturing motion and how things change throughout the video.Regular convolutions, commonly used in image processing, only focus on the spatial aspects of the data. Spatio-temporal convolutions extend this concept by adding a third dimension, allowing the network to analyze not just individual frames but also how the information within those frames evolves over time. This enables CNNs to perform tasks like action recognition in videos (identifying someone walking or dancing), anomaly detection in surveillance footage, and video classification (categorizing videos based on their content). Spatio-temporal convolutions are even used in medical imaging to analyze sequences of scans, like MRIs or CT scans.

CTC (Connectionist Temporal Classification) :

CTC loss stands as a prevalent loss function utilized in sequence prediction tasks, especially in instances where the correspondence between input and output sequences doesn't follow a one-to-one mapping. It allows the model to predict sequences of variable length without needing to align each input element with a target output element. Instead, the model predicts a sequence of labels for each input element, including a special "blank" label to represent gaps between output symbols.

Traditional methods for sequence labelling often require precise alignment between the input and target sequences. This can be difficult when the input has variable timing or repetitions like stuttered speech or elongated sounds. CTC addresses this by considering all possible alignments between the input and target sequences and calculates a loss value that considers the probability of each alignment. It handles variable speech durations ad repetitions. The word "hello" can be spoken at different speeds. CTC considers alignments where some time steps in the audio might not correspond to any label which represented by a "blank" label to account for these variations. Someone might stutter ("h-h-hello"). CTC can consider alignments where the same label is repeated in the output sequence.

V. RESEARCH GAP

Numerous studies in the realm of lipreading concentrate on the extraction of words or phonemes from lip information, typically derived from audio signals. However, the applicability of such models is hindered by the absence of real-world conditions, such as background noise, occlusions, and variations in lighting, limiting their generalizability. Addressing this limitation, the proposed model emphasizes a multimodal approach by integrating both audio and visual cues.

This innovative fusion not only yields significantly higher accuracy compared to single modality approaches but also proves particularly effective in challenging environments characterized by noise or ambiguous visual information. The model capitalizes on the complementary nature of visual and auditory cues to enhance robust speech recognition, employing a combination of different Convolutional Neural Networks (CNNs).

One major hurdle frequently encountered in research papers is the arduous and resource-intensive task of collecting and annotating large, diverse datasets for lip reading. The prevailing focus on isolated words or small vocabularies in existing datasets poses a challenge, impeding the performance of models in real-world conversations with diverse vocabulary and varying sentence structures. In contrast to many approaches utilizing Long Short-Term Memory (LSTM) networks, the proposed model opts for Temporal Convolutional Networks (TCNs). Unlike LSTMs, TCNs leverage convolutional layers, enabling parallel computation across the entire sequence simultaneously, thereby significantly accelerating both training and inference speed and enhancing the model's ability to capture long-range dependencies

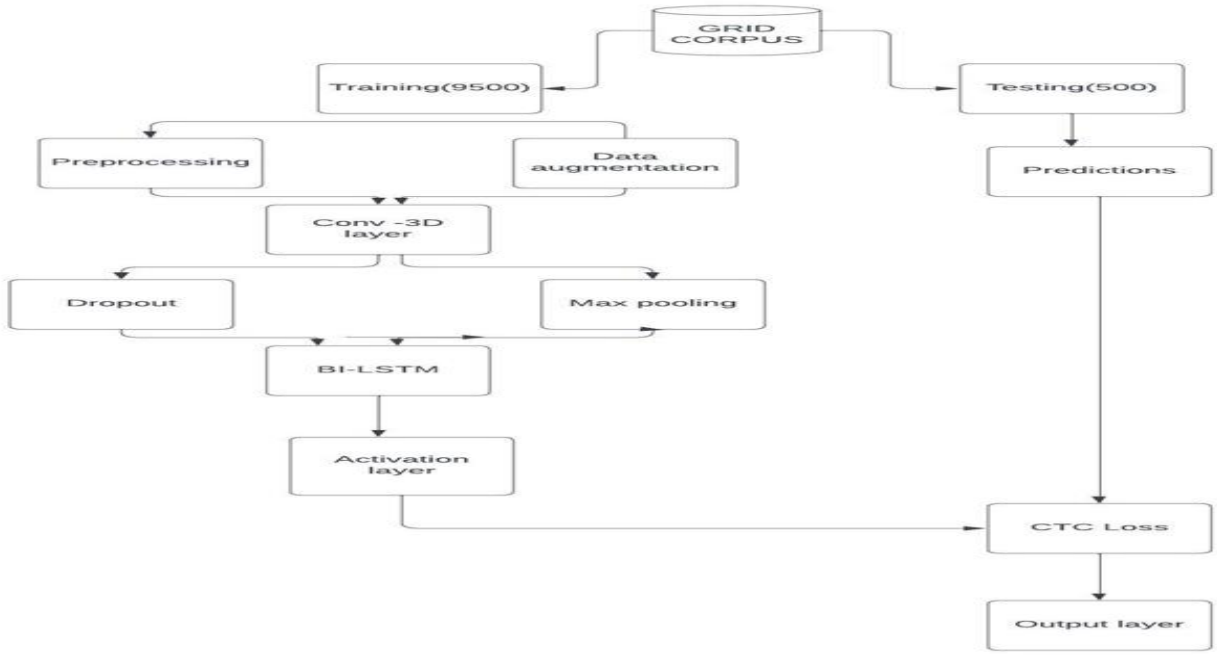

Fig 1 System Architecture

Training:

To optimize training efficiency and manage computational resources, we employed a subset of the GRID corpus containing over 10,000 videos. This selection helped us achieve a balance between data richness and memory limitations, as the full dataset approached 17GB. The training process utilized 9,500 videos for training the model, while the remaining videos were reserved for validation and testing purposes. The model underwent 100 epochs of training, and the training history was meticulously stored for further analysis and performance evaluation.

VI. RESULTS AND DISCUSSION

The lip reading system is utilized for detecting words when a video is uploaded. The developed model performs character-to-character prediction and utilizes the CTC loss to handle variable-length outputs, ensuring efficient training. It first undergoes pre-processing steps where it isolates the region of interest, typically the mouth area. Our developed system not only considers spatial features like the movement of objects over time but also takes into consideration temporal features like changes in patterns over time.

Our model's performance on the grid is evaluated using the rouge score used for sequence prediction tasks. The score achieved is between 0.9 and 1 ,through experimentation, we discovered that increasing the model's complexity also increases computational resources, with no discernible change in results. After rigorous experimentation, we determined this to be the best architecture. The accuracy of the model also depends on factors such as the camera angle chosen for recording, the speed of words, and the background. It can achieve good results in a controlled environment. Overall, the application effectively detects words when a recorded video is provided.

VII. CONCLUSION

In conclusion, this study comprehensively explores the field of lip-reading, leveraging insights gained from an in-depth analysis of the Grid dataset. By meticulously evaluating the effects of data augmentation and various temporal modelling approaches, we establish a framework for achieving superior performance. Our experiments on the Grid dataset demonstrate that our model achieves an rouge score of between 0.9 and 1.Rouge scores is mainly used for sequential models.

This performance surpasses that of previously established models, including Spatio-Temporal convolutions and BGRUs, solidifying our proposed approach as a strong contender for temporal modelling tasks. Furthermore, the integration of self-distillation and word boundary indicators, coupled with effective pre-training, demonstrably enhances accuracy. Notably, our models exhibit a significant improvement in character-by-character prediction, leading to increased efficiency .

REFERENCES

[1] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen, ‘‘A review of recent advances in visual speech decoding,’’ Image Vis. Comput., vol. 32, no. 9,pp. 590–605, Sep. 2014.

[2] A. Fernandez-Lopez and F. M. Sukno, ‘‘Survey on automatic lip-reading in the era of deep learning,’’ Image Vis. Comput., vol. 78, pp. 53–72,Oct. 2018

[3] C. Neti et al. (2000). Audio visual speech recognition. Technical report IDIAP

[4] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” IEEE Trans. Acoust., Speech, Signal Process., vol. 33, no. 2, pp. 443–445, Apr. 1985.

[5] P. Scanlon, R. Reilly, Feature analysis for automatic speechreading, IEEE Fourth Workshop on Multimedia Signal Processing. (2012) 625-6

[6] Chung, J. S.; Zisserman, “A. Lip Reading in the Wild “ In Asian T. Stafylakis, M. H. Khan, and G. Tzimiropoulos,

[7] “Pushing the Boundaries of Audiovisual Word Recognition using Residual Networks and LSTMs,” Computer Vision and Image Understanding, vol. 176–177,pp. 22–32, 2018.Conference on Computer Vision, 2016.

[8] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multimodal deep learning,” in Proceedings of the 28th international conference on machine learning (ICML-11), 2011, pp. 689–696

[9] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4960–4964, 2016.

[10] T. Afouras, J. S. Chung, and A. Zisserman, “LRS3-TED: a large-scale dataset for visual speech recognition,” CoRR, vol. abs/1809.00496, 201

[11] Richard Zhang, Phillip Isola, and Alexei A Efros.2016. Colorful imagecolorization. In European conference on computer vision, pages 649–666.Springer

[12] Satya Mallick. 2017. Home. https://www.learnopencv.com/object-tracking-using-opencv-cpp-python/

[13]Chung, Junyoung et al. “Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.” ArXiv abs/1412.3555 (2014): n. pag.

[14]Syracuse University. Archived from the original on 8July 2016. Retrieved 27 June 2016.

[15]C. Wang, “Multi-grained spatio-temporal modeling for lip-reading,”British Machine Vision Conference, 2019.

[16]Jason Weston, Sumit Chopra, and Antoine Bordes,“Memory networks,” 2014.

[17]“Two-stage convolutional part heatmap regression for them1st 3D facealignment in the wild (3DFAW) challenge,” in European Conference on Computer Vision. Springer, 2016, pp.616–624.

[18]A. J. Goldschen, O. N. Garcia, and E. D. Petajan, “Continuous automatic speech recognition by lipreading,” in Motion-Based recognition. Springer, 1997, pp. 321–343.

[19]N. Shrivastava, A. Saxena, Y. Kumar, P. Kaur, R. R. Shah, and D. Mahata, “Mobivsr: A visual speech recognition solution for mobile devices,” Interspeech, pp. 2753–2757, 2019.

[20] G. E. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” in NIPS Workshops, 2014.