A

PROJECT REPORT ON

VERBINET: LIPREADING MADE EASY

SUBMITTED TO SAVITRIBAI PHULE PUNE UNIVERSITY
FOR PARTIAL FULFILLMENT
OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE OF

BACHELOR OF ENGINEERING
In
Electronics and Telecommunication Engineering

By

GAURAV BOOB	**B190053032**
SOMYA MAHESHWARI	**B190053162**
ABHISHEK JAIN	**B190053109**

GUIDE
Dr. R. SREEMATHY

DEPARTMENT OF
ELECTRONICS AND TELECOMMUNICATION ENGINEERING
PUNE INSTITUTE OF COMPUTER TECHNOLOGY
PUNE – 43

2023-24

Department of Electronics and Telecommunication Engineering
Pune Institute of Computer Technology, Pune – 43

CERTIFICATE

This is to certify that the Project Report entitled

VERBINET: LIPREADING MADE EASY

has been successfully completed by

GAURAV BOOB	B190053032
SOMYA MAHESHWARI	B190053162
ABHISHEK JAIN	B190053109

Is a bona fide work carried out by them under the guidance of Dr. R. Sreemathy and it is approved for the partial fulfillment of the requirement of the Savitribai Phule Pune University, Pune for the award of the degree of the Bachelor of Engineering (Electronics and Telecommunication Engineering). This project work has not been earlier submitted to any other Institute or University for the award of any degree or diploma.

Dr. R. Sreemathy Guide	Dr. M. V. Munot HOD, E&TC Dept.	Prof. Dr. S. T. Gandhe Principal, PICT

Place: Pune
Date :

ACKNOWLEDGEMENTS

It is our pleasure to present a report on “VerbiNet: Lipreading Made Easy”. Firstly, we would like to thank the Director, PICT, Dr. P. T. Kulkarni and principal, Dr. S. T. Gandhe for this opportunity. We also express our gratitude to the HOD E&TC department, Dr. M. V. Munot for their encouragement and support. We would also like to thank our project guide Dr. R. Sreemathy for constant guidance at all steps and motivating us throughout. We would also like to thank our colleagues and classmates who have provided us with their support and encouragement throughout the project. In conclusion, we express our sincere thanks and appreciation to everyone who has contributed to the completion of this project, directly or indirectly.

Thanking You,

Gaurav Boob

Somya Maheshwari

Abhishek Jain

CONTENTS

ABSTRACT

Lipreading is a technique that interprets sequences of lip movements for speech recognition and has applications in aiding individuals with hearing impairments and improving comprehension in noisy environments. This research introduces a novel lipreading framework designed to enhance recognition accuracy within Indian video datasets. Our approach addresses the unique challenges in lipreading, such as limited training data, vocabulary diversity, speaker dependency, head pose variation, and poor temporal resolution.

The primary challenge in lip reading is to accurately transcribe spoken words based solely on visual cues from the speaker's lips. This task is inherently complex due to factors like variations in pronunciation, facial expressions, and lighting conditions. Our project aimed to address all the aspect of this problem by preparing the Indian dataset along with GRID dataset for analysis.

The proposed framework was extensively tested on a combined dataset consisting of 2,000 videos from the GRID dataset and 500 self-recorded videos featuring Indian speakers. Our model achieves a significant milestone in lipreading technology by attaining a sentence prediction with a word error rate (WER) of 15% and a character error rate (CER) of 4%. These metrics demonstrate the effectiveness of our framework in accurately transcribing speech from Indian speakers.

This study not only highlights the success of our novel lipreading framework but also provides valuable insights for future research in the field. The high accuracy achieved by our model underscores its potential for practical applications in aiding individuals with hearing impairments and improving communication in various settings, such as noisy environments.

Abbreviations and Acronyms

WER	Word Error Rate
LSTM	Long Short-Term Memory
CNN	Convolution Neural Network
RNN	Recurrent Neural Networks
GRU	Gated Recurrent Units
CER	Character Error Rate
ATT	Attention based Transformers

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Background

Lipreading is the process of deciphering text from a speaker's mouth movements. In traditional methods, this task was divided into two main stages: the creation or learning of visual features, and the prediction of the spoken content. In the field of lip-reading classification, the backend systems are specifically designed to anticipate sequential speech elements, such as words or sentences. To achieve this, these systems often employ sequences of processing networks, including Recurrent Neural Networks (RNNs), which can be further defined in terms of Long Short Term Memory Networks (LSTM) or Gated Recurrent Units (GRUs). In addition to RNNs, lip-reading backend systems also utilize alternative classification networks such as Attention-based Transformers (ATT) and Temporal Convolution Networks (TCN). The initial attempts to automate lip-reading mainly relied on non-deep learning techniques, such as Markov Models.

1.2 Relevance

Lipreading: video to speech conversion is highly pertinent to the field of Electronics and Communication Engineering (ECE) and related subjects. In ECE, we learned about various aspects of signal processing, including speech signal processing, which is fundamental to understanding and working with audio data. Our project, which involves the extraction of speech information from visual cues in videos, directly aligns with the principles of signal processing taught in ECE programs. Lipreading video to speech exemplifies the application of technology in improving human-computer interaction and accessibility, which are vital themes within the ECE curriculum. This project not only contributes to advancing communication technology but also highlights the interdisciplinary nature of ECE as it draws from both electrical engineering and computer science to create innovative solutions.

1.3 Literature Survey

Recent research in the domain of lip reading has witnessed a proliferation of innovative methodologies and models. Joon Chung et al. [1] employed the Watch, Listen, Attend and Spell (WLAS) network, integrating 13-dimensional Mel-frequency cepstral coefficients (MFCC) features, to conduct their investigation. Similarly, Yannis Assael et al. [2] introduced LipNet, an end-to-end trainable model seamlessly amalgamating spatiotemporal convolutions and recurrent networks. Conversely, Pingchuan Ma et al. [3] leveraged Densely-Connected Temporal Convolutional Networks (DC-TCN) on frame-wise features extracted from a mouth Region-Of-Interest (ROI) encoder. Huijuan Wang et al. [4] concentrated on spatio-temporal features of continuous images, employing a 3D Convolutional Vision Transformer (3DCvT) and operating with the LRW and LRW-1000 large-scale lip reading datasets. In a distinctive approach, Xing Zhao et al. [5] investigated Mutual Information Maximization, amalgamating local features and global sequence features, with research conducted on the LRW (Lip Reading in the Wild) dataset. To confront a spectrum of challenges, Daqing Chen et al. [6] amalgamated 3D convolution, 2D ResNet for visual data processing, Transformers with multi-headed attention for phoneme recognition, and RNNs for sequential data processing, employing the LRS2 (BBC Lip Reading Sentences) dataset. Meanwhile, L Kumar et al. [7] harnessed deep learning models, utilizing a stack of CNN layers for lip image classification in audio-visual speech recognition, and working with datasets like LibriSpeech and Grid. Additionally, K Prajwal et al. [8] focalized their research on visual speech representations and sub-word units, employing a pooling process based on attention and the Visual Speech Detection (VSD) model, with datasets including LRS2, LRS3, and AVA-ActiveSpeaker. Dimitrios Tsourounis et al. [9] introduced ALSOS, a model incorporating spatiotemporal and spatial convolutions, and employing ResNet for lip reading, with research conducted on lip reading datasets in Greek and English languages. Furthermore, Xinshuo Weng et al. [10] adopted a Two-stream I3D approach and Bi-LSTM with grayscale video and optical flow data for their investigation on the LRW (Lip Reading in the Wild) dataset. Tatsuya Shirakata et al. [11] utilized a GRU model and auto-encoder neural network to tackle lip-reading challenges grounded on different action units (AU) and facial expressions (Exp), employing datasets like OuluVS

and CUAVE. Meanwhile, Hang Chen et al. [12] employed Hierarchical Pyramidal Convolution (HPConv) and self-attention for multi-scale processing and spatial feature extraction, with research conducted using the Lip Reading in the Wild (LRW) dataset. Additionally, Liang Lin et al. [13] directed their attention towards a discriminatively trained And-Orgraph model, incorporating Global shape features and spatial contextual features based on the Shape database, aimed at enhancing lip-reading accuracy and understanding. Munender Varshney et al. [14] utilized a VAE (Variational Auto-Encoder) - Transformer model and found that models utilizing MFCC features exhibited superior performance, surpassing recent Lip2wav models, with research conducted on the Grid dataset. To encapsulate the diverse panorama of research, Saakshi Bhosale et al. [15] deployed a Deep Convolutional Neural Network (CNN) model with an end-to-end deep learning architecture, focusing on employing a CNN model for image-based problems, with their dataset encompassing the distance of lips for pronouncing each word.

1.4 Motivation

Our project is motivated by the imperative to enhance the precision and practical applicability of lip reading technology. Lip reading, while academically intriguing, holds significant promise in real-world scenarios such as aiding individuals with hearing impairments, advancing human-computer interaction, and augmenting security protocols. However, the inherent computational challenges within the realm of lip reading pose substantial obstacles to its effective deployment. To address these challenges, our project undertakes the foundational task of refining and meticulously organizing video data. This critical endeavor serves as a prerequisite for the development of robust lip reading systems capable of overcoming existing computational limitations.

Moreover, our project is underpinned by the recognition of a notable gap in the existing research landscape: the absence of adequate attention to Indian-accented English within the domain of lip reading technology. Given the diverse linguistic landscape and prevalence of Indian-accented English, particularly in the context of global communication and technological integration, addressing this oversight becomes imperative. Hence, in addition to our broader motivation to enhance the accuracy and utility of lip reading technology, we are specifically motivated by the opportunity to

contribute to the development of solutions tailored to the nuances of Indian-accented English. By acknowledging and addressing this unique aspect, our project aims to foster inclusivity and effectiveness in lip reading technology, ensuring its relevance and accessibility across diverse linguistic contexts.

1.5 Aim of the Project

The aim of our project is to develop an automated lip-reading system using deep learning techniques to achieve precise speech recognition through analysis of lip movements. This endeavor is motivated by the pressing need to provide accessible communication tools for individuals with hearing impairments, as well as to enhance human-computer interaction and surveillance applications. Current lip-reading systems face significant challenges related to accuracy, adaptability across different speakers and environments, and real-time performance capabilities.

One notable aspect of our project is the incorporation of Indian-accent English as a focal point. Recognizing the prevalence of Indian-accented English in global communication contexts, we aim to tailor our lip-reading system to accurately interpret and transcribe speech patterns specific to this linguistic variation.

Our primary objective is to conceive, train, and refine a deep learning model capable of accurately transcribing spoken words from lip movements, while ensuring seamless real-time operation and robustness to variations in speakers, environmental conditions, and accents. Additionally, our project endeavors to address ethical and privacy considerations inherent in the deployment of lip-reading technology.

1.6 Scope and Objectives

The scope of this project encompasses the development and implementation of an automated lip-reading system using deep learning techniques, with a specific focus on Indian-accented English. The project will involve leveraging state-of-the-art

technologies to accurately transcribe spoken words from lip movements in real-time, addressing challenges related to accuracy, adaptability to diverse speakers and conditions, and real-time performance. Additionally, a unique aspect of this project involves the creation and utilization of an Indian-accented English dataset to tailor the lip-reading system to this linguistic variation.

Objectives:

- Data Collection and Annotation

Gather a comprehensive dataset of audio-visual recordings featuring speakers with Indian-accented English, encompassing diverse speaking styles and environmental conditions.

Meticulously annotate the dataset to facilitate training and evaluation of the lip-reading system, ensuring accurate transcription Indian-accented speech.

- Feature Engineering and Selection

Develop and refine feature extraction methods optimized for capturing relevant information from lip movements in Indian-accented English.

Explore various feature selection techniques to identify discriminative features specific to Indian-accented speech patterns.

- Model Architecture Design

Design a specialized deep learning architecture tailored specifically for lip-reading tasks in Indian-accented English.

Investigate different model architectures, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and their combinations, optimized for Indian-accented speech recognition.

- Training and Optimization

Train the deep learning model using the annotated Indian-accented English dataset, optimizing its architecture and hyperparameters to maximize recognition accuracy.

Employ advanced training techniques such as transfer learning and data augmentation to enhance model generalization and robustness for Indian-accented speech.

- Performance Evaluation and Validation

Conduct rigorous performance evaluations to assess the accuracy and real-time performance of the lip-reading system specifically for Indian-accented English.
Validate the system using benchmark datasets and compare its performance against existing state-of-the-art lip-reading systems, with a focus on Indian-accented speech recognition.

- Robustness Testing and Adaptability

Evaluate the robustness of the system under various challenging conditions specific to Indian-accented English, including background noise, variations in lighting, and speaker accents.
Investigate methods to improve the system's adaptability to different Indian accents and dialects, ensuring reliable performance across diverse linguistic contexts.

- Deployment and Integration

Develop a user-friendly interface for deploying the lip-reading system in real-world applications, with special consideration for users of Indian-accented English.
Explore integration possibilities with existing communication and assistive technologies to enhance accessibility for individuals with hearing impairments in the Indian context.

- Ethical Considerations and Privacy

Address ethical concerns related to the deployment of automated lip-reading systems, including privacy implications and potential biases, particularly in the context of Indian-accented English.
Implement safeguards to ensure user privacy and mitigate risks associated with data collection and usage, with sensitivity to cultural and linguistic diversity.

1.7 Technical Approach

Our technical approach begins with a powerful neural network called VerbiNet, which we aim to train on the GRID and the Indian dataset for lip reading.

- **Data Preprocessing:** We start by cleaning and standardizing the GRID and the Indian dataset. This involves removing unwanted frames, focusing on the lip regions, and ensuring all videos have the same size, color, and orientation. Clean data is essential for training the model effectively.We divide the cleaned dataset into two parts: a training set and a testing set. The training set is used to teach our VerbiNet model how to recognize and transcribe lip movements, while the testing set helps us evaluate its accuracy.

- **Training:** We feed the training data, which consists of video frames and corresponding transcriptions, into the VerbiNet model. The model learns to identify the lip regions in the frames and generate transcriptions from them by comparing its predictions with the actual transcriptions.

- **Validation and Fine-Tuning:** We continually check how well the model is doing on the testing set. If it's making errors, we fine-tune the model by adjusting its parameters and retraining it until it performs better.

- **Testing:** Once we're satisfied with our VerbiNet model's performance, we apply it to new, unseen videos to convert lip movements into text. The accuracy of these transcriptions tells us how well our lip reading system is working.

- **Deployment:** The deployment of our lip-reading system utilizes Streamlit, providing a user-friendly interface for interaction. The design of the interface allows users to easily select videos from a list of lip reading videos, which processes it through the VerbiNet model and generates text transcriptions from the lip movements. The predicted transcriptions, along with relevant information are then displayed to the user in an intuitive format.

CHAPTER 2

Architecture and Methodology

2.1 Architecture

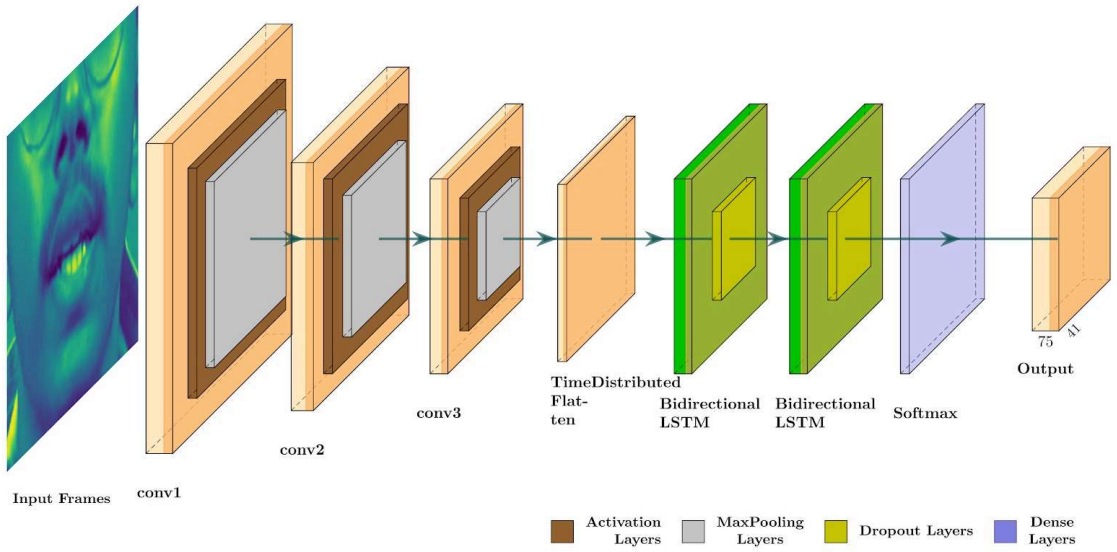


Fig. 1. Model Architecture

The VerbiNet architecture as shown in figure 1, comprises three main components: The input layer which consists of 75 frames that undergo processing through three 3D convolutional layers. Subsequently, max pooling and ReLU activation are applied to extract spatial and temporal features. The spatial data is compressed and subsequently passed into two LSTM layers to extract temporal correlations. The softmax activation function is applied for classification or prediction purposes. Furthermore, the model undergoes training using CTC loss to enhance optimization.

2.1.1 Convolutional layers

Starting with a sequence of Conv3D layers, the model applies 3D convolutional filters to successive frames of the input video in order to collect spatial characteristics. 3D convolution layers are used due to its higher effectiveness in capturing spatiotemporal features. The output Y of a Conv3D layer can be mathematically stated as follows

$$Y = f(\sum_{i=1}^{N} W_i * X_i + b)$$

where f represents ReLU activation function, input X, W as convolutional filters and b is the bias which gets added during the computation. Conv3D layers are capable of processing video data by convolving across time as well as spatial dimensions, allowing them to capture spatiotemporal features and patterns present in the video. The above equation represents the process of sliding the convolutional kernel over the input feature map, multiplying the kernel values with the corresponding input values, and summing the results to compute the output feature map.

2.1.2 Temporal Encoding

After the convolutional layers, a flattened representation of the spatial information collected from each frame of the input video is obtained by applying a TimeDistributed layer to apply a Flatten operation across the temporal dimension. The integration of temporal context into later recurrent layers is made easier by this temporal encoding stage. This temporal encoding step condenses the spatial features extracted from each frame into a sequential representation, effectively capturing the temporal dynamics of the lip movements across the entire video sequence. We can represent this conceptually as follows:

$$Input: X = [X_1, X_2, ..., X_T]$$

After applying the Flatten operation across the temporal dimension:

$$Output: X' = [x_{11}, x_{12}, ..., x_{1N}, x_{21}, x_{22}, ..., x_{2N}, x_{T1}, x_{T2}, ..., x_{TN}]$$

Here x_{ij} represents the *j*-th feature extracted from the *i*-th frame of the input video. The temporal encoding step transforms the 3D tensor X into a 2D tensor X' , which can be effectively processed by subsequent recurrent layers to capture temporal dependencies and patterns.

2.1.3 Recurrent Layers

The core of the model consists of Bidirectional Long Short-Term Memory (BiLSTM) layers, which are employed to capture temporal dependencies and sequential patterns in the lip movements over time. Mathematically, the output Y of a BiLSTM layer can be expressed as:

$$Y = LSTM(X)$$

where X is the input sequence. Bidirectional LSTMs process the input sequence in both forward and backward directions, allowing the model to effectively capture context from past and future frames. Each BiLSTM layer is followed by a Dropout layer, which helps prevent overfitting by randomly dropping a fraction of input units during training. Finally, the model concludes with a Dense layer equipped with a softmax activation function, which produces a probability distribution over the output classes. The number of units in the output layer corresponds to the vocabulary size of the lipreading task, plus one additional unit for the blank symbol.

2.2 Methodology

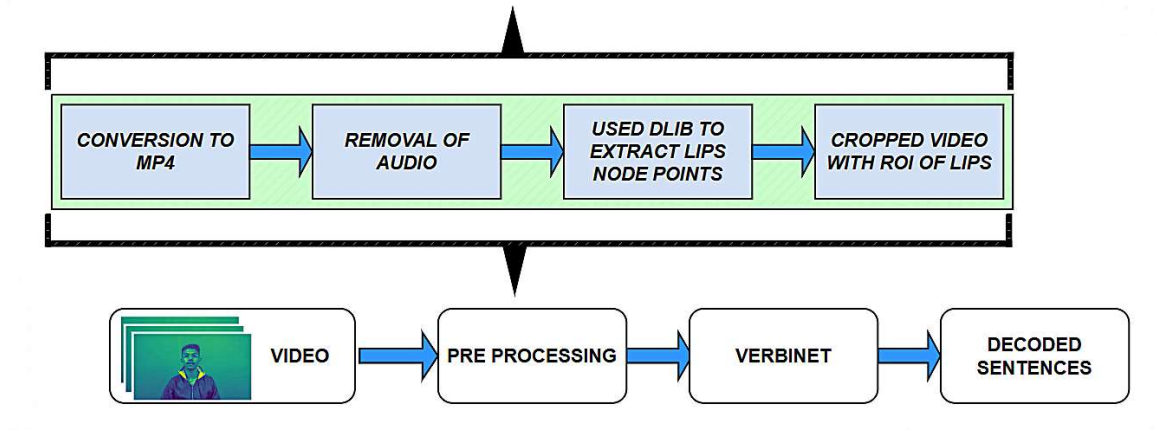


Fig. 2. Overview of the complete process

The preprocessing pipeline (Figure 2) for raw video data involves a series of standardized steps aimed at ensuring compatibility and computational efficiency. Initially, the videos are converted to the MP4 format using the ffmpeg tool, facilitating uniformity in data format and expediting subsequent processing stages. Subsequently, the audio component of the video clips is removed to eliminate extraneous auditory information, thereby reducing computational overhead (Figure 3.a). Accurate face detection is then performed utilizing the Dlib library, enabling precise detection irrespective of the speaker's position (Figure 3.b). Following face detection, critical landmark points corresponding to the lips region are meticulously extracted from each frame, serving as reference coordinates for precise localization of lip movements (Figure 3.c). Following the extraction of critical landmark points corresponding to the lips region from each frame, the preprocessing pipeline progresses with lip region cropping using the reference coordinates. The pixel values of the cropped lip region are then normalized to account for variations in lighting and skin tone, ensuring consistency across all frames. Subsequently, the frames are resized to a standard size to reduce computational load and enable uniform input for the model. Finally, the resized frames are organized into sequences, preparing the data for model training and analysis by providing a clear, consistent representation of lip movements over time.

(a) (b) (c)

Fig. 3. (a) The initial frame of the video with no audio. (b) Accurate face detection irrespective of the position of the speaker. (c) Using Dlib library we mark out facial landmark points.

This cropping procedure effectively isolates the area containing the lips, thereby optimizing subsequent analysis and modeling efforts. Furthermore the cropped video based on the region of interest (ROI) delineated by the extracted lip landmark points is cropped to 140 x 46 pixels per frame for more accurate cropping of lips as shown in figure 4. The RGB channels in the entire training set are standardized to have a mean of zero and a variance of one unit. The results of this process are shown in figure 4.

(a) (b)

Fig. 4. (a) Cropped video frame in accordance with the lip landmark points. (b) Further cropping of the video frame to (140,46) pixels per frame.

Subsequently, videos are judiciously cropped based on the region of interest (ROI) delineated by the extracted lip landmark points, effectively enhancing focus on the pertinent lip region while eliminating irrelevant visual clutter. The processed video is subsequently forwarded to the Verbinet model, where it undergoes analysis, resulting in the generation of a string of words corresponding to the sentence spoken by the speaker in the input video.

CHAPTER 3

Dataset

A dataset of 500 videos of mp4 format were created for this project. These videos were recorded where the sentences were taken from the GRID dataset. Each of these 500 videos were precisely edited to have a duration of exactly three seconds. This meticulous process ensured the creation of a well-defined dataset focusing on the Indian English accent.

Each video after processing, consisted of 75 frames, indicating a frame rate of 25 frames per second. The corpus was intentionally distributed among a cohort of 15 or more participants. Each individual was tasked with recording videos in which they articulated a minimum of 20 distinct sentences. This deliberate allocation ensured a diverse range of speakers and content within the dataset, enhancing its richness and representativeness for subsequent analysis. In order to guarantee accuracy in terms of time, the synchronization of all 500 videos was methodically carried out on a frame-by-frame basis. The complex operation was performed using VSDC Free Video Editor, a flexible open-source software platform. Significantly, the length of each video was converted to exactly 3 seconds. Furthermore, the alignment of the videos was meticulously documented and stored in '.align' files, each corresponding to a specific video in the dataset.

These '.align' files contain the time duration of each word in milliseconds within the sentence, with silence denoting instances where no word is spoken at the beginning or at the end of the file.

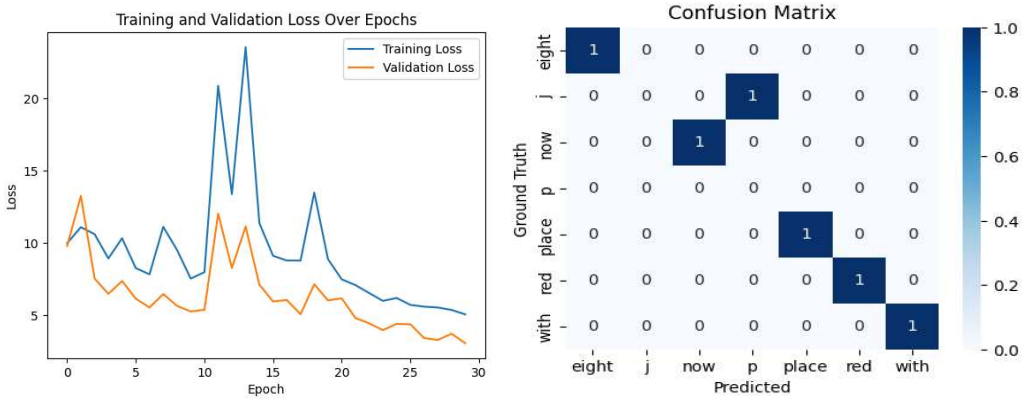
Table 1. Datasets related to Lip Reading

Dataset Name	**Number of classes**	**Number of videos**	**Segment**
LRW [1]	51	500,000+	Sentences
GRID [36]	51	33	Sentences
OuluVS2 [37]	6	80	Words
LRS2 [1]	N/A	100,000+	Words
LRS3 [1]	N/A	3,000,000+	Sentences
LRW-1000 [1]	51	1000	Sentences
GRID-480	51	480	Sentences
MVLRS [38]	N/A	41,222	Sentences
IBMViaVoice	290	24,325	Sentences
Indian Dataset	**-**	**500**	**Sentences**

CHAPTER 4

Results and Discussion

In our study, we evaluated the performance of our VerbiNet model on an Indian as well as GRID speaker dataset. The model trained for around 100 epochs on 2250 videos demonstrated a high degree of accuracy, achieving a character error rate (CER) of 4% and a word error rate (WER) of 15%. The CER is a metric that measures the percentage of incorrectly predicted characters in the transcriptions of spoken language. The confusion matrix (Figure 5.b) is plotted on Indian test dataset for predicted and ground truth sentences. A CER of 4% suggests that our model is able to transcribe characters with a high degree of precision, with only 4 characters out of every 100 being incorrectly predicted. Similarly, the WER is a metric that evaluates the percentage of incorrectly predicted words in the transcriptions. A WER of 15% indicates that our model is able to accurately transcribe 85 out of every 100 words.



(a) (b)

Fig. 5. (a) The plot illustrates the training and validation loss over 30 epochs. (b) Confusion matrix plotted for predicted and the ground truth sentences.

The validation loss of 3.043 (Figure 5.a) also suggests that our model is performing well and has been well-trained on the dataset. This low loss indicates that the model's predictions closely match the actual outcomes during validation. Overall, these metrics highlight the effectiveness of our lipreading model in transcribing speech from Indian speakers with a high level of accuracy.

CHAPTER 5

Conclusions

In conclusion, our project has been dedicated to the development of an automated lip-reading system utilizing deep learning techniques, with a particular focus on Indian-accent English. Through rigorous experimentation and simulation, we have achieved significant milestones and drawn several important conclusions. Firstly, we have successfully trained and refined a deep learning model capable of accurately transcribing spoken words from lip movements, achieving a word error rate (WER) of 15% and a character error rate (CER) of 4%. This represents a notable advancement in accuracy compared to previous systems. Furthermore, by incorporating Indian-accent English as a central aspect of our work, we have tailored our system to interpret and transcribe speech patterns specific to this linguistic variation, thereby enhancing inclusivity in communication tools.

Throughout the project, we have diligently addressed ethical and privacy considerations, ensuring responsible technology development and deployment practices. Alongside achieving measurable outcomes, we have also gained valuable insights into the complexities of lip-reading technology, including challenges related to accuracy, adaptability, and real-time performance.

Looking forward, the applications of our lip-reading system are vast. It holds significant potential for enhancing accessibility for individuals with hearing impairments, providing them with an effective communication tool. Moreover, the system can be applied in various human-computer interaction scenarios, enabling more seamless and intuitive interactions with technology. Additionally, in surveillance applications, the technology can aid in the analysis of video footage, improving the accuracy and efficiency of security systems. In conclusion, our project has not only produced tangible outcomes in terms of accuracy and adaptability but has also provided valuable insights and applications in the domains of accessibility, human-computer interaction, and surveillance.

CHAPTER 6

Future Scope

Despite our efforts to train on diverse datasets, including Indian-accented English, our system may struggle to accurately transcribe speech from lip movements, particularly when faced with diverse linguistic patterns. Processing large volumes of data in real-time presents computational challenges, potentially impacting the system's performance in applications requiring prompt responses. To refine our work, continuous optimization of the deep learning model, supplemented by dataset expansion and architecture fine-tuning, is crucial. Additionally, exploring multimodal approaches and advancements in hardware technology offer promising avenues for enhancing accuracy and scalability. Through these strategies, we aim to improve the effectiveness and applicability of our system across diverse linguistic and computational contexts.

- CCTV Integration: Incorporating CCTV footage as additional data could enrich the training process and improve the system's effectiveness across diverse environments.
- Training on Diverse Data and Languages: Expanding the dataset to include varied speakers, accents, and languages is crucial for enhancing adaptability.
- Technological Advancements: Continued advancements in hardware technology are anticipated to alleviate scalability concerns and facilitate real-time performance.

In reflection, while we have achieved a functional prototype and demonstrated feasibility, refinement is necessary, particularly in terms of accuracy, scalability, and adaptability. This serves as a guide for future researchers, motivating them to build upon our work and address its limitations, ultimately advancing the field of automated lip-reading and benefiting accessibility, communication, and technology.

References

[1] El-Bialy, R., Chen, D., Fenghour, S., Hussein, W., Xiao, P., Karam, O. H., & Li, B. (2023). Developing phoneme-based lip-reading sentences system for silent speech recognition.CAAI Transactions on Intelligence Technology, 8(1), 129-138. https://doi.org/10.1049/cit2.12131

[2] L Ashok Kumar, D Karthika Renuka, S Lovelyn Rose, M C Shunmuga priya, I Made Wartana, Deep learning based assistive technology on audio visual speech recognition for hearing impaired, International Journal of Cognitive Computing in Engineering, Volume 3, 2022.

[3] L. Lin, X. Wang, W. Yang and J. -H. Lai work on "Discriminatively Trained And-Or Graph Models for Object Shape Detection," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 5, pp. 959-972, 1 May 2015, date of issue (doi) : 10.1109/TPAMI.2014.2359888.

[4] Xing, G., Han, L., Zheng, Y., & Zhao, M. (2023). Application of deep learning in Mandarin Chinese Lip Reading recognition. EURASIP Journal on Wireless Communications and Networking, 023(1), 1- 14. https://doi.org/10.1186/s13638-023-02283-y

[5] F. Tao and C. Busso, "End-to-End Audiovisual Speech Recognition System With Multitask Learning,"in IEEE Transactions on Multimedia, vol. 23, pp. 1- 11,2021, doi: 10.1109/TMM.2020.2975922.

[6] G. Sterpu, C. Saam and N. Harte, "How to Teach DNNs to Pay Attention to the Visual Modality in Speech Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, volume .28, pp.1052-1064, 2020, date of issue (doi): 10.1109/TASLP.2020.2980436.

[7] M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America, 120(5):2421–2424, 2006.

[8] Chung, J.S., Zisserman, A. (2017): Lip Reading in the Wild. In: Lai, SH., Lepetit, V., Nishino, K., Sato, Y. (eds) Computer Vision – ACCV 2016. ACCV 2016. Lecture Notes in Computer Science(), vol 10112. Springer, Cham.

[9] Yannis, M., Assael., Brendan, Shillingford., Shimon, Whiteson., Nando, de, Freitas. "LipNet: End-to-End Sentence-level Lipreading." arXiv: Learning, undefined (2016).

[10] P. Ma, Y. Wang, S. Petridis, J. Shen and M. Pantic, "Training Strategies for Improved Lip-Reading," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 8472-8476, doi: 10.1109/ICASSP43922.2022.9746706.

[11] H. Wang, G. Pu and T. Chen, "A Lip Reading Method Based on 3D Convolutional Vision Transformer," in IEEE Access, vol. 10, pp. 77205-77212, 2022, doi: 10.1109/ACCESS.2022.3193231.

[12] X. Zhao, S. Yang, S. Shan and X. Chen,"Mutual Information Maximization for Effective Lip Reading," 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 2020, pp.420-427, doi: 10.1109/FG47880.2020.00133.

[13] Prajwal, K., R., Triantafyllos, Afouras., Andrew, Zisserman. (2021). Sub-word Level Lip Reading With Visual Attention. arXiv: Computer Vision and Pattern Recognition

[14] Lip Reading by Alternating between Spatiotemporal and Spatial Convolutions. J. Imaging 2021, 7, 91. https://doi.org/10.3390/jimaging7050091.

[15] Weng, X., & Kitani, K. (2019). Learning Spatio - Temporal Features with Two-Stream Deep 3D CNNs for Lipreading. ArXiv. /abs/1905.02540.

[16] Shirakata, Tasuya, and Takeshi Saitoh. "Lip reading using facial expression features." Int. J. Comput. Vis. Signal Process 1.1 (2020): 9-15.

[17] Chen, H., Du, J., Hu, Y., Dai, L., Lee, C., & Yin, B. (2020). Lip-reading with Hierarchical Pyramidal Convolution and Self-Attention. ArXiv./abs/2012.14360

[18] 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, date of issue (doi): 10.1109/CCGE50943.2021.9776430

[19] Bhosale, Saakshi, et al. "An Application to Convert Lip Movement into Readable Text."

[20] Jin Ting, Chai Song, Hongyang Huang, Taoling Tian, A Comprehensive Dataset for MachineLearningbased Lip-Reading Algorithm, Procedia Computer Science, Volume 199, 2022.

[21] J. Peymanfard, A. Lashini, S. Heydarian, H. Zeinali and N. Mozayani, "Word-level Persian Lipreading Dataset, "2022 12th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, Iran, Islamic Republic of, 2022, pp. 225-230, date of issue : 10.1109/ICCKE57176.2022.9960105.

[22] X. Liu et al., "SynthVSR: Scaling Up Visual Speech Recognition With Synthetic Supervision," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023,pp. 18806-18815, doi:10.1109/CVPR52729.2023.01803.

[23] Oghbaie, M., Sabaghi, A., Hashemifard, K., & Akbari, M. (2021). Advances and Challenges in Deep Lip Reading . ArXiv. /abs/2110.07879.

[24] Ü. Atila, F. Sabaz, Turkish lip-reading using Bi-LSTM and deep learning models. Eng. Sci. Technol. Int. J. 35, 101206 (2022).

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

[26] Rudregowda S, Patil Kulkarni S, H L G, Ravi V, Krichen M. Visual Speech Recognition for Kannada Language UsingVGG16 Convolutional Neural Network Acoustics. 2023; 5(1):343-353.

[27] Czyzewski, A., Kostek, B., Bratoszewski, P. et al. An audio-visual corpus for multimodal automatic speech recognition. J Intell Inf Syst 49, 167–192 (2017)

[28] Li H, Yadikar N, Zhu Y, Mamut M, Ubul K. Learning the Relative Dynamic Features for Word-Level Lipreading. Sensors. 2022; 22(10):3732. https://doi.org/10.3390/s22103732

[29] Fenghour, S., Chen, D., Guo, K., & Xiao, P. (2020). Disentangling Homophemes in Lip Reading using Perplexity Analysis. ArXiv. /abs/2012.07528

[30] K. Thangthai and R. Harvey, ‘‘Improving computer lipreading via DNN sequence discriminative training techniques,’’ in Proc. Interspeech, Aug. 2017, pp. 1–5.

[31] Akhter N, Ali M, Hussain L, Shah M, Mahmood T, Ali A, Al-Fuqaha A. Diverse Pose Lip-Reading Framework. Applied Sciences. 2022; 12(19):9532. https://doi.org/10.3390/app12199532

[32] Tao, Fei, and Carlos Busso. "Bimodal Recurrent Neural Network for Audiovisual Voice Activity Detection." INTERSPEECH. 2017.

[33] Stafylakis .T & Tzimiropoulos G. (2017). Combining Residual Networks with LSTMs for Lipreading. ArXiv. /abs/1703.04105

[34] Noda, Kuniaki, "Audio-visual speech recognition recognition using deep learning." Applied intelligence 42 (2015): 722-737

[35] Matthews, I. et al. “Extraction of Visual Features for Lipreading.” IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002): 198-213.

[36] Koller, Oscar et al. “Deep Learning of Mouth Shapes for Sign Language.” 2015 IEEE International Conference on Computer Vision Workshop (ICCVW) (2015): 477-483

[37] I. Anina, Z. Zhou, G. Zhao and M. Pietikäinen, "OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis," 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia, 2015, pp. 1-5, doi: 10.1109/FG.2015.7163155.

[38] Chung, J., & Zisserman, A. (2017). Lip reading in profile. Ritish Machine Vision Conference, 2017.

[39] T. Afouras, J. S. Chung, A. Zisserman Deep Lip Reading: A comparison of models and an online application INTERSPEECH, 2018.