# VerbiNet: Lip Reading for Indian-Accented English

Gaurav Boob[1], Somya Maheshwari[1], Abhishek Jain[1], and R Sreemathy[2]

[1]Department of Electronics and Telecommunication Engineering, SCTR's Pune Institute of Computer Technology, Pune, India
{gauravboob5,som.maheshwari2002,abhishekjainindore24}@gmail.com

[2]Department of Electronics and Telecommunication Engineering, SCTR's Pune Institute of Computer Technology, Pune, India
rsreemathy@pict.edu

**Abstract.** Lipreading, a technique employed to decipher sequences of lip movements for speech recognition, finds widespread application across diverse contexts, notably aiding individuals with hearing impairments and facilitating comprehension in noisy environments. This research introduces a novel lip-reading framework tailored to enhance recognition accuracy within Indian video datasets. Through extensive testing on a combined dataset of 2000 videos sourced from the GRID dataset and 500 self-recorded videos featuring an Indian accent, our proposed framework achieves notable success. Specifically, it achieves a sentence prediction with a word-error rate (WER) of 15% and a character-error rate (CER) of 4%, a milestone on the Indian dataset for lipreading. These findings underscore the efficacy of our approach in advancing the field of lipreading technology.

**Keywords:** Lip reading, machine learning, long-short-term memory.

## 1 Introduction

The practice of lip-reading involves the meticulous observation of facial movements to decipher spoken language. Research findings reveal that individuals proficient in lip-reading typically achieve an accuracy rate of approximately 20%, indicating a modest level of performance. Notably, those with hearing impairments exhibit even lower levels of accuracy, with reported rates ranging around 17±12% for a curated set of 30 monosyllabic words and approximately 21±11% for an equivalent number of compound words. These empirical findings, documented by Easton and Basala in 1982 [2], underscore the considerable challenges inherent in lip-reading, particularly among individuals with auditory deficiencies.

The findings of the paper by Amy Irvin [3] suggest that accent type may have an influence on visual speech intelligibility and as such may impact the design, and results, of tests of speechreading ability. In this paper, we propose a novel dataset for lip-reading in Indian English. Despite the existence of studies in this domain across different languages, there is presently a lack of research and datasets focused on the Indian English accent. Therefore, this study endeavors to assess the capabilities of state-of-the-art deep learning models in the realm of Indian English lip-reading.

Lip-reading methods include manual techniques and advanced computational approaches with machine learning (ML) algorithms. Manual lip-reading involves human analysis of visual speech signals, interpreting lip, tongue, and facial movements. Computational methods use CNNs for spatial features, RNNs (LSTM, GRU) for temporal dependencies, and transformers for hierarchical representations. Trained on annotated data, these algorithms improve accuracy over manual methods, offering applications across various domains.

Developing a comprehensive lip-reading dataset and creating deep learning models specifically tailored for Indian English holds significant practical implications. These include enhancing assistive technologies for individuals with hearing impairments, refining the accuracy of speech recognition systems, supporting language learning through visual feedback mechanisms, and strengthening applications in security, telecommunications, and medical fields. Integration with hardware like wearable devices or smart glasses could greatly improve real-time communication by offering immediate visual translations of spoken language, thereby facilitating seamless interaction in environments with challenging auditory conditions. This research aims to drive pioneering advancements and foster innovative applications across these diverse domains.

Our research's primary contributions include:

- Development of a novel framework designed to enhance the accuracy of speech recognition within Indian video datasets.
- Creating a Comprehensive Video Dataset: Curating and utilizing a dataset of 500 self-recorded videos featuring speakers with Indian accents, serving as a crucial resource for training and evaluating our lip-reading model.
- Development of the Verbinet model tailored for lip-reading in Indian English, marking a significant advancement aimed at improving accessibility tools for the hearing impaired, enhancing communication technologies, and supporting applications in security, telecommunications, and healthcare.
- Achieving notable performance metrics, including a sentence prediction accuracy with a Word-Error Rate (WER) of 15% and a Character-Error Rate (CER) of 4%, underscoring its capability to accurately transcribe spoken language from visual cues.

## 2 Literature Survey

The exploration of lip reading research commenced in the late 1990s with the seminal paper[7] "Automatic Lip Reading Using Hidden Markov Models." This pioneering study introduced HMMs for automated lip reading, demonstrating an 8.7% reduction in error compared to vector quantization methods. However, it faced challenges in processing complex sentences and handling variability in lip shapes, marking early limitations in the field.

By 2004[8], research advanced with "Asymmetrically Boosted HMM for Speech Reading," which improved feature extraction through Eigenlips combined with HMMs. Evaluated against traditional AdaBoost and HMM classifiers using video and MoCap data, it achieved a notable 19% accuracy increase compared to HMM models alone. Even when compared to AdaBoost and Symmetrically boosted HMM, its performance remained noteworthy. While successful in addressing certain variability issues, limitations persisted concerning vocabulary, phonemes, and variations in lip shapes. Then the paper[9] "Lipreading using Profile versus frontal views" came which focused on extracting visual speech information from the speaker's profile view, utilizing the AVASR system. The study demonstrated that an audio visual automatic speech recognition (AVASR) system is feasible from profile views, albeit with moderate performance degradation compared to frontal video data. In 2007[10], the paper titled "Enhanced Level Building Algorithm for the Movement Epenthesis Problem in Sign Language Recognition" tackled the movement epenthesis (me) challenge in automated sign language recognition. This phenomenon involves gestures bridging consecutive signs with diverse hand features and movements. By enhancing the Level Building algorithm to handle unmodeled me segments and incorporating a trigram grammar model, the study achieved an 83% word level recognition rate on a single-view video dataset. This marked a substantial improvement compared to the 20% accuracy achieved by classical methods. In 2008[11], "Continuous Pose-Invariant Lipreading" introduced a solution for recognizing visual speech despite head movements. The study developed an audio-visual automatic speech recognition (AVASR) system, tested against the CUAVE database, achieving a Word Error Rate (WER) of 61.20%. The integration of local spatiotemporal descriptors for representing and recognizing spoken isolated phrases using visual input was introduced in 2009[12] achieved accuracies of 62% and 70% in speaker-independent and speaker-dependent recognition, respectively. This approach offered advantages such as localized processing and resilience to gradual gray-scale variations. Importantly, it addressed challenges associated with error-prone segmentation of moving lips.

A significant advancement occurred in 2011[13] with 'Lip Reading Using Deep Neural Networks,' which pioneered the application of DNNs for lip reading. This approach achieved a Word Error Rate (WER) of 30% and Character Error Rate (CER) of 12% on the GRID dataset, signifying a substantial improvement over

Hidden Markov Models (HMMs). Additionally, an efficient method for lip detection using OpenCV was introduced, enabling rapid and precise extraction of the lip area and enhancing segmentation accuracy. By 2013 [14], the adoption of Recurrent Neural Networks (RNNs) in "Speech Recognition with Deep RNNs" marked a significant development, leveraging temporal dependencies to achieve a test set error of 17.7% on the TIMIT phoneme recognition benchmark when trained end-to-end with appropriate regularization techniques. The introduction of attention mechanisms in "Listen, Attend, and Spell" [15] represented a pivotal advancement by enhancing performance on extended sequences and intricate sentences. Unlike earlier end-to-end Connectionist Temporal Classification (CTC) models, this approach enabled the network to generate character sequences without presuming independence among the characters. In evaluations on a subset of the Google voice search task, LAS achieved notable results with a word error rate (WER) of 14.1% and 10.3% when employing language model rescoring over the top 32 beams, underscoring its efficacy in real-world applications. In 2016, a groundbreaking development emerged with [16] "LipNet: End-to-End Sentence-Level Lipreading," which integrated Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) while employing Connectionist Temporal Classification (CTC) loss. This approach achieved remarkable accuracy, reaching 95.2% on the GRID dataset. Unlike previous methods focused on word classification, LipNet was trained end-to-end for sentence-level sequence prediction, marking a significant leap in lipreading technology. The integration of audio and visual data using Long Short-Term Memory networks (LSTMs) in "End-to-End Visual Speech Recognition with LSTMs" [17] yielded significant advancements. This approach demonstrated an absolute improvement of 9.7% on the OuluVS2 database and 1.5% on the CUAVE database compared to baseline methods utilizing a similar visual front-end. In 2018[18], "Large-Scale Visual Speech Recognition" showcased improved generalization capabilities through extensive training on the largest visual speech recognition dataset to date. This dataset comprised pairs of text and video clips featuring speakers (adding up to 3,886 hours of video) and achieved a Word Error Rate (WER) of 40.9%.

Recent advancements have concentrated on multi-modal integration and innovations such as self-supervised learning[19], neural networks along with LSTM[20] and transformer models[21] have progressively reduced WERs and Character Error Rates (CERs). However, these approaches still encounter challenges related to real-world noise and high computational costs. Advancements in lip reading technology have prioritized optimizing models for real-time applications, carefully balancing accuracy and computational efficiency. These efforts are pivotal in rendering lip reading technology more practical and applicable for everyday use.

For word-level lip reading, Chung et al. [22] explored various temporal fusion techniques for word-level VSR networks, achieving 61.1% accuracy on the LRW dataset and 25.7% on the LRW-1000 dataset using VGG-M as the frontend network. Stafylakis et al. [23] introduced a C3D-ResNet34 frontend network combined with a BiConvLSTM backend, reporting 83.5% accuracy on LRW and 38.2% on the LRW-1000 dataset. Wang et al. [24] employed ResNet34 and 3D-DenseNet52 as frontend networks, and BiConvLSTM as the backend, obtaining 83.3% and 36.9% accuracy on the LRW and LRW-1000 datasets, respectively. Liu et al. [25] utilized a GCN-based network as the frontend and BiGRU as the backend, achieving 84.25% accuracy on LRW. Martinez et al. [26] used C3D-ResNet18 for the frontend and a multi-scale TCN for the backend, achieving 85.3% accuracy on LRW and 41.4% on LRW-1000. Sheng et al. [27] proposed a lip semantic encoding method, eliminating the need for a predefined lip graph, and achieved 85.7% accuracy on LRW using C3D-ResNet18 and ASST-GCN as the frontend, and MS-TCN as the backend. Ma et al. [28] improved VSR generalization through knowledge distillation, achieving state-of-the-art results with 87.7% accuracy on LRW and 43.2% on LRW-1000 using a C3D-ResNet and MS-TCN architecture. Feng et al. [29] presented a unified framework for audio-visual speech recognition and synthesis, achieving 85.0% accuracy on LRW and 48.0% on LRW-1000 using SE-C3D-ResNet18 and Bi-GRU. Yang et al. [30] introduced a squeeze-and-extract module, achieving state-of-the-art results with 88.5% accuracy on LRW and 50.5% on LRW-1000, using C3D-ResNet18 and ResNet18 with cross-modal mutual learning. Finally, Koumparoulis et al. [31] proposed a resource-efficient network, achieving 89.5% accuracy on LRW with EfficientNetV2-L as the frontend and TCN and Transformer as the backend

For sentence-level lip reading, Assael et al. [32] introduced the first end-to-end sentence-level VSR model, achieving a 1.9% CER and 4.8% WER on the GRID dataset using ST-CNN as the frontend network, Bi-GRU as the backend network, and CTC loss as the learning paradigm. Xu et al. [33] addressed the limitations of the CTC approach, achieving a 1.3% CER and 2.9% WER on the GRID dataset using C3D and HighwayNet as the frontend network, Bi-GRU as the backend, and CTC loss. Afouras et al. [34] compared the CTC and seq2seq models, employing C3D-ResNet18 as the frontend network and a Transformer as the backend, achieving 54.7% CER and 66.3% WER on the LRS2 and LRS3 datasets with CTC loss, and 48.3% CER and 58.9% WER on the same datasets with seq2seq loss. Shillingford et al. [35] proposed using phonemes as output symbols and introduced the largest dataset, LSVSR, achieving 28.3% CER and 40.9% WER on LSVSR, and 55.1% WER on LRS3, utilizing ST-CNN as the frontend and Bi-LSTM as the backend with CTC loss. Zhang et al. [36] integrated causal convolution into the Transformer, achieving 1.3% WER on the GRID dataset, 51.7% WER on LRS2, and 60.1% WER on LRS3 using C3D-ResNet18 as the frontend network and TF-blocks as the backend with seq2seq loss. Makino et al. [37] developed an RNN-T based VSR system, achieving 33.6% WER on LRS3 with ST-CNN as the frontend and RNN-T (BiLSTM) as the backend. Ma et al. [38] proposed a hybrid CTC/Attention model, achieving state-of-the-art results with 37.9% WER on LRS2 and 43.3% WER on LRS3, utilizing C3D-ResNet18 as the frontend and a Conformer and Transformer as the backend with combined CTC and seq2seq loss. Prajwal et al. [39] introduced sub-word output symbols and replaced 2DCNN with a visual transformer, achieving 22.6% WER on LRS2 and 30.7% WER on LRS3 using 3DCNN and VTP as the frontend and Transformer as the backend with seq2seq loss.

# 3 Proposed Method

The proposed method aims to enhance the accuracy and efficiency of lip-reading systems through a comprehensive video processing and neural network-based approach. The video undergoes frame by frame preprocessing to enhance quality and format. It then enters the Verbinet model, which applies advanced neural network algorithms to analyze frames, detect patterns, and generate a concise decoded sentence summarizing the video's content or context.

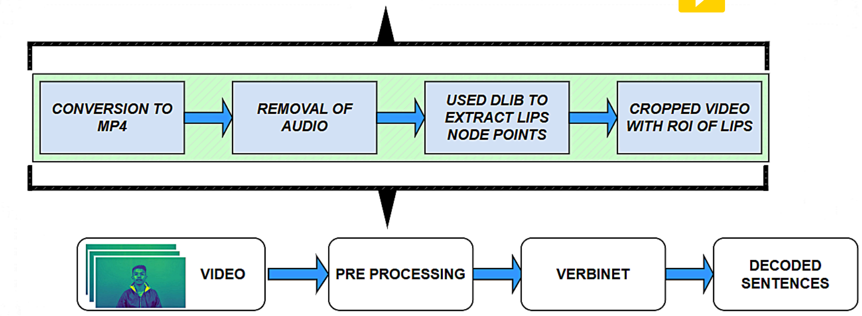


**Fig. 1.** Overview of the complete process.

## 3.1 Dataset

In this research endeavor, a corpus of 500 short Indian videos, each spanning 3 seconds, was meticulously curated. Each video comprised 75 frames, implying a frame rate of 25 frames per second. These videos captured individuals conversing in Indian English, with the verbal content drawn from the GRID dataset. The corpus was strategically partitioned among a cohort of 15 or more individuals, each tasked with delivering a minimum of 20 distinct sentences. To ensure temporal precision, the alignment of each of the 500 videos was meticulously performed frame by frame. This intricate task was executed using VSDC Free Video Editor, a versatile open-source software platform. Notably, the duration of each video was standardized to precisely 3 seconds. Furthermore, the alignments of these videos were meticulously documented and stored in '.align' files, each corresponding to a distinct video within the dataset. Table 1 represents all existing datasets available for lip reading research.

**Table 1.** Overview of the existing dataset in Lip Reading and creation of our own dataset

| Dataset Name | Number of Classes | Number of Videos | Segments |
|---|---|---|---|
| LRW[1] | 51 | 50,000+ | Sentences |
| GRID[6] | 51 | 33 | Sentences |
| OuluVS2[5] | 6 | 80 | Words |
| LRS2[1] | N/A | 100,000 | Words |
| LRS3[1] | N/A | 3,000,000+ | Sentences |
| LRW-1000 | 51 | 1000 | Sentences |
| GRID | 51 | 480 | Sentences |
| MVLRS | N/A | 41,222 | Sentences |
| IBMViaVoice | 290 | 24,325 | Sentences |
| **Indian Dataset** | - | **500** | **Sentences** |

### 3.2 Preprocessing

The GRID Corpus dataset [4] consists of 1000 sentences that have been narrated by 34 speakers each. Nevertheless, we have employed 2000 sentences from a solitary speaker in our dataset, which has led to the construction of 2000 videos retrieved from the GRID dataset. In addition, our dataset comprises 500 videos showcasing Indian speakers who possess similar features to the GRID dataset but with an Indian accent. Each video is a 3-second long video with a frame rate of 25fps. Videos from the GRID dataset are in .mpg format whereas the Indian dataset videos are recorded in .mp4 format. To normalize we first convert all the videos to .mp4 format and then the true preprocessing phase is activated. Figure 1 offers an overview of the preprocessing stages.

Initially, to ensure compatibility and facilitate computational efficiency, the raw video data, composed of 3-second clips encompassing 75 frames, endures a series of standardized preprocessing steps. Firstly, the videos are converted to the extensively supported MP4 format utilizing the FFmpeg tool. This conversion not only ensures uniformity in data format but also streamlines subsequent processing stages. Subsequently, the audio component of the video clips is removed using FFmpeg, as auditory information is extraneous to the task of lip reading and may introduce unnecessary computational overhead. As seen in Figure 2, accurate face detection takes place irrespective of the position of the speaker and with the help of comprehensive capabilities of the Dlib library, the critical landmark points corresponding to the lips region are meticulously extracted from each frame of the video. These landmark points serve as indispensable reference coordinates, facilitating precise localization and monitoring of lip movements across the video sequence. In addition, to enhance the focus on the pertinent lip region and eradicate irrelevant visual clutter, the videos are judiciously cropped based on the region of interest (ROI) delineated by the extracted lip landmark points as seen in Figure 3.

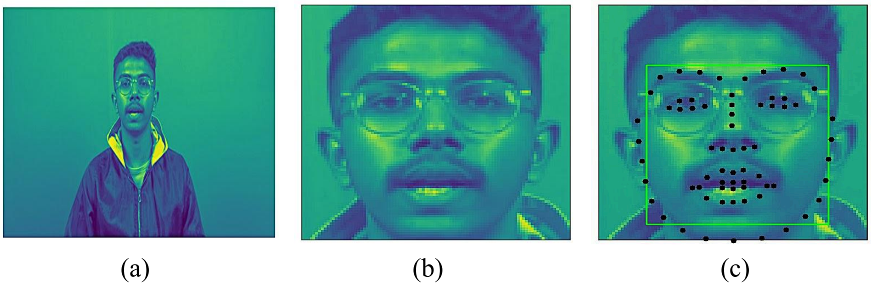

(a) (b) (c)

**Fig. 2.** (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. The RGB channels in the entire training set are standardized to have a mean of zero and a variance of one unit.

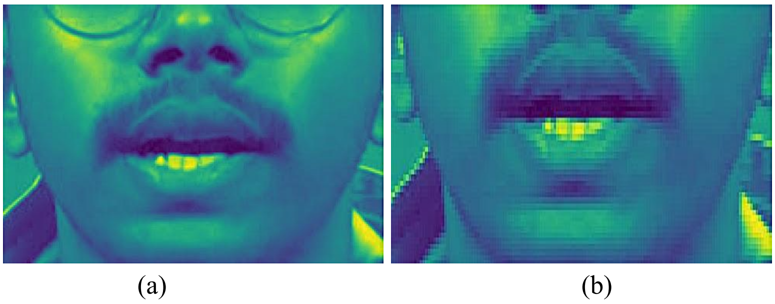

(a) (b)

**Fig. 3.** (a) Cropped video frame in accordance with the lip landmark points. (b) Further cropping of the video frame to (140,46) pixels per frame.

### 3.3 Verbinet

The goal of the VerbiNet model architecture is to reliably identify speech from visual clues of lip movements in lipreading challenges. The architecture utilizes the spatial and temporal information recorded in video sequences, employing multiple convolutional and recurrent neural network layers. In this section, we describe VerbiNet's architecture.

#### 3.3.1 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 their 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 the 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.

#### 3.3.2 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 $j^{\text{th}}$ 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.

### 3.3.3 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.

# 4 Architecture

Figure 4 represents VerbiNet Architecture. The model architecture consists of three primary components: 75 input frames that are processed by three 3D convolutional layers, which are then followed by max pooling and ReLU activation to extract spatial and temporal properties. The spatial information is compressed and then fed into two LSTM layers to capture temporal relationships. The softmax activation function is employed for classification or prediction purposes. Additionally, the model is trained with CTC loss for further optimization.

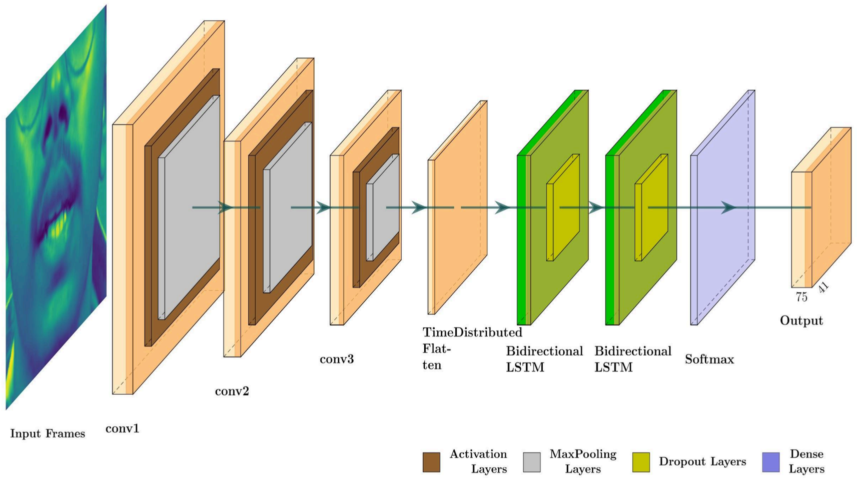


**Fig. 4.** The 75 input frames that make up VerbiNet's architecture are processed by three 3D convolutional layers, each followed by a max pooling layer and ReLU activation. The two LSTM layers receive input from the TimeDistributed layer, which converts 3D tensors to 2D. Softmax is used as the activation function to process the LSTM output after each step.

## 4.1 CTC Loss Function

The model is trained using the backpropagation algorithm and the Adam optimizer with a learning rate of 0.0001. The training process aims to minimize the Connectionist Temporal Classification (CTC) [4] loss function. The CTC loss function computes the negative log-likelihood of the right alignment between the input and output sequences. Without the need for explicit alignment information, the network is trained to align the predicted sequence of characters (or phonemes) with the ground truth sequence of characters in the training data using the Connectionist Temporal Classification (CTC) loss function. Given a sequence of input features X = $(x_1, x_2, ... x_T)$ and a sequence of target labels Y = $(y_1, y_2, ... y_U)$ The CTC loss function computes the negative log-likelihood of the correct alignment between the input and output sequences. The CTC loss function can be expressed as:

$$L_{CTC} = \; - log( \sum_{\pi\varepsilon\beta^{-1}(Y)} P(\pi|X) )$$

where $\beta^{-1}(Y)$ is the set of all possible alignments of Y, $P(\pi|X)$ is the probability of alignment π given input sequence X and the summation is taken over all possible alignments in $\beta^{-1}(Y)$. In actuality, computing $P(\pi|X)$ entails marginalizing across all feasible paths that pass through the same input sequence, and map to the same output sequence, Y.

## 5 Lip Reading Evaluation

Assessing the performance of lip-reading systems is crucial in order to comprehend their efficacy and precision. Word Error Rate (WER) and Character Error Rate (CER) are two often used metrics in the domains of automated speech recognition (ASR) and LipReading. These metrics measure the disparity between the predicted sequences and the reference (ground truth) sequences. WER is a standard metric used to measure the performance of automatic speech recognition (ASR) systems, including LipReading models. It is defined as the percentage of words that are incorrectly predicted. WER accounts for the insertions, deletions, and substitutions required to transform the predicted sequence into the reference sequence. The formula for WER is given by:

$$WER \; = \frac{W_S + W_D + W_I}{N}$$

where $W_S$ is the number of word substitutions, $W_D$ is the number of word deletions, $W_I$ is the number of word insertions, and $N$ is the total number of words in the reference sentence. Similarly, CER is another evaluation metric that is particularly useful for languages with complex word structures or when fine-grained accuracy is required. It is defined similarly to WER but at the character level instead of the word level. CER is given as:

$$CER = \frac{C_S + C_D + C_I}{N}$$

where $C_S$ is the number of character substitutions, $C_D$ is the number of character deletions, $C_I$ is the number of character insertions, and $N$ is the total number of characters in the reference sentence.

# 6 Results

This study aimed to assess the efficacy of our Verbinet model on a dataset containing Indian-accented English, with a particular emphasis on scenarios with overlapping speakers. The Verbinet model demonstrated a character error rate (CER) of 4% and a word error rate (WER) of 15%, suggesting a satisfactory performance considering the challenges associated with accented speech. The model was trained over 120 epochs within intervals of 30 epochs and using checkpoints for further training. The training and validation loss over 30 epochs are illustrated in Figure 5. The training loss demonstrates a general downward trend with some fluctuations, particularly in the earlier epochs, suggesting that the model was learning effectively but encountered some challenges in optimization. The validation loss, on the other hand, initially exhibits a higher degree of fluctuation, which stabilizes in the latter epochs, indicating the model's improved generalization to unseen data over time. Notably, both the training and validation losses converge, highlighting the model's robustness and reduced overfitting. The confusion matrix, as shown in Figure 6, provides insight into the model's prediction accuracy across different words. The matrix is normalized to show the proportion of correct predictions (represented by the diagonal elements) against the actual ground truth. The high accuracy across all classes demonstrates that the model effectively distinguishes between different words despite the accented speech. Each word in the test set was accurately predicted, as indicated by the diagonal values being 1.0, underscoring the model's precise recognition capability.

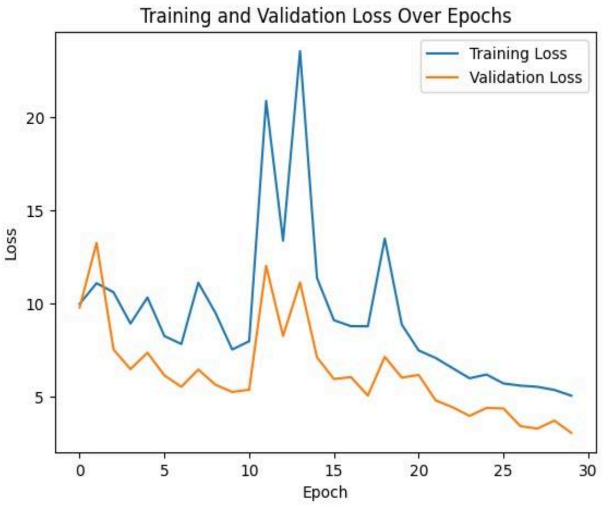
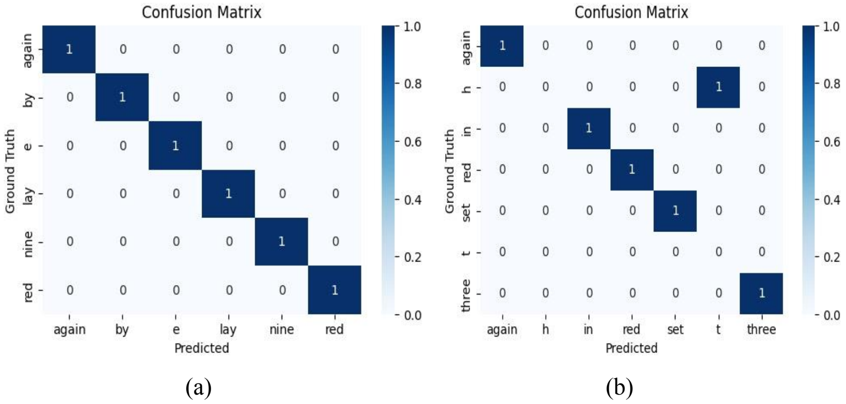


**Fig. 5.** Training loss and Validation loss curve.

The confusion matrices in Figures 6.a and 6.b provide insight into the model's prediction accuracy across different words. These matrices are normalized to show the proportion of correct predictions (represented by the diagonal elements) against the actual ground truth.



| Ground Truth \ Predicted | again | by | e | lay | nine | red |
|---|---|---|---|---|---|---|
| again | 1 | 0 | 0 | 0 | 0 | 0 |
| by | 0 | 1 | 0 | 0 | 0 | 0 |
| e | 0 | 0 | 1 | 0 | 0 | 0 |
| lay | 0 | 0 | 0 | 1 | 0 | 0 |
| nine | 0 | 0 | 0 | 0 | 1 | 0 |
| red | 0 | 0 | 0 | 0 | 0 | 1 |

(a)

| Ground Truth \ Predicted | again | h | in | red | set | t | three |
|---|---|---|---|---|---|---|---|
| again | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| h | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| in | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| red | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| set | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| t | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| three | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

(b)

**Fig. 6.** (a) Confusion matrix representing perfectly predicted characters. (b) Confusion matrix representing some variation in predicting particular characters.

The Verbinet model demonstrates its proficiency in managing Indian accented English speech, while there is still potential for improvement. Based on these criteria, it is evident that the model is highly efficient. However, it is recommended that future efforts concentrate on enlarging the dataset to encompass a broader spectrum of speakers with Indian accents. This will aid in enhancing the model's resilience and precision. Furthermore, conducting tests on speakers who have not been previously encountered will offer a more thorough assessment of the model's ability to generalize. Our study highlights the importance of developing and including a wider range of datasets that represent different accents and dialects in order to improve the performance and usefulness of speech recognition models, especially considering the scarcity of Indian-accented English datasets.

**Table 2.** Comparison With Other Methods

| Method | Dataset | WER |
|---|---|---|
| Lan et al.[41] | GRID CORPUS | 35.0% |
| Wand et al.[42] | GRID CORPUS | 20.4% |
| Chung and Zisserman[22] | LRW | 38.9% |
| Joon Son Chung[40] | LRW | 23.8% |
| **Proposed Method** | **Indian dataset + GRID CORPUS** | **15%** |

Using our novel approach of preprocessing and then applying the VerbiNet model for Lip Reading increases the accuracy tremendously. A comparison of WER with different methods and datasets is shown in Table 2.

## 7 Conclusion

In this work, we have successfully introduced a novel lip-reading framework, VerbiNet, tailored to enhance recognition accuracy within Indian video datasets. Lipreading, a technique employed to decipher sequences of lip movements for speech recognition, is of paramount importance in various contexts, particularly in aiding individuals with hearing impairments and facilitating comprehension in noisy environments. Our approach addresses these needs by focusing on the unique challenges posed by Indian-accented English. The VerbiNet model architecture is meticulously designed to identify speech from visual cues of lip movements using an integrated combination of convolutional and recurrent neural networks. In this research, we curated a comprehensive dataset comprising 500 short Indian videos, each 3 seconds long, alongside 2000 videos from the GRID dataset. Each video was meticulously aligned frame by frame to ensure temporal precision. Our findings demonstrate the efficacy of the VerbiNet model in advancing lipreading technology, particularly in the context of Indian-accented English. The success of this framework paves the way for further research and development, with significant potential applications in improving communication for hearing-impaired individuals and enhancing speech recognition in noisy environments. Future work could explore expanding the dataset, refining the model architecture, and applying the framework to other languages and accents to further enhance its applicability and performance. Additionally, integrating our model with hardware such as CCTV systems and other real-time monitoring devices could extend its utility in surveillance, security, and public safety applications, thereby broadening the scope and impact of lipreading technology.

## References

1. 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.
2. Easton, R.D., Basala, M. Perceptual dominance during lipreading. Perception & Psychophysics 32, 562–570 (1982).
3. Amy Irwin, Michael Pilling, Sharon M. Thomas, An analysis of British regional accent and contextual cue effects on speechreading performance, Speech Communication, Volume 53, Issue 6, 2011, pp. 807-817.
4. A. Graves, F. Gomez, Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, in: International Conference on Machine Learning, 2006, pp. 369–376.
5. Chung, J., & Zisserman, A. (2017). Lip reading in profile. Ritish Machine Vision Conference, 2017.
6. The GRID audiovisual sentence corpus. https://spandh.dcs.shef.ac.uk/gridcorpus/. Accessed 19 June 2024.
7. E. Yamamoto, S. Nakamura and K. Shikano, "Lip movement synthesis from speech based on hidden Markov models," *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, 1998, pp. 154-159, doi: 10.1109/AFGR.1998.670941.
8. J. Rehg, I. Essa and P. Yin, "Asymmetrically Boosted HMM for Speech Reading," in Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 2004 pp. 755-761. doi: 10.1109/CVPR.2004.37
9. Lucey, Patrick, and Gerasimos Potamianos. "Lipreading using profile versus frontal views." In *2006 IEEE Workshop on Multimedia Signal Processing*, pp. 24-28. IEEE, 2006.
10. Yang, Ruiduo, Sudeep Sarkar, and Barbara Loeding. "Enhanced level building algorithm for the movement epenthesis problem in sign language recognition." In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8. IEEE, 2007.
11. Lucey, Patrick, Sridha Sridharan, and David Dean. "Continuous pose-invariant lipreading." In *Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech 2008) incorporating the 12th Australasian International Conference on Speech Science and Technology (SST 2008)*, pp. 2679-2682. International Speech Communication Association, 2008.
12. Zhao, Guoying, Mark Barnard, and Matti Pietikainen. "Lipreading with local spatiotemporal descriptors." *IEEE Transactions on Multimedia* 11, no. 7 (2009): 1254-1265.
13. Pingxian, Yang, Guo Rong, Guo Peng and Fang Zhaoju. "Research on lip detection based on Opencv." *Proceedings 2011 International Conference on Transportation, Mechanical, and Electrical Engineering (TMEE)* (2011): 1465-1468.

14. Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton. "Speech recognition with deep recurrent neural networks." In *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 6645-6649. Ieee, 2013.
15. Chan, William, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. "Listen, attend and spell." *arXiv preprint arXiv:1508.01211* (2015).
16. Assael, Yannis M., Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. "Lipnet: End-to-end sentence-level lipreading." *arXiv preprint arXiv:1611.01599* (2016).
17. Petridis, Stavros, Zuwei Li, and Maja Pantic. "End-to-end visual speech recognition with LSTMs." In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 2592-2596. IEEE, 2017.
18. Shillingford, Brendan, Yannis Assael, Matthew W. Hoffman, Thomas Paine, Cían Hughes, Utsav Prabhu, Hank Liao et al. "Large-scale visual speech recognition." *arXiv preprint arXiv:1807.05162* (2018).
19. Sheng, Changchong, Matti Pietikäinen, Qi Tian, and Li Liu. "Cross-modal self-supervised learning for lip reading: When contrastive learning meets adversarial training." In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 2456-2464. 2021.
20. Exarchos, Themis, Georgios N. Dimitrakopoulos, Aristidis G. Vrahatis, Georgios Chrysovitsiotis, Zoi Zachou, and Efthymios Kyrodimos. "Lip-Reading Advancements: A 3D Convolutional Neural Network/Long Short-Term Memory Fusion for Precise Word Recognition." *BioMedInformatics* 4, no. 1 (2024): 410-422.
21. Wang, Huijuan, Gangqiang Pu, and Tingyu Chen. "A lip reading method based on 3D convolutional vision transformer." *IEEE Access* 10 (2022): 77205-77212.
22. J. S. Chung and A. Zisserman, "Lip reading in the wild," in ACCV,2016, pp. 87–103.
23. T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," in Interspeech, 2017.
24. C. Wang, "Multi-grained spatio-temporal modeling for lip-reading," arXiv:1908.11618, 2019.
25. H. Liu, Z. Chen, and B. Yang, "Lip graph assisted audio-visual speech recognition using bidirectional synchronous fusion." in interspeech, 2020, pp. 3520–3524.
26. B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in ICASSP, 2020, pp. 6319–6323.
27. C. Sheng, X. Zhu, H. Xu, M. Pietikainen, and L. Liu, "Adaptive semantic-spatio-temporal graph convolutional network for lip reading," IEEE TMM, 2021.
28. P. Ma, B. Martinez, S. Petridis, and M. Pantic, "Towards practical lipreading with distilled and efficient models," in ICASSP, 2021, pp. 7608–7612.
29. D. Feng, S. Yang, S. Shan, and X. Chen, "Learn an effective lip reading model without pains," arXiv:2011.07557, 2020.
30. C.-C. Yang, W.-C. Fan, C.-F. Yang, and Y.-C. F. Wang, "Crossmodal mutual learning for audio-visual speech recognition and manipulation," in AAAI, 2022.
31. A. Koumparoulis and G. Potamianos, "Accurate and resource-efficient lipreading with efficientnetv2 and transformers," in ICASSP. IEEE, 2022, pp. 8467–8471.
32. Y. M. Assael, B. Shillingford, S. Whiteson, and N. De Freitas, "Lipnet: End-to-end sentence-level lipreading," arXiv:1611.01599, 2016.
33. K. Xu, D. Li, N. Cassimatis, and X. Wang, "Lcanet: End-to-end lipreading with cascaded attention-ctc," in FG, 2018, pp. 548–555.
34. T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," IEEE TPAMI, 2018.

35. B. Shillingford, Y. Assael, M. W. Hoffman, T. Paine, C. Hughes, U. Prabhu, H. Liao, H. Sak, K. Rao, L. Bennett et al., "Large-scale visual speech recognition," arXiv:1807.05162, 2018.
36. X. Zhang, F. Cheng, and S. Wang, "Spatio-temporal fusion based convolutional sequence learning for lip reading," in ICCV, 2019, pp. 713–722.
37. T. Makino, H. Liao, Y. Assael, B. Shillingford, B. Garcia, O. Braga, and O. Siohan, "Recurrent neural network transducer for audio-visual speech recognition," in IEEE ASRU workshop, 2019, pp. 905–912.
38. P. Ma, S. Petridis, and M. Pantic, "End-to-end audio-visual speech recognition with conformers," in ICASSP, 2021, pp. 7613–7617.
39. K. Prajwal, T. Afouras, and A. Zisserman, "Sub-word level lip reading with visual attention," in CVPR, 2022, pp. 5162–5172.
40. Son Chung, J., Senior, A., Vinyals, O., Zisserman, A.: Lip reading sentences in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6447–6456 (2017)
41. Lan, Y., Harvey, R., Theobald, B., Ong, E.-J., Bowden, R.: Comparing visual features for lipreading. In: International Conference on Auditory-Visual Speech Processing 2009, pp. 102–106 (2009)
42. Wand, M., Koutn´ık, J., Schmidhuber, J.: Lipreading with long short-term memory. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6115–6119 (2016). IEEE