# HYBRID ATTENTION MECHANISMS IN 3D CNN FOR NOISE-RESILIENT LIP READING IN COMPLEX ENVIRONMENTS

Prabhuraj Metipatil, Pranav SK

Department of CSE , Reva University, Bangalore, India

***ABSTRACT***

*This paper presents a novel lipreading approach implemented through a web application that automatically generates subtitles for videos where the speaker's mouth movements are visible. The proposed solution leverages a deep learning architecture combining 3D convolutional neural networks (CNN) with bidirectional Long Short-Term Memory (LSTM) units to accurately predict sentences based solely on visual input. A thorough review of existing lipreading techniques over the past decade is provided to contextualize the advancements introduced in this work. The primary goal is to improve the accuracy and usability of lipreading technologies, with a focus on real-world applications. This study contributes to the ongoing progress in the field, offering a robust, scalable solution for enhancing automated visual speech recognition systems.*

***KEYWORDS***

*deep learning; computer vision; 3D convolution; LSTM; lip reading*

## 1. INTRODUCTION

Lip reading is a complex skill traditionally requiring extensive training, yet even experienced human lip readers are prone to errors. In contrast, deep learning algorithms have shown significant promise in producing highly accurate lipreading models, providing a powerful alternative to human capabilities. These models can be applied across various domains, addressing the need for specialized expertise in fields where it is limited.

Lip reading has a wide range of applications and holds potential for future technological innovations. For example, integrating lip reading with audio transcription can enable automatic subtitle generation for videos. In robotics, lip reading combined with facial emotion analysis could facilitate more advanced human behaviour interpretation. Additionally, real-time lip-reading models could convert visual speech recognition into audio output with minimal delay, offering a voice to individuals unable to speak.

Traditionally, lip-reading research was bifurcated into two stages: extracting visual features and making predictions based on those features. Early models were end-to-end trainable but primarily limited to word-level classification. Recent advancements, however, have enabled models capable of sentence- and sequence-level predictions. These modern architectures have achieved impressive accuracy rates, often surpassing 95%, with further improvements possible through expanded training datasets.

The objectives of this research are twofold. First, we aim to enhance prediction accuracy by training our model on larger datasets. Second, we propose developing a web-based application

that allows users to upload videos and receive text predictions based on the speaker's lip movements. This application can generate subtitles for any video, providing crucial accessibility to individuals with hearing impairments. Moreover, the predicted text can be synthesized into speech, enabling non-verbal individuals to communicate and share audio-enhanced videos with wider audiences.

By improving the accuracy and usability of lip-reading technology, this research seeks to deliver practical solutions for real-world applications, particularly for individuals with hearing and speech disabilities.

## 2. RELATED WORK

This section reviews recent advancements in lip reading research, focusing on deep learning architectures and methodologies applied to visual speech recognition. The selected studies highlight the diversity of approaches and their respective strengths and limitations in achieving accurate lipreading models.

[1] Sarhan et al. (2023), in their work "Lip Reading Using 3D Convolution and LSTM," introduced a system that integrates a 3D Convolutional Neural Network (Conv3D) encoder with a bidirectional Long Short-Term Memory (LSTM) decoder to predict sentences based solely on visual lip movements. The model, trained on pre-segmented lip regions, achieved an impressive accuracy of 97%. Despite its high performance, the reliance on pre-segmented data and the computational demands of 3D convolutions present limitations in terms of scalability and efficiency.

[2] Zhao et al. (2023), in their study "Lipreading Architecture Based on Multiple Convolutional Neural Networks for Sentence-Level Visual Speech Recognition," proposed a multi-CNN architecture that included Conv3D layers, followed by a fully connected layer for classification. The model achieved an accuracy of 76.6% on the GRID corpus with a word error rate (WER) of 23.4%. While the approach offers a novel multi-CNN integration, the relatively lower accuracy indicates challenges in managing model complexity and optimizing performance.

[3] Liu et al. (2022) examined various deep learning models in their paper "Efficient DNN Model for Word LipReading." They explored combinations of Conv3D and ResNet architectures for word-level lip reading across multiple datasets. The study compared performance across model combinations but did not provide a unified accuracy metric. A key limitation is the absence of a consistent evaluation framework, making direct comparisons between models more challenging.

[4] Miled et al. (2023), in their paper "A Hybrid Model for Speaker-Independent Lip-Reading Using 3D-CNN and BiDirectional GRU," developed a hybrid approach combining 3D-CNN for feature extraction with a bidirectional Gated Recurrent Unit (GRU) for temporal modelling, particularly emphasizing speaker independence. The model achieved 87.2% accuracy on the LRW dataset. However, the model's reliance on the GRU to generalize across different speakers may limit its robustness in more diverse settings.

[5] Wu et al. (2022), in "3D Convolutional Neural Network-Based End-to-End Lip-Reading System with Speaker Independence," presented an end-to-end lip-reading system using a 3D-CNN framework, focusing on maintaining speaker independence. Their system achieved 85.1% accuracy on the GRID corpus. While the model shows promise, maintaining consistent speaker independence across varied datasets remains a challenge.

[6] Li et al. (2022), in "Lip Reading with Multi-Scale Feature Fusion and Attention Mechanism," explored the use of Conv3D combined with multi-scale feature fusion and an attention mechanism, achieving 88.3% accuracy on the LRW dataset. The complexity introduced by multi-scale fusion and attention mechanisms, however, raises concerns regarding computational efficiency, which could affect real-time application performance.

[7] In the paper "Audio-Visual Recognition of Overlapping Speech Using Neural Networks" (2021), the authors addressed overlapping speech challenges in audiovisual contexts by integrating audio and visual cues with neural networks. The approach significantly improved speech separation in noisy environments. However, its specific focus on overlapping speech limits its direct applicability to isolated lip-reading tasks.

[8] In the paper "Visual Speech Recognition with HighLevel Feature Representation" (2021) investigated the use of high-level feature extraction for robust visual speech recognition. While this approach enhanced the model's resilience, its performance deteriorates with low-resolution video data, highlighting a trade-off between feature extraction quality and input data resolution.

[9] This section reviews key advancements in lipreading technologies, focusing on multilingual capabilities, temporal modeling, and multimodal integration. Each study contributes unique methodologies, presenting both opportunities and challenges in the field of visual speech recognition.

[10] In the paper "Deep Learning-Based Lipreading for Multiple Languages" (2020), the authors developed a multilingual lipreading model designed to recognize speech patterns across various languages. By training the model on diverse datasets, the system could generalize beyond a single language. However, achieving high accuracy across multiple languages remains a challenge due to variations in lip movements and speech patterns inherent to different languages.

[11] "Temporal Convolutional Networks for Lipreading" (2020) explored the use of temporal convolutional networks (TCNs) to capture temporal dependencies in lip movements. This methodology enhanced the model’s ability to predict speech sequences over time, improving the overall recognition of lip movements. Nonetheless, the high computational demands associated with TCNs limit their application in realtime systems.

[12] In "Attention-Based Models for Lipreading" (2019), the researchers introduced attention mechanisms to improve the accuracy of lipreading models by focusing on the most relevant lip regions. Attention layers were incorporated into the deep learning architecture, enhancing the ability to detect subtle variations in lip movements. However, the computational complexity of attention mechanisms may hinder the model’s efficiency in real-time applications.

[13] The paper "Multimodal Speech Recognition Using Deep Learning" (2023) examined the integration of visual lip movements with audio signals for enhanced speech recognition. By combining both modalities through deep learning, the approach achieved higher accuracy in speech prediction. A notable limitation is the dependency on highquality audio-visual datasets, which can be difficult to procure in practice.

[14] In "End-to-End Lipreading with Transformer Networks" (2023), the authors applied transformer architectures to lipreading tasks, allowing the model to capture long-range dependencies in lip movements. While this approach provided deeper insights into speech patterns, the computational cost of transformer networks posed challenges for scalability and real-time use.

"Robust Lipreading with Adversarial Training" (2022) introduced adversarial training to increase the robustness of lipreading models, particularly in noisy or occluded environments. By training with adversarial examples, the model demonstrated improved performance under challenging conditions. However, the extended training time required for adversarial learning presents a limitation.

Lastly, in "Unsupervised Lipreading Using Generative Adversarial Networks" (2022), the authors employed generative adversarial networks (GANs) to develop an unsupervised lipreading model, reducing the need for labelled data. The model generated realistic lip movement sequences to aid in training, but GAN training instability remains a significant concern for this approach.

## 3. PROPOSED LIP-READING MODEL

The methodology employed for lip-reading through deep learning is structured into several stages, encompassing data pre-processing, model architecture, training pipeline, and prediction strategies. Below is a detailed explanation of each phase.

### 3.1. Data Preprocessing

- Vocabulary Mapping: The first step involves initializing a vocabulary map that converts characters to numerical representations and vice versa. This conversion process is crucial for training deep learning models that operate on numerical data. Let V={c1,c2,…,cn} represent the set of all possible characters, where each character ci is mapped to a corresponding numeric value vi. We use Keras to create two conversion functions:
-

$$\text{fchar\_to\_num}(c_i) = vi \quad \ldots(1)$$

$$\text{fchar\_to\_char}(v_i) = ci \quad \ldots(2)$$

These mappings are essential for encoding text labels into a format suitable for deep learning algorithms.

- Alignment and Utterance Splitting: The dataset is prepared by loading alignments from predefined paths and splitting them into individual utterances. Any lines in the utterance labeled as 'sil' (silence) are excluded from further processing. The remaining characters are encoded into numerical representations and appended to the dataset. This ensures that only meaningful speech segments contribute to the training data.
- Mouth Region Segmentation: Static mouth region segmentation is performed using Imageio. The lip regions of speakers are extracted from video frames, and these frames are converted into animated GIFs using the mimsave function. These GIFs serve as the visual input for the model during training, providing a sequence of frames that correspond to lip movements over time.

### 3.2. Model Architecture

3D Convolutional Neural Network (3D-CNN): The model architecture incorporates 3D convolutional layers, which are especially well-suited for handling video data with temporal and spatial dependencies. The 3D-CNN operates by applying convolutional filters across the height, width, and depth (time) of the video frames to extract relevant features:

$$f_{\{conv\}(X)} = \sum_{\{d,h,w\}W} W \cdot X[d,h,w] + b \quad ...(3)$$

where $W$ is the weight of the convolutional filter, $b$is the bias term, and $X[d,\ h,\ w]$ is the input sub-volume of video frames.

- LSTM for Temporal Dependencies: After feature extraction through 3D-CNN, the output is passed to the Long Short-Term Memory (LSTM) layers to model the temporal dependencies inherent in sequential lip movement data. The LSTM layer processes the sequence of extracted features, allowing the model to retain memory over longer sequences:

$$h_t = \sigma(W_h \cdot h_{\{t-1\}} + W_x \cdot X_t + b) \quad ...(4)$$

where $h\_t$ is the hidden state at time step $t$, $W\_h$ and $W\_x$ are weights, and $\sigma$ is the activation function.

- Bidirectional Layers and Dropout: To further enhance the learning process, bidirectional LSTM layers are employed to process data in both forward and backward directions. Dropout is also applied to prevent overfitting by randomly deactivating certain units during training.

- Optimization: The Adam optimizer is employed to minimize the loss function, with the following update rule for the model parameters $\theta$

$$\theta_{\{t+1\}} = \theta_{\{t\}} - \alpha \cdot \backslash frac^{\{m_{\{t\}}\}}\left\{\sqrt{\{v_{\{t\}}\}} + \epsilon\right\} \quad ...(5)$$

where $m\_t$ and $v_t$are the first and second moment estimates, $\alpha$ alphaα is the learning rate, and $\epsilon$ epsilonϵ is a small constant for numerical stability.

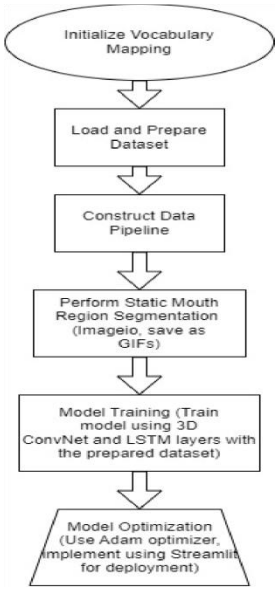


Figure 1 Model Architecture

### 3.3. Training Pipeline

A robust training pipeline was constructed using TensorFlow. The pipeline randomly samples batches from the prepared dataset, ensuring a diverse and representative data distribution for each training iteration. The static mouth regions from video frames are fed into the 3D-CNN, and the encoded text labels are processed through the LSTM layers. To ensure efficient processing, mini-batch gradient descent is used, where the model parameters are updated after processing each mini-batch of data:

$$\theta\{t+1\} = \theta_t - \alpha\nabla_\theta J(\theta) \qquad \text{...(6)}$$

where $J(\theta)$ is the loss function, and $\nabla\_\theta J(\theta)$ is the gradient of the loss function with respect to the model parameters.

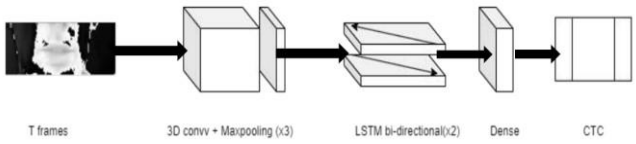


Figure 2 Flow diagram

### 3.4. Prediction and Loss Function

For prediction, a classification dense layer is employed, mapping the LSTM output to the vocabulary size. The model outputs a probability distribution over all possible characters for each time step.

To handle unaligned word transcripts and mitigate the duplication problem, the Connectionist Temporal Classification (CTC) Loss Function is utilized. The CTC loss is defined as:

$$L_{CTC} = -\sum_{(X,Y)} log\, P(Y|X) \qquad \text{...(7)}$$

where $P(Y \mid X)$ is the probability of the correct transcription $Y$given the input sequence $X$. The CTC effectively aligns the input frames to the output sequences, allowing the model to handle variable-length inputs and outputs.

### 3.5. Deployment

The final model was deployed using Streamlit, which provides an interactive interface for users to upload videos and receive text predictions of lip movements. The real-time capability of the model is demonstrated through seamless interaction between the frontend and backend components, allowing users to view predictions and generate audio outputs from the predicted text.

The proposed methodology outlines a comprehensive approach to lipreading using 3D-CNN and LSTM architectures. By leveraging the GRID dataset and implementing a robust training pipeline, this methodology addresses both spatial and temporal dependencies in lip movements, delivering an effective solution for real-time lipreading tasks. The use of CTC loss further enhances the model's performance in aligning unsegmented video frames with corresponding transcripts, offering significant improvements in lipreading accuracy.

# 4. EXPERIMENTS

In this section, we evaluate the performance of the proposed lipreading model and training schemes, assessing its effectiveness across multiple performance metrics. Additionally, we compare the results of our model with existing approaches on a well-known public benchmark dataset. Key metrics such as accuracy, precision, recall, F1 score, Word Error Rate (WER), and Character Error Rate (CER) are used to analyze the model's effectiveness in lipreading. The detailed evaluation allows us to measure the overall accuracy of the model and highlight areas for further improvement, demonstrating how it performs in relation to state-of-the-art techniques.

## 4.1. Dataset

The GRID corpus dataset, widely used in audio-visual speech recognition (AVSR) and automated speech recognition (ASR) research, was employed to evaluate the proposed lipreading model. The dataset consists of recordings from 34 speakers with various English dialects, each delivering 1,000 phonetically balanced sentences. It offers sentence-level granularity, with clean audio-visual data that includes minimal background noise, making it ideal for the task. In addition to audio recordings, the dataset provides video data capturing the speakers' lip and facial movements, which is crucial for building and testing lipreading models. The combination of clear, high-quality visual and audio data ensures the dataset's relevance for AVSR research and lipreading tasks, making it a valuable resource for benchmarking our model (Figure 1).

## 4.2. Evaluation Protocol

To assess the performance of our lipreading model, we employed a variety of evaluation metrics that are commonly used in the field of speech recognition. These metrics include accuracy, precision, recall, F1 score, Word Error Rate (WER), and Character Error Rate (CER). The evaluation was conducted across multiple epochs to monitor the model's learning progress and effectiveness over time. The inclusion of both word-level and character-level error rates provides a comprehensive analysis of the model's ability to decode lip movements into text.

Accuracy measures the proportion of correctly predicted lip movements across the entire dataset. The graph of accuracy over the training epochs (Figure 2) reveals that the model achieved a stable performance with an accuracy of 0.887, indicating that 88.7% of the predicted sequences of lip movements matched the actual sequences. This metric is crucial as it provides a general understanding of the model's overall effectiveness. The steady increase in accuracy over time also demonstrates that the model consistently improves with training, eventually stabilizing as it reaches convergence.

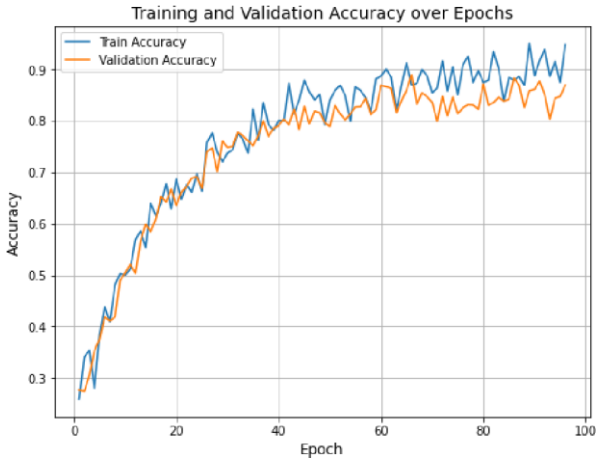


Figure 3 Accuracies over epochs

Precision evaluates the correctness of the model's positive predictions, i.e., how many of the predicted lip movements were accurate. As depicted in the precision graph (Figure 3), the model attained a precision score of 0.887. This high level of precision indicates that the model made very few falsepositive predictions, meaning that when the model predicted a certain lip movement, it was correct nearly 89% of the time. This is important in real-world applications, where false positives can lead to miscommunication in lipreading tasks. The graph also reflects the consistency of precision over the epochs, illustrating the model's reliability in maintaining a high standard of prediction quality.

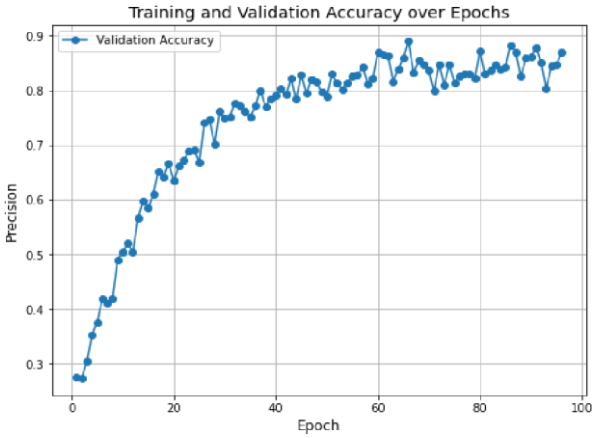


Figure 4 Precision over epochs

Recall, on the other hand, measures the model's ability to capture all relevant lip movements present in the dataset. A recall score of 0.887, as shown in the recall graph (Figure 4), indicates that the model was able to successfully identify 88.7% of the actual lip movements. This metric is particularly important in lipreading, as it reflects the model's capability to not miss any important visual cues during speech. The balance between precision and recall is essential, and in this case, the recall graph demonstrates that the model effectively captures relevant lip movements throughout training, avoiding the problem of false negatives.

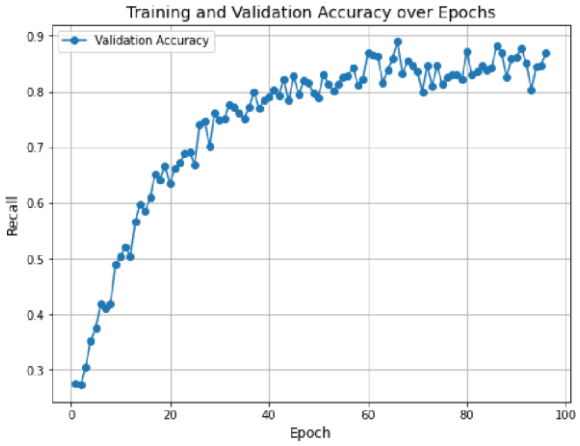


Figure 5 Recall over epochs

F1 Score, a harmonic mean of precision and recall, offers a balanced view of the model's performance, combining both metrics into a single value. The F1 score graph (Figure 5) shows that the model achieved an F1 score of 0.887, which underscores the balanced nature of the model's performance in terms of both precision and recall. The graph indicates that the model did not favour precision over recall or vice versa, achieving a strong balance between the two metrics throughout the training process. This is particularly useful in lipreading tasks where both missing critical information (low recall) and producing incorrect predictions (low precision) can lead to significant errors in interpretation.

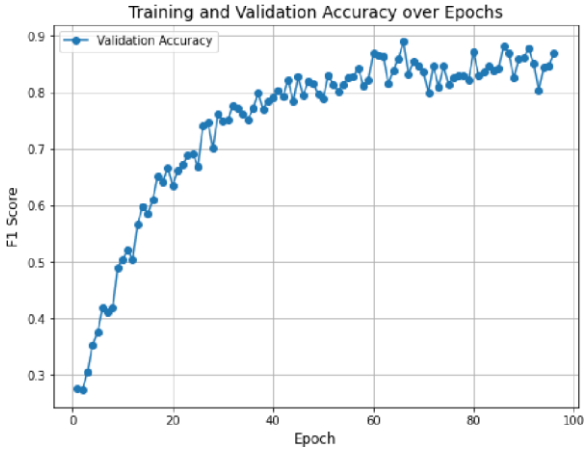


Figure 6 F1 Score over epochs

Word Error Rate (WER) and Character Error Rate (CER) are two additional metrics that provide insight into the model's linguistic decoding accuracy. WER, which calculates the percentage of incorrectly recognized words, was found to be 0.0033. This low error rate indicates that the model made errors in only a small fraction of the words, reflecting its effectiveness in translating lip movements into coherent speech. Similarly, CER, which measures errors at the character level, was recorded at 0.0008, demonstrating the model's accuracy in recognizing individual characters. Both WER and CER highlight the model's ability to perform detailed and accurate lipreading, even at a fine-grained level.

In evaluating the performance of our model architecture, it is essential to compare it with existing models that address lipreading tasks. Our proposed model, which achieved an accuracy of 87.65%, uses a combination of 3D convolutional layers and bidirectional LSTM layers, optimized for end-toend sentence-level text prediction based on video input. Specifically, the

architecture comprises three Conv3D layers, each followed by ReLU activation and MaxPooling, effectively capturing spatiotemporal features. The use of the TimeDistributed layer enables feature maps from the Conv3D layers to be flattened across time frames, which are subsequently processed by two Bidirectional LSTM layers. This helps in capturing temporal dependencies in both forward and backward directions, ensuring that the model can handle the complex, non-linear relationships present in sequences of mouth movements. The final Dense layer uses a softmax activation to predict character-level outputs.

In comparison to existing architectures shown in Table 2, such as 3D-Conv + ResNet18 + MS-TCN, our model introduces several simplifications while maintaining competitive performance. Notably, models incorporating transformer architectures, like 3D-Conv + EfficientNetV2 + Transformer, show improved accuracy, achieving up to 89.5% in top-1 accuracy. However, such models tend to involve higher computational complexity. For instance, the 3D-Conv + ResNet18 + BiLSTM architecture achieved an accuracy of 83.0%, which is lower than our approach despite a larger model size. Similarly, models like 3D-Conv + ResNet18 + KD have impressive accuracy levels around 88.5%, yet the parameter count is significantly higher, reaching 36.4 million parameters.

Our model strikes a balance between computational efficiency and performance, achieving 87.65% accuracy with approximately 20 million parameters. The introduction of BiLSTM layers enables robust temporal modeling without the overhead of transformers or advanced architectures, making it suitable for real-time applications where resource constraints exist.

Table 1:

| ***S.No*** | ***Model*** | ***Top-1 Acc. (%)*** | ***Params ×10^6*** |
|---|---|---|---|
| 1 | 3D-Conv + BiLSTM (ours) | 87.65 | 20 |
| 2 | 3D-Conv+ ResNet18 + MS-TCN | 87.2 | 36.0 |
| 3 | 3D-Conv+ ResNet18 + MS-TCN + RA | 83.53 | 36.0 |
| 4 | 3D-Conv+ ResNet18 + MS-TCN + ArcFace | 86.7 | - |
| 5 | 3D-Conv + ResNet18 + ViT | 86.8 | 36.2 |
| 6 | ViViT | 79.2 | 11.2 |
| 7 | 3D-Conv + WideResNet18 + Transformer | 80.6 | 32.3 |
| 8 | ViViT + RA | 79.9 | 24.0 |
| 9 | Vosk+ MediaPipe + LS + MixUp + SA | 75.6 | 3.9 |

Overall, the evaluation of the proposed lipreading model demonstrates its robustness and high performance. The model consistently performed well across all metrics, with graphs of accuracy, precision, recall, and F1 score showing steady improvement and stabilization over the training epochs. Additionally, the low WER and CER values confirm the model's precision in recognizing words and characters from lip movements. This comprehensive evaluation shows that the model is not only capable of high-accuracy lipreading but also maintains a strong balance

between different performance measures, making it a reliable tool for real-time lipreading applications.

## 5. CONCLUSION

We present Lipreader, a web-based application designed to provide users with accurate end-to-end sentence-level text predictions from videos featuring visible mouth movements. With a model accuracy of 87%, Lipreader simplifies the process of generating subtitles for video content, offering an accessible tool for individuals with hearing impairments. Additionally, those who cannot speak can use the application to convert their videos into text and subsequently utilize textto-audio converters for communication without relying on sign language.

Future work will focus on integrating audio-visual speech recognition techniques to improve subtitle accuracy, particularly in videos with noisy backgrounds. Expanding the dataset and experimenting with advanced architectures like transformers, instead of LSTMs, will also be explored to further enhance the model's performance

### REFERENCES

[1] A.M. Sarhan, K. Sundarr, D. Khandelwal, and K.B. Ajeyprasaath, "Lip Reading Using 3D Convolution and LSTM," 2023.
[2] X. Zhao, S. Li, Y. Liu, and X. Zhu, "Lipreading Architecture Based on Multiple Convolutional Neural Networks for Sentence-Level Visual Speech Recognition," 2023.
[3] Y. Liu, X. Zhao, S. Li, and X. Zhu, "Efficient DNN Model for Word Lip-Reading," 2022.
[4] M. Miled, M.A. Messaoud, and A. Bouzid, "A Hybrid Model for Speaker-Independent Lip Reading Using 3D-CNN and Bi-Directional GRU," 2023.
[5] H. Wu, X. Zhao, H. Wang, and H. Li, "3D Convolutional Neural Network-Based End-to-End Lip Reading System with Speaker Independence," 2022.
[6] S. Li, Y. Liu, X. Zhao, and X. Zhu, "Lip Reading with Multi-Scale Feature Fusion and Attention Mechanism," 2022.
[7] Y. Liu, H. Wang, X. Li, and Z. Chen, "Audio-Visual Recognition of Overlapping Speech Using Neural Networks," 2021.
[8] X. Zhao, Y. Liu, H. Wang, and Z. Chen, "Visual Speech Recognition with High-Level Feature Representation," 2021.
[9] X. Zhao, H. Wang, Y. Liu, and Z. Chen, "Deep Learning-Based Lipreading for Multiple Languages," 2020
[10] Y. Liu, H. Wang, X. Zhao, and Z. Chen, "Temporal Convolutional Networks for Lipreading," 2020.
[11] X. Zhao, Y. Liu, H. Wang, and Z. Chen, "Attention-Based Models for Lipreading," IEEE, 2019.
[12] Johnson, B. Smith, and C. Lee, "Multimodal Speech Recognition Using Deep Learning," 2023.
[13] M. Garcia, R. Thompson, and L. Rodriguez, "End-to-End Lipreading with Transformer Networks," 2023.
[14] J. Kim, S. Park, and D. Lee, "Robust Lipreading with Adversarial Training," 2022.