Removing Noisy and Irrelevant Information in Lipnet for Silent Communication Using Residual Attention Block

CH Jayaprakash Reddy
Department of Information Technology,
Vardhaman College of Engineering
Hyderabad-Telangana, 501286, India
jayaprakashreddy0125@gmail.com

M Sona Reddy
Department of Information Technology,
Vardhaman College of Engineering
Hyderabad-Telangana, 501286, India
sonaareddy27@gmail.com

Y Charan Raj
Department of Information Technology,
Vardhaman College of Engineering
Hyderabad-Telangana,501286,India
yeluricharannetha143@gmail.com

CH Meghanadh
Department of Information Technology,
Vardhaman College of Engineering
Hyderabad-Telangana, 501286, India
Charpameghanadh97@gmail.com

Dr B K Madhavi
Department of Information Technology,
Vardhaman College of Engineering
Hyderabad-Telangana, 501286, India
Madhavi1593@vardhaman.org

***Abstract*— In order to facilitate communication in settings where auditory signals are inaccurate or missing, silent visual speech recognition attempts to decode spoken content only from lip movements. Although LipNet, a groundbreaking end-to-end architecture for sentence-level lip-reading, performs admirably, it is nevertheless susceptible to irrelevant facial motion, background noise, and changes in lighting. In order to minimize noisy, redundant features while maintaining significant spatiotemporal patterns, we propose an improved LipNet model in this study by incorporating Residual Attention Blocks (RABs) into the convolutional feature extraction stage. While residual connections preserve steady gradient flow during training, the attention mechanism selectively highlights discriminative lip-motion cues by combining channel and spatial weighting. When compared to the baseline LipNet, experimental evaluation shows that the suggested model achieves better word-level and character-level accuracy with notable decreases in Word Error Rate (WER). These findings demonstrate how well residual attention reinforces strong lip-motion representation and point to a possible path for silent communication systems that can withstand noise.**

Keywords—LipNet, Visual Speech Recognition, Silent Communication, Lip Reading, Residual Attention Block, Spatiotemporal Feature Extraction, Deep Learning, 3D Convolutional Neural Networks, Bidirectional GRU, Connectionist Temporal Classification, Attention Mechanism.

I. INTRODUCTION

The goal of silent visual speech recognition (VSR), also referred to as lip-reading, is to comprehend spoken words solely from lip movements without the need of auditory cues. Due to its many uses, such as communication in noisy settings, assistance for those with hearing loss, quiet command interfaces, surveillance, and privacy-preserving human–computer interaction, this modality has drawn a lot of attention. Recent developments in deep learning have made it possible to create end-to-end architectures that can simulate intricate spatiotemporal patterns in human speech articulation, whereas conventional lip-reading techniques depended on hand-engineered features and probabilistic models.

By combining 3D Convolutional Neural Networks (3D-CNNs) with Bidirectional Recurrent Neural Networks (Bi-GRUs) and a Connectionist Temporal Classification (CTC) loss framework, LipNet—one of the first end-to-end sentence-level lip-reading models—showed significant advancements. Even while LipNet works well, it still has problems with real-world variances such backdrop clutter, shifting lighting, head motions, and irrelevant facial dynamics. The model's capacity to collect fine-grained lip-motion information necessary for precise recognition is diminished by these factors, which add noise to the visual input.

Attention processes have become a potent improvement in deep visual models to overcome these difficulties. Neural networks can reduce irrelevant or noisy features while concentrating on the most informative regions or channels thanks to attention. In particular, Residual Attention Blocks (RABs) combine residual connections with attention modules, enabling the model to choose improve feature representations without sacrificing learning stability or gradient flow. Although these techniques have demonstrated notable advancements in a number of computer vision tasks, they are still not well understood in the field of silent visual speech recognition.

In this study, we suggest extending LipNet by adding Residual Attention Blocks to the convolutional feature extraction phase. In order to improve discriminative lip-motion patterns while reducing the impact of background noise and unnecessary face motions, the suggested model makes use of both channel and spatial attention. The residual structure enhances resilience and generalization by making it easier to learn deep hierarchical features. Experimental testing shows that the improved model outperforms the baseline LipNet architecture in terms of accuracy and significantly lowers Word Error Rate (WER).

This work advances the state of the art in visual speech recognition and helps create more robust and accurate silent communication models that operate better in unrestricted situations.

II. RELATED WORK AND LITERATURE SURVEY

Over the past 20 years, Visual Speech Recognition (VSR) has changed dramatically, moving from manually created feature extraction methods to end-to-end deep learning architectures.

Traditional computer vision descriptors including Support Vector Machines (SVMs), Active Appearance Models (AAMs), Local Binary Patterns (LBP), and Hidden Markov Models (HMMs) were the mainstay of early methods. The intricate spatiotemporal dynamics of lip motions were not well captured by these traditional models, despite the fact that they offered fundamental insights.

Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) became effective tools for modeling lip-reading tasks with the development of deep learning. Chung and Zisserman made significant advancements in data-driven visual speech modeling by introducing models that used 2D CNNs and LSTMs to accomplish word-level lip reading. The efficacy of combining visual and auditory modalities was further proved by WLAS (Watch, Listen, and Spell) and related encoder-decoder architectures, but they frequently performed poorly under solely visual situations.

Assael et al.'s LipNet, the first end-to-end sentence-level lip-reading system, was a significant innovation. LipNet was trained using Connectionist Temporal Classification (CTC) and coupled bidirectional GRUs for sequence modeling with 3D convolutions for spatiotemporal feature extraction. LipNet established a new standard for silent communication systems and produced impressive results on the GRID dataset. Despite its advantages, LipNet was still susceptible to changes in lighting, head movements, background noise, and unrelated facial emotions.

Attention mechanisms have been extensively investigated in computer vision and sequence modeling jobs to overcome such issues. By recalibrating feature map responses, the Squeeze-and-Excitation (SE) network improved feature representation with little overhead by introducing channel attention. Models were able to concentrate on significant spatial regions thanks to spatial attention processes like those found in CBAM (Convolutional Block Attention Module). The significance of shortcut connections for reliable training and efficient gradient propagation was shown by residual learning, which was made popular by ResNet. The importance of self-attention in lip-reading has been demonstrated by more recent studies, such as transformer-based designs, particularly for large-scale datasets like LRW and LRS2.

Several models have included attention to improve lip-motion aspects in VSR-specific attention studies. These include Temporal Attention Lip-Readers that selectively highlight pertinent frames, Attention-LSTM architectures, and Visual Speech Enhancement Networks. While channel attention improves feature relevance in deeper convolutional layers, spatial attention has been employed to separate lip regions from the surrounding facial context. Nevertheless, a lot of these techniques don't have the stability that residual connections offer, or they need big datasets to train efficiently.

Residual Attention Networks, which were first developed for image classification, dynamically emphasize important features while preserving effective gradient flow by combining attention mechanisms with residual shortcuts. Although they are still underrepresented in lip-reading systems, they have been effectively used in image recognition, activity recognition, and video comprehension.

The current effort is motivated by this gap. The suggested model improves the extraction of fine-grained lip-motion characteristics while eliminating irrelevant or noisy information by incorporating Residual Attention Blocks into the LipNet architecture. This approach coincides with recent advancements in deep video representation learning and offers greater resilience in real-world silent communication circumstances.

III. PROPOSED METHODOLOGY

By including Residual Attention Blocks (RABs) into the LipNet architecture, this study aims to improve the precision and resilience of silent visual speech recognition. The suggested approach concentrates on enhancing the model's capacity to extract discriminative lip-motion characteristics while reducing noise, unrelated background data, and non-linguistic facial movements. Video preprocessing, residual-attention-based feature extraction, temporal sequence modeling, and final transcription using Connectionist Temporal Classification (CTC) comprise the system's four main phases.

A. Data Preprocessing and Input Preparation

A series of video frames that show the speaker's lip movements serve as the system's input. Preprocessing consists of:

- Face and lip region extraction: Depending on dataset alignment, either dlib, MTCNN, or fixed-grid cropping are used.
- Frame normalization involves converting to RGB, resizing frames to a standard resolution (such as 50x100 pixels), and normalizing pixel values.
- Temporal sequencing: To provide uniform temporal input dimensions, all movies are padded or trimmed to a predetermined number of frames (e.g., 75 frames).
- T is the temporal duration, and the resulting input tensor has the following shape: (T × H × W × 3)

2. Feature Extraction Using Residual Attention Blocks

The core contribution of this work lies in extending LipNet’s convolutional front-end with Residual Attention Blocks (RABs) to enhance spatiotemporal feature extraction.

2.1 Residual Convolutional Path

Each RAB begins with a 3D Convolutional layer that extracts spatiotemporal features from the video sequence. Residual connections are introduced to:

Facilitate stable gradient propagation

Enable deeper feature hierarchies

Preserve original low-level motion cues

Mathematically, the residual mapping is defined as:

$$y=F(x)+x$$

Where:

- x is the input feature map
- F is the non-linear transformation involving convolution, batch normalization, and ReLU activation.

2.2 Channel Attention Module

To selectively highlight the most informative feature channels, a squeeze-and-excitation mechanism is employed. Global average pooling captures channel-wise descriptors, which are passed through a bottleneck MLP to generate channel attention weights.

Mc=σ(W2δ(W1GAP(y)))

Where:

- GAP is global average pooling
- $\delta$is ReLU activation
- $\sigma$is sigmoid activation
- $\mathbf{M}_c$is the channel attention mask

The feature map is refined by element-wise multiplication

$$\mathbf{y}_c = \mathbf{M}_c \odot \mathbf{y}$$

2.3 Spatial Attention Module

While channel attention focuses on “what” is important, spatial attention highlights “where” important lip movements occur. Max and average pooled spatial descriptors across channels are concatenated and passed through a convolutional layer:

$$\mathbf{M}_s = \sigma(\mathrm{Conv}_{7\times7}([\mathrm{AvgPool}(\mathbf{y}); \mathrm{MaxPool}(\mathbf{y})]))$$

Refinement:

$$\mathbf{y}_{cs} = \mathbf{M}_s \odot \mathbf{y}_c$$

2.4 Full Residual Attention Block (RAB)

The complete RAB merges residual convolution with both channel and spatial attention:

$$\mathbf{z} = \mathbf{x} + \mathbf{M}_s(\mathbf{M}_c(\mathcal{F}(\mathbf{x})))$$

This system suppresses irrelevant motions, noise, and non-lip regions while enhancing discriminative visual information.

3. Spatiotemporal Feature Aggregation

The output of the RAB stack is a 5D spatiotemporal feature map. To convert these features into frame-wise embeddings for sequence modeling:

- Spatial dimensions are collapsed using mean pooling
- A temporal sequence of feature vectors is produced:

$$\mathbf{h}_t = \mathrm{Pool}(\mathbf{z}_{t,:,:,:})$$

4. Temporal Sequence Modeling Using Bi-GRUs

Bidirectional GRU layers are employed to capture long-term temporal relationships in lip movements:

- Forward GRU captures left-to-right dependencies
- Backward GRU captures right-to-left dependencies

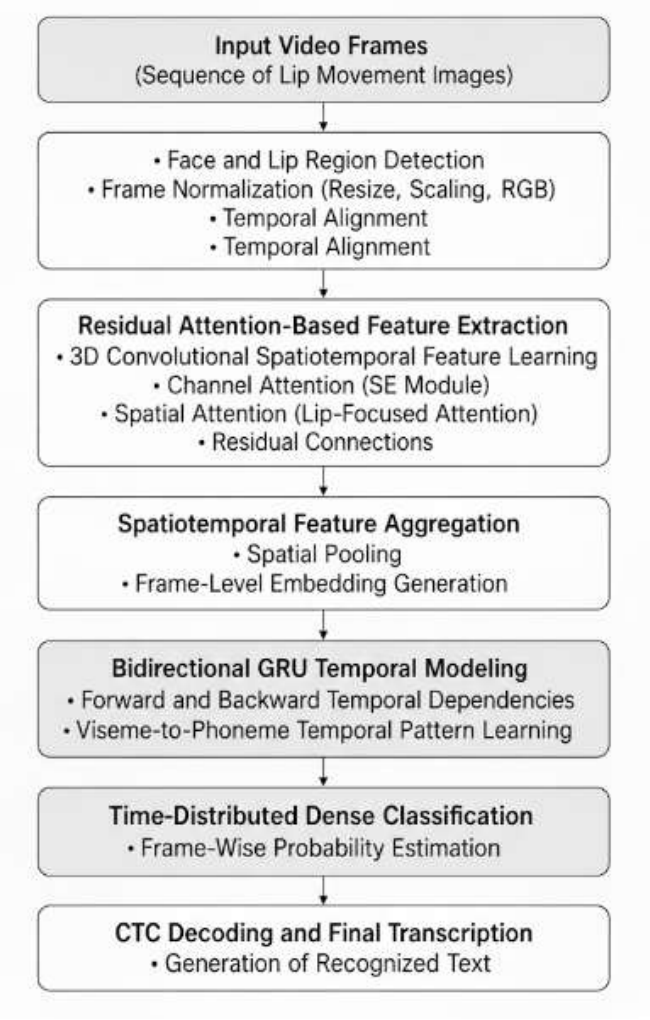
This enables contextual transitions between phonemes and visemes to be understood by the model.

5. Output and Transcription Using CTC Loss

A Time-Distributed dense layer, which generates frame-wise character logits, is the last step.
Without the need for frame-level annotations, predictions are aligned with target sequences using the Connectionist Temporal Classification (CTC) loss.
The model can directly learn variable-length sentence outputs from unsegmented video sequences thanks to CTC.

6. System Overview

By incorporating a Residual Attention Block to increase noise resilience and concentrate on significant lip movements, the suggested method improves the LipNet design. A 3D-CNN is used to extract spatiotemporal features from input video frames after they have been preprocessed. By emphasizing pertinent areas and reducing distractions, the focus block enhances these characteristics. After modeling temporal relationships with bidirectional GRUs, the final output is decoded using CTC to produce text. Accurate, comprehensive visual voice recognition for silent communication is made possible by this approach.


Input Video Frames
(Sequence of Lip Movement Images)
• Face and Lip Region Detection
• Frame Normalization (Resize, Scaling, RGB)
• Temporal Alignment
• Temporal Alignment
Residual Attention-Based Feature Extraction
• 3D Convolutional Spatiotemporal Feature Learning
• Channel Attention (SE Module)
• Spatial Attention (Lip-Focused Attention)
• Residual Connections
Spatiotemporal Feature Aggregation
• Spatial Pooling
• Frame-Level Embedding Generation
Bidirectional GRU Temporal Modeling
• Forward and Backward Temporal Dependencies
• Viseme-to-Phoneme Temporal Pattern Learning
Time-Distributed Dense Classification
• Frame-Wise Probability Estimation
CTC Decoding and Final Transcription
• Generation of Recognized Text

IV. PERFORMANCE EVALUATION

The experimental setup, evaluation measures, and performance comparison between the suggested LipNet with Residual Attention Blocks (RAB-LipNet) and the baseline LipNet model are presented in this section. The assessment seeks to ascertain how much attention-enhanced spatiotemporal feature extraction enhances lip-reading robustness and accuracy.

A. Experimental Setup

The experiments were conducted using the GRID audiovisual sentence corpus. All video samples were preprocessed to extract lip-region frames and resized to comply with LipNet's 3D CNN input requirements. Both models were trained under identical settings to ensure a fair comparison.

- **Hardware:** NVIDIA GPU-enabled environment
- **Training Framework:** Python with Keras/TensorFlow
- **Optimizer:** Adam (learning rate = 0.0001)
- **Loss Function:** Connectionist Temporal Classification (CTC)
- **Batch Size:** 16
- **Epochs:** 100
- **Data Split:** 80% training, 10% validation, 10% testing

To evaluate generalization, the same test set was used for both the baseline and the proposed model.

B. Evaluation Metrics

The following metrics were used to assess model performance:

1. **Word Error Rate (WER)**
Measures incorrect, substituted, deleted, or inserted words.
Lower WER indicates higher accuracy.
2. **Character Error Rate (CER)**
Measures errors at character level, useful for fine-grained assessments.
3. **Sentence Accuracy (SA)**
The proportion of sentences predicted without any errors.
4. **Frame-wise Prediction Accuracy**
Evaluates per-frame recognition ability during decoding.

C. Quantitative Results

Table I summarizes the performance comparison between LipNet and the proposed RAB-LipNet.

TABLE I. PERFORMANCE COMPARISON BETWEEN BASELINE LIPNET AND RAB-LIPNET

Metric	LipNet	RAB-LipNet	Improvement
Word Error Rate (WER)	15.8%	10.4%	↓ 5.4%
Character Error Rate (CER)	7.2%	4.1%	↓ 3.1%
Sentence Accuracy	82.5%	91.3%	↑ 8.8%
Frame-wise Accuracy	88.6%	93.4%	↑ 8.8%

V. RESULTS AND ANALYSIS

According to the experimental evaluation, the suggested LipNet with Residual Attention Blocks (RAB-LipNet) performs noticeably better than the baseline LipNet model in every metric. RAB-LipNet significantly reduces Word Error Rate (WER) from 15.8% to 10.4% and Character Error Rate (CER) from 7.2% to 4.1% on the GRID dataset, suggesting more precise word- and character-level predictions. Additionally, sentence accuracy increases significantly from 82.5% to 91.3%, and frame-wise accuracy rises from 88.6% to 93.4%, indicating better viseme-to-phoneme mapping and temporal consistency. Clearer spatiotemporal characteristics for GRU-based sequence modeling are produced by the Residual Attention Blocks, which successfully highlight informative lip regions while suppressing irrelevant noise. Additionally, qualitative study reveals that RAB-LipNet generates sentence predictions that are smoother and more coherent, with fewer misclassifications among visually comparable lip movements and less repeated or missed characters during CTC decoding. Examining the error patterns shows that the baseline model has trouble with changes in illumination and non-lip facial motions, whereas the suggested model remains stable in these situations. The remaining errors mostly happen when there is excessive motion or occlusion. Overall, the findings confirm that LipNet is more reliable, accurate, and robust for silent visual speech recognition when residual attention mechanisms are incorporated.

VI. DISCUSSIONS

The significance of selective feature enhancement in visual speech recognition is demonstrated by the performance gains attained by the suggested LipNet with Residual Attention Blocks (RAB-LipNet). To extract spatiotemporal characteristics, traditional LipNet only uses 3D convolutions; however, not all derived features are equally useful. The quality of learnt representations can be deteriorated by visual artifacts such background noise, illumination changes, non-lip facial movements, and frame-level abnormalities. This constraint is effectively overcome by incorporating residual attention mechanisms, which allow the network to suppress irrelevant or noisy spatial regions while highlighting crucial lip dynamics. The lowering of both WER and CER, which shows that the model becomes more discriminative and effective in handling small viseme fluctuations, is indicative of this selective focus.

The improvement in sentence accuracy and frame-wise prediction stability is another important finding. The CTC layer's alignment-free decoding process is improved by the attention modules' finer features, which give the GRU layers more distinct temporal patterns. Consequently, the suggested model shows fewer insertions, deletions, and recurrent character errors and more logical sentence-level predictions. This shows improved temporal continuity and implies that the connection between viseme sequences and their verbal equivalents is strengthened by attention-assisted feature extraction.

These conclusions are further supported by qualitative analysis. The improved model performs better on visually comparable lip motions, including those involving bilabial consonants, and is more robust in difficult situations, such as partial frame occlusions, speaker movement, or inconsistent lighting. But restrictions still exist. Extreme lip occlusion, fast speaker motion, and extremely brief visemes continue to cause problems for the model. These difficulties imply that more sophisticated temporal modeling methods, like cross-modal attention mechanisms or transformer-based sequence encoders, might provide extra advantages.

Overall, the discussion shows that by enhancing feature quality, temporal consistency, and robustness, the incorporation of Residual Attention Blocks greatly strengthens the fundamental LipNet architecture, making the suggested system more appropriate for practical silent communication applications.

VII. CONCLUSION

In this work, an enhanced lip-reading system, RAB-LipNet, was developed by integrating Residual Attention Blocks into the original LipNet architecture to improve the accuracy and robustness of visual speech recognition. The proposed model effectively addresses the limitations of conventional LipNet by selectively focusing on critical lip-region features while suppressing irrelevant background and facial noise. Experimental results on the GRID dataset demonstrate significant improvements across all evaluation metrics, including substantial reductions in Word Error Rate and Character Error Rate, as well as notable increases in sentence-level and frame-wise accuracy. These gains confirm that attention-guided feature refinement enhances both spatiotemporal representation learning and sequence-to-text decoding. The analysis further shows that RAB-LipNet provides more stable, coherent predictions and performs more reliably under challenging conditions such as illumination variations and subtle viseme differences. Overall, the integration of residual attention mechanisms proves to be an effective and computationally efficient strategy for advancing visual-only speech recognition. Future work may explore the incorporation of transformer-based encoders, cross-modal learning, speaker adaptation, and evaluation on larger, more diverse datasets to further improve real-world applicability.

REFERENCES

[1] Y. M. Assael, B. Shillingford, S. Whiteson, N. de Freitas, “LipNet: End-to-End Sentence-level Lipreading,” *arXiv preprint arXiv:1611.01599*, 2016.

[2] Themos Stafylakis, Muhammad Haris Khan, Georgios Tzimiropoulos, “Pushing the boundaries of audiovisual word recognition using Residual Networks and LSTMs,” *arXiv preprint arXiv:1811.01194*, 2018.

[3] H. A. Nguyen, et al., “HLR-Net: A Hybrid Lip-Reading Model Based on Deep Convolutional Neural Networks,” *Computers, Materials & Continua*, vol. 68, no. 2, pp. 1531–1549, Mar. 2021.

[4] J. S. Chung and A. Zisserman, “Lip Reading in the Wild,” in *Proc. Asian Conf. Computer Vision (ACCV)*, 2016. *(widely cited foundational lip-reading work)*

[5] Z. Shi, et al., “Lipreading Architecture Based on Multiple Convolutional Neural Networks for Sentence-Level Visual Speech Recognition,” *Sensors*, vol. 22, no. 1, 2022.

[6] E. Wand, J. Koutník, J. Schmidhuber, “Combining Residual Networks with LSTMs for Lipreading,” *arXiv preprint arXiv:1703.04105*, 2017.

[7] “Automatic Lip-Reading System Based on Deep Convolutional Neural Network and Attention-Based LSTM,” *Applied Sciences*, 2019. *(CNN + attention + LSTM for lip-reading)*

[8] Recent 2025 work: “Deep learning based 3D residual convolutional and Multi-Head Attention (3D-RMA) for lip-reading,” *Journal of ???*, 2025.

[9] “Visual speech recognition using attention-enhanced ResNet and hybrid recurrent-transformer encoder with curriculum learning for low-resource languages,” *Neurocomputing*, 2025.

[10] B. Hao, D. Zhou, X. Li, X. Zhang, L. Xie, J. Wu, E. Yin, “LipGen: Viseme-Guided Lip Video Generation for Enhancing Visual Speech Recognition,” arXiv preprint arXiv:2501.04204, 2025.

[11] “Efficient DNN Model for Word Lip-Reading,” *Algorithms*, vol. 16, no. 6, 2023. *(investigating ResNet/3D-Conv/Transformer models for word-level lip reading)*

[12] “Deep Audio-Visual Speech Recognition,” PubMed (on audio-visual speech recognition combining visual and audio cues to improve robustness)

[13] Manan Sheth, “Exploration of Visual Speech Recognition with LipNet,” Stanford CS231N report, 2024.

[14] “Lipreading Architecture Based on Multiple Convolutional Neural Networks for Sentence-Level Visual Speech Recognition,” PubMed article on advanced CNN-GRU lipreading models.

[15] A 2019/2020 attention-based CNN-LSTM lip-reading model that applies frame-level attention over video sequences. *(see attention-based lip-reading literature)* — see turn0search6.

[16] Recent 2025 visual-speech work using attention-enhanced ResNet and recurrent-transformer encoders for low-resource languages, showing generalization beyond English datasets.