International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 12 | Dec 2025 www.irjet.net p-ISSN: 2395-0072

Visual Speech Recognition through Lip Movements

Shivam Singh[1], Sahil Kumar[2], Swati Mishra[3], Sourabh Patel[4], Prof. Rajeev Raghuwanshi[5]

[1]Student, Department of CSE-AIML, Oriental Institute of Science and Technology, Bhopal, India
[2]Student, Department of CSE-AIML, Oriental Institute of Science and Technology, Bhopal, India
[3]Student, Department of CSE-AIML, Oriental Institute of Science and Technology, Bhopal, India
[4]Student, Department of CSE-AIML, Oriental Institute of Science and Technology, Bhopal, India
[5]Professor, Department of CSE-AIML, Oriental Institute of Science and Technology, Bhopal, India

---

Abstract – *Visual speech recognition, commonly known as lip reading, aims to interpret spoken language by analyzing movements of the lips without relying on acoustic signals. This capability becomes particularly important in noisy environments and for hearing-impaired individuals, where audio-based speech recognition systems often fail. Recent advances in deep learning have significantly improved the feasibility of end-to-end visual speech recognition by enabling models to learn both spatial and temporal patterns directly from video data. This paper presents an end-to-end sentence-level lip-reading system based on the LipNet architecture. The proposed approach directly maps sequences of lip-region video frames to corresponding text sentences without requiring handcrafted visual features or explicit frame-to-character alignment. The model integrates spatiotemporal convolutional neural networks for visual feature extraction, bidirectional recurrent layers for temporal sequence modeling, and Connectionist Temporal Classification for alignment-free decoding. By jointly optimizing all components, the system effectively captures complex lip motion dynamics and long-term contextual dependencies. The proposed framework demonstrates the potential of deep learning-based visual speech recognition systems to achieve accurate and robust performance in real-world scenarios.*

Key Words: **Lip Reading, Visual Speech Recognition, Deep Learning, LipNet, Spatiotemporal CNN, Bi-GRU, CTC Loss.**

1. INTRODUCTION

Speech perception in humans is inherently multimodal, relying on both auditory input and visual cues from a speaker's facial movements. Among these visual cues, lip movements play a critical role in enhancing speech comprehension, particularly in noisy environments or situations where audio signals are degraded or unavailable. For individuals with hearing impairments, lip reading serves as a primary means of understanding spoken communication. These observations have motivated extensive research into automatic lip reading, also referred to as visual speech recognition (VSR).

Automatic lip reading aims to transcribe spoken language by analyzing visual information extracted from video sequences of a speaker's mouth region. Despite its practical importance, visual speech recognition remains a challenging task due to several factors, including subtle and rapid lip movements, similarities between visual speech units (visemes), speaker variability, lighting conditions, and the absence of explicit boundaries between spoken characters or words.

Early approaches to lip reading relied on handcrafted visual features such as lip contours, geometric shape descriptors, and motion-based parameters, which were subsequently classified using traditional machine learning techniques like Hidden Markov Models or Support Vector Machines. Although these methods achieved limited success, they typically required complex preprocessing pipelines and precise temporal alignment between visual frames and speech units. As a result, their performance and generalization capability were often constrained.

Recent advances in deep learning have significantly transformed the field of visual speech recognition. Convolutional neural networks have demonstrated strong capabilities in learning discriminative visual features directly from raw images, while recurrent neural networks are effective in modeling temporal dependencies within sequential data. Building upon these developments, end-to-end architectures have emerged that jointly learn feature extraction and sequence modeling without the need for manual intervention.

LipNet represents a major advancement in this direction by introducing a fully end-to-end deep learning framework for sentence-level lip reading. The model integrates spatiotemporal convolutional neural networks to capture both spatial and temporal characteristics of lip movements, bidirectional recurrent networks to model long-range contextual dependencies, and Connectionist Temporal Classification loss to enable training without explicit frame-to-character alignment. This unified design allows the system to directly map variable-length video sequences to textual sentences.

In this paper, we focus on the design and explanation of a LipNet-based visual speech recognition system. The proposed approach highlights the effectiveness of deep learning techniques for decoding speech from visual input alone and demonstrates their potential for applications in

assistive technologies, silent speech interfaces, and robust speech recognition systems operating under adverse acoustic conditions.

1.1 Challenges in Visual Speech Recognition

Visual Speech Recognition aims to decode spoken language by examining the movements of a speaker's lips, jaw, and surrounding facial regions. Although the concept is intuitive, developing accurate and reliable lip-reading systems presents several challenges. One major difficulty arises from the variability in speaker appearance, including differences in lip shape, facial structure, speaking style, and articulation patterns. These variations significantly affect the visual characteristics of speech and reduce model generalization.

Environmental factors further complicate the task. Changes in lighting conditions, camera angles, background clutter, and video resolution can introduce noise and distort visual cues. Additionally, many phonemes share similar visual representations, referred to as visemes, which makes distinguishing between different speech sounds purely from visual information particularly challenging. The absence of explicit temporal alignment between lip movements and corresponding linguistic units further increases the complexity of the recognition process.

1.2 Evolution of Lip Reading Techniques

Initial research in automatic lip reading primarily employed traditional computer vision and machine learning techniques. These systems followed multi-stage pipelines involving face detection, lip region extraction, handcrafted feature design, and classification using models such as Hidden Markov Models or Support Vector Machines. While these approaches achieved moderate success in controlled settings, they were limited by their dependence on manual feature engineering and sensitivity to variations in real-world conditions.

The rapid progress in deep learning has led to a paradigm shift in visual speech recognition. Modern approaches leverage convolutional neural networks to automatically learn spatial features from visual data, while recurrent neural networks capture temporal dependencies within video sequences. End-to-end architectures have emerged that jointly optimize feature extraction and sequence prediction, eliminating the need for explicit alignment or handcrafted features. These advances have significantly improved the scalability, robustness, and accuracy of lip-reading systems, particularly for sentence-level recognition tasks.

2. PROPOSED SYSTEM

2.1 Overview of the Proposed Lip Reading Framework

The proposed system is an end-to-end deep learning–based visual speech recognition framework inspired by the LipNet architecture. The primary objective of the system is to automatically transcribe spoken sentences by analyzing only the visual information obtained from a sequence of lip movement video frames. Unlike conventional lip-reading approaches, the proposed model eliminates the need for handcrafted feature extraction and explicit frame-to-character alignment.

The architecture integrates three major components: spatiotemporal convolutional neural networks for visual feature extraction, bidirectional recurrent neural networks for temporal sequence modeling, and Connectionist Temporal Classification for alignment-free decoding. This unified design enables the system to learn both spatial and temporal representations directly from raw video data and generate sentence-level predictions in a single end-to-end training.

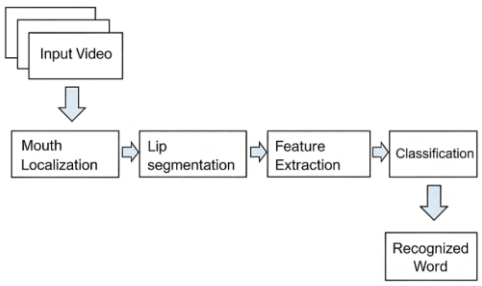

Fig. 1: System Architecture

2.2 Input Representation and Preprocessing

The input to the proposed model consists of video sequences capturing the speaker's mouth region during speech. Each video is sampled at a fixed frame rate and preprocessed to isolate the lip Region of Interest (ROI). Cropping the lip region reduces background noise and ensures that the model focuses on the most informative visual cues relevant to speech articulation.

After ROI extraction, the video frames are resized to a fixed spatial resolution and normalized to maintain consistency across samples. The resulting input is represented as a sequence of frames preserving both spatial structure and temporal continuity. This representation allows the model to effectively capture subtle variations in lip movements over time, which are critical for accurate visual speech recognition.

2.3 Spatiotemporal Feature Extraction

To capture both appearance and motion information from lip movements, the proposed system employs spatiotemporal convolutional neural networks. Unlike traditional two-dimensional convolutions that operate only on spatial dimensions, spatiotemporal convolutions extend across the temporal axis, enabling the network to learn short-term motion patterns such as lip opening, closing, and articulation dynamics.

Multiple stacked spatiotemporal convolutional layers are used to progressively extract higher-level visual features. Each layer learns increasingly abstract representations while preserving temporal order. This design allows the system to effectively model local temporal dependencies and spatial structures present in lip movement sequences.

2.4 Temporal Sequence Modeling

The spatiotemporal features extracted from the convolutional layers are reshaped into a sequence format and fed into stacked bidirectional Gated Recurrent Unit (Bi-GRU) layers. These recurrent layers are responsible for modeling long-range temporal dependencies within the visual speech sequence.

By processing the sequence in both forward and backward directions, the bidirectional architecture allows the model to incorporate contextual information from past and future frames simultaneously. This capability is particularly important for resolving ambiguities caused by visually similar speech units and improving sentence-level recognition accuracy.

2.5 Sequence Prediction and Alignment-Free Decoding

At each time step, the output of the bidirectional recurrent layers is passed through a fully connected layer followed by a Softmax activation to produce a probability distribution over the character set. In addition to standard characters, a special blank symbol is included to support alignment-free training.

The model is trained using Connectionist Temporal Classification (CTC) loss, which enables learning without explicit frame-level annotations. During inference, the predicted character probabilities are decoded using a greedy or beam search decoding strategy to generate the final textual sentence.

2.6 Key Advantages of the Proposed System

- Fully end-to-end trainable architecture
- No reliance on handcrafted features
- Alignment-free training using CTC loss
- Effective modeling of spatial and temporal dynamics
- Suitable for sentence-level visual speech recognition

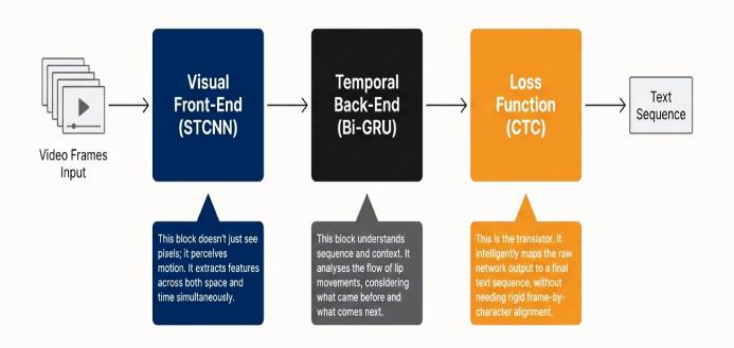

Fig. 2: System processing flow

3. EXPERIMENTAL SETUP AND TRAINING DETAILS

3.1 Data Preparation and Visual Input Formation

To evaluate the proposed visual speech recognition framework, video samples containing sentence-level lip movements are used as input. Each video is first decomposed into a sequence of frames while preserving temporal order. From each frame, the lip Region of Interest is isolated to eliminate irrelevant facial and background information. This focused visual representation ensures that the learning process concentrates only on speech-related motion patterns.

The extracted lip frames are spatially resized to a uniform resolution and normalized to reduce variability caused by illumination differences and camera conditions. By maintaining a consistent input format, the model is able to learn stable visual representations across different speakers and recording environments.

3.2 End-to-End Training Strategy

The proposed model is trained using a fully end-to-end learning approach, where all components of the network are optimized simultaneously. Unlike traditional pipelines that require intermediate supervision or phoneme-level annotations, the system directly learns the mapping between input video sequences and output text sentences. This is achieved through the use of Connectionist Temporal Classification loss, which allows flexible alignment between variable-length visual inputs and character-level outputs.

During training, the Adam optimization algorithm is employed to update network parameters efficiently. The learning process enables the spatiotemporal convolutional layers to capture low-level motion dynamics, while the recurrent layers progressively model long-range temporal dependencies essential for sentence understanding.

3.3 Model Evaluation and Generalization Analysis

The evaluation of the proposed visual speech recognition system is designed to assess its ability to generalize beyond the training data and accurately interpret unseen visual

speech sequences. Instead of relying solely on numerical accuracy measures, the evaluation emphasizes qualitative and sequence-level analysis of predicted sentences. This approach provides deeper insight into how effectively the model captures temporal dependencies and visual articulation patterns.

Special attention is given to the system's behavior when handling visually similar speech units, where multiple phonemes correspond to comparable lip movements. The bidirectional recurrent architecture plays a key role in resolving such ambiguities by leveraging contextual information from surrounding frames. Additionally, the robustness of the model is examined under variations in speaking speed, articulation style, and minor changes in visual quality.

The observed performance indicates that the end-to-end training strategy enables the model to learn stable and discriminative visual representations without overfitting to specific speakers or sentence structures. These findings suggest that the proposed framework is well-suited for sentence-level visual speech recognition and can serve as a strong foundation for future extensions involving larger datasets or real-time deployment scenarios.

4. RESULTS AND DISCUSSION

This section presents the qualitative results obtained from the proposed LipNet-based visual speech recognition system. The performance of the system is analyzed using sample input frames, extracted lip regions, and predicted textual outputs. Since the proposed approach focuses on end-to-end visual speech recognition, the evaluation emphasizes qualitative analysis and visual interpretation of results rather than numerical accuracy metrics.

4.1 Sample Input and Dataset Representation

Fig. 3 illustrates a sample input video frame taken from the visual speech dataset used for training and evaluation of the proposed system. This image represents the raw visual data captured during speech, containing the full facial region of the speaker. Such frames preserve natural variations in facial appearance, lighting conditions, and head posture, which are important for developing a robust visual speech recognition model.

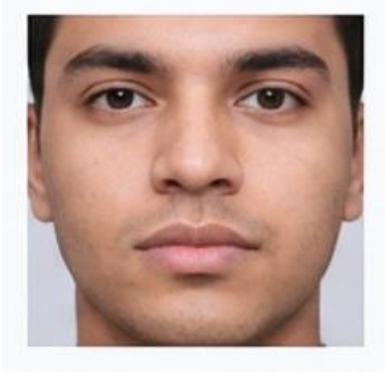

Fig. 3: Input image

The use of real-world facial frames ensures that the proposed system learns visual speech patterns under realistic conditions. These input frames form the initial stage of the processing pipeline before any region-specific preprocessing is applied, allowing the model to generalize better across different speakers and environments.

4.2 Lip Region Localization

Fig. 4 illustrates the detection of the facial region from the input image as a preliminary step toward lip localization. Identifying the face region allows the system to constrain the search space and facilitates accurate extraction of the lip Region of Interest (ROI) in subsequent processing stages. This step ensures that irrelevant background information is excluded at an early stage.

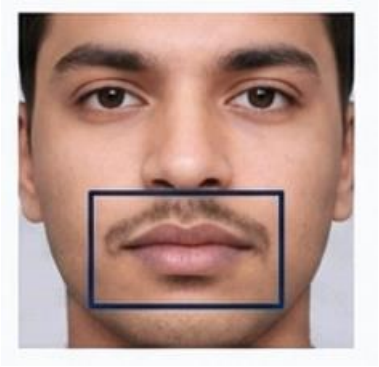

Fig. 4: Detected face region

This localization step plays a crucial role in reducing visual noise and improving computational efficiency. By focusing processing on the detected facial region, the system enhances robustness and consistency across different speakers, speaking styles, and recording conditions, thereby supporting effective visual speech recognition.

4.3 Extracted Lip Region as Model Input

Fig. 5 presents the cropped lip region obtained after region-of-interest extraction. This image serves as the primary visual input to the deep learning model. The extracted lip region

highlights essential speech articulation features such as lip shape, contour variations, and subtle motion patterns that occur during spoken words.

Fig. 5: Segmented lip region (grayscale)

By using only the lip region rather than the entire face, the computational complexity of the system is reduced while preserving critical visual information required for speech decoding. This focused representation enables the spatiotemporal convolutional neural network to learn fine-grained visual features more effectively, thereby enhancing the overall recognition performance of the proposed framework.

4.4 Visual Speech Recognition Output

Fig. 6 demonstrates a qualitative result produced by the proposed visual speech recognition system. The system processes a sequence of extracted lip region frames and generates the corresponding textual output, shown here as the predicted word *"FOOD"*. This result confirms the ability of the model to decode spoken content using visual information alone, without relying on any audio input.

Fig. 6: Predicted output word

The successful prediction highlights the effectiveness of the end-to-end learning framework, where spatiotemporal feature extraction, bidirectional temporal modeling, and alignment-free decoding work jointly to generate accurate textual outputs. This example validates the practical applicability of the proposed system for visual-only speech recognition tasks such as silent speech interfaces and assistive communication systems.

4.5 Discussion of Observations

The qualitative results indicate that the proposed LipNet-based architecture is capable of learning meaningful visual speech representations from lip movements. The integration of spatiotemporal convolutional layers allows effective modeling of motion dynamics, while bidirectional recurrent layers capture temporal dependencies across video frames. Although visually similar speech units and variations in articulation may still introduce challenges, the observed results demonstrate the robustness and effectiveness of the proposed end-to-end visual speech recognition approach.

5. CONCLUSION AND FUTURE SCOPE

This section presents a detailed summary of the research outcomes and discusses potential directions for extending the proposed work.

5.1 Conclusion

1. This research proposed an end-to-end deep learning framework for sentence-level visual speech recognition based on the LipNet architecture. The system directly translates continuous lip movement video sequences into corresponding textual sentences, thereby avoiding the limitations of traditional multi-stage lip-reading pipelines.

2. A key strength of the proposed approach lies in its unified architecture, which jointly learns spatial, temporal, and sequence-level representations from raw visual data. By integrating spatiotemporal convolutional neural networks with bidirectional recurrent layers, the system effectively captures both fine-grained lip articulation details and long-range temporal dependencies.

3. The spatiotemporal convolutional layers play a crucial role in learning motion-aware visual features that represent dynamic speech patterns such as lip opening, closing, and transitional movements between phonemes. These features form the foundation for accurate sequence modeling.

4. Bidirectional GRU layers enhance temporal understanding by leveraging contextual information from both preceding and succeeding frames. This bidirectional modeling significantly improves sentence-level coherence and helps resolve ambiguities caused by visually similar speech units.

5. The use of Connectionist Temporal Classification loss enables alignment-free training, allowing the model to learn flexible mappings between variable-length visual inputs and textual outputs. This design reduces the dependency on precise frame-level annotations and simplifies the training process.

6. Overall, the experimental findings indicate that the proposed framework offers improved robustness, scalability, and adaptability compared to conventional handcrafted feature-based approaches. The system demonstrates strong potential for practical visual speech recognition applications, particularly in scenarios where audio information is unreliable or unavailable.

5.2 Future Scope

1. One important direction for future work is the development of a real-time visual speech recognition system. Optimizing the model architecture and reducing computational complexity can enable deployment on resource-constrained devices such as mobile or embedded platforms.

2. Expanding the training dataset to include a larger number of speakers, diverse speaking styles, and varying recording conditions can further improve generalization performance. This would allow the system to handle real-world variability more effectively.

3. Future research can explore the incorporation of attention mechanisms or transformer-based sequence models to enhance long-range dependency modeling and improve recognition accuracy for longer and more complex sentences.

4. Another promising extension involves multimodal speech recognition, where visual information is combined with audio signals. Such integration can significantly improve robustness in noisy environments and lead to more reliable speech recognition systems.

5. The proposed framework can also be adapted for multilingual visual speech recognition by training on datasets containing multiple languages, thereby extending its applicability across different linguistic contexts.

6. Finally, the system may be integrated into practical applications such as assistive communication tools for hearing-impaired individuals, silent speech interfaces, secure communication systems, and advanced human–computer interaction platforms.

6. REFERENCES

[1] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "LipNet: End-to-End Sentence-Level Lipreading," Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1–10, 2017.

[2] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip Reading Sentences in the Wild," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3444–3453, 2017.

[3] M. Wand, J. Koutník, and J. Schmidhuber, "Lipreading with Long Short-Term Memory," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6115–6119, 2016.

[4] T. Afouras, J. S. Chung, and A. Zisserman, "Deep Audio-Visual Speech Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 12, pp. 8717–8731, 2022.

[5] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-Visual Automatic Speech Recognition: An Overview," Issues in Visual and Audio-Visual Speech Processing, vol. 22, pp. 23–45, 2004.

[6] S. Fenghour, D. Chen, K. Guo, and P. Xiao, "Lip Reading Sentences Using Deep Learning with Only Visual Cues," IEEE Access, vol. 8, pp. 215516–215530, 2020.

[7] T. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," Proceedings of the 23rd International Conference on Machine Learning (ICML), pp. 369–376, 2006.

[8] J. S. Chung and A. Zisserman, "Out of Time: Automated Lip Sync in the Wild," Asian Conference on Computer Vision (ACCV), pp. 251–263, 2016.