IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

LIP READING MODEL – CONVERT SILENT VIDEO TO TEXT

[1]Mrs. Pradnya Kasture, [2]Vishwa Patil, [3]Titiksha Chaudhari, [4]Snehal Chougule

[1]Professor, [2]Student, [3]Student, [4]Student

[1]Computer Engineering Department,

[1]RMD Sinhgad Technical Institute Campus, Pune, Maharashtra, India.

Abstract: Visual Speech Recognition or lip reading is a growing area of research in artificial intelligence and computer vision. The main goal of lip-reading systems is to recognize spoken words by looking at lip movements without using sound. These systems are really helpful in places, surveillance systems, generating subtitles for videos and technology for people who are hard of hearing. This paper presents a lip-reading model that uses learning to turn silent videos into text. The model works by finding the mouth in a video then using a special architecture that combines three types of networks Three-Dimensional Convolutional Neural Networks or 3D-CNN, which look at the video and find patterns in space and time. EfficientNetB0, which helps find the important features in the video. Bidirectional Long Short-Term Memory or Bi-LSTM, networks, which look at how lip movements change over time. The model uses something called Connectionist Temporal Classification loss to predict sequences of words without needing to line up each frame of video with a word. Tests show that the model can recognize speech from videos and produce accurate text. The model can be used in areas, such, as Helping people communicate, Security monitoring, Human-computer interaction, Making multimedia more accessible.

Index Terms - Artificial Intelligence, Visual Speech Recognition, Deep Learning Lip Reading, 3D-CNN, EfficientNetB0, Bi-LSTM, Computer Vision, Silent Video Processing.

I. INTRODUCTION

Communication plays a vital role in everyday life, and speech is one of the most effective ways through which people express their thoughts and ideas. In recent years, Automatic Speech Recognition (ASR) systems have become increasingly popular due to their ability to convert spoken language into text. These systems are widely used in virtual assistants, customer service applications, and voice-controlled devices. However, traditional speech recognition systems mainly depend on audio signals, which limits their performance in noisy environments or situations where audio information is unavailable. To overcome this limitation, researchers have explored Visual Speech Recognition (VSR), also known as lip reading, which focuses on recognizing speech by analyzing lip movements and facial expressions. Recent advances in deep learning and computer vision have significantly improved the ability of machines to understand visual speech, making lip-reading systems a promising solution for silent communication and speech recognition without audio input.

To address this limitation, researchers have explored Visual Speech Recognition (VSR), commonly known as lip reading. Lip reading is the process of understanding speech by observing the movement of a speaker's lips, mouth, and facial expressions. Humans naturally use visual cues to improve speech understanding, especially in noisy surroundings. Inspired by this ability, modern computer vision and deep learning techniques have been developed to enable machines to recognize speech directly from visual information.

The growing demand for intelligent communication systems has increased interest in lip-reading technology. Such systems can be extremely useful for hearing-impaired individuals, silent communication applications, surveillance systems, video caption generation, and human-computer interaction. In many real-world situations, speech may need to be recognized even when audio recordings are unclear or completely absent. Therefore, developing an efficient lip-reading system has become an important research challenge. Recent advancements in Deep Learning have significantly improved the performance of visual speech recognition systems. Convolutional Neural Networks (CNNs) have proven highly effective in extracting important visual features from images and videos, while Long Short-Term Memory (LSTM) networks are capable of learning sequential patterns over time. By combining these technologies, it becomes possible to accurately analyze lip movements and convert them into meaningful text.

This research proposes a LipReading Model that converts silent videos into text using advanced deep learning techniques. The system first extracts frames from an input video and identifies the mouth region using facial landmark detection. The extracted mouth sequences are then processed through a 3D Convolutional Neural Network (3D-CNN) to capture spatial and temporal information. EfficientNetB0 is incorporated to improve feature extraction efficiency, while a Bidirectional Long Short-Term Memory (Bi-LSTM) network is used to understand the sequence of lip movements. Finally, Connectionist Temporal Classification (CTC) decoding generates the corresponding text output.

The proposed model aims to provide an accurate and efficient solution for silent speech recognition. Unlike conventional speech recognition systems, it does not depend on audio signals, making it suitable for challenging environments where sound quality is poor. Furthermore, the system can contribute to improving accessibility and communication for individuals with hearing difficulties.

II. RESEARCH METHODOLOGY

The LipSyncNet system follows an end-to-end deep learning pipeline for visual speech recognition from silent video input. The process begins with video acquisition, where silent video frames are extracted using OpenCV to capture temporal variations in lip movement. Face and lip regions are then detected using dlib facial landmark to isolate the region of interest for accurate analysis. The extracted frames are preprocessed through cropping, resizing, and normalization to standardize input dimensions and enhance model consistency. Feature extraction is carried out using a 3D Convolutional Neural Network (3D-CNN) combined with EfficientNetB0, which captures both spatial and temporal features from video frames. These extracted features are passed into a Bidirectional Long Short-Term Memory (Bi-LSTM) or Transformer network for sequence modeling, allowing the system to learn contextual and temporal dependencies between consecutive frames. Finally, decoding is performed using Connectionist Temporal Classification (CTC) or Attention mechanisms to translate the learned visual sequences into text. The final output is the generated text representing spoken words derived solely from visual input.

2.1 Data Acquisition

The LipSyncNet model uses the GRID Corpus dataset**,** a well-known benchmark for audio-visual speech recognition.The dataset consists of videos from 34 speakers, each uttering 1,000 structured six-word sentences. These videos provide synchronized audio and visual data, though only the visual component is used for lip-reading. The dataset ensures diversity in speaker characteristics, lighting conditions, and mouth movements, enabling the model to learn robust visual features. Silent video input is utilized to capture lip movements corresponding to spoken words, forming the foundation for visual speech interpretation.

2.2 Frame Extraction

Each video from the dataset is divided into individual frames using OpenCV. This frame extraction process converts continuous video sequences into a set of discrete image frames, typically capturing 25 to 30 frames per second. The frame-by-frame representation allows the model to analyze minute temporal variations in lip motion. This sequential input serves as the basis for understanding speech dynamics, enabling the model to detect subtle transitions between phonemes and words.

2.3 Lip Region Detection

To focus exclusively on relevant visual information, the Region of Interest (ROI) is localized around the mouth area. Facial landmark detection is performed using the Dlib library, which identifies key facial points such as the mouth, eyes, and jawline. Based on these landmarks, the mouth region is cropped from each frame, removing unnecessary background and non-verbal facial features. This ensures that the model's attention is concentrated solely on lip movement patterns, which are critical for accurate speech recognition.

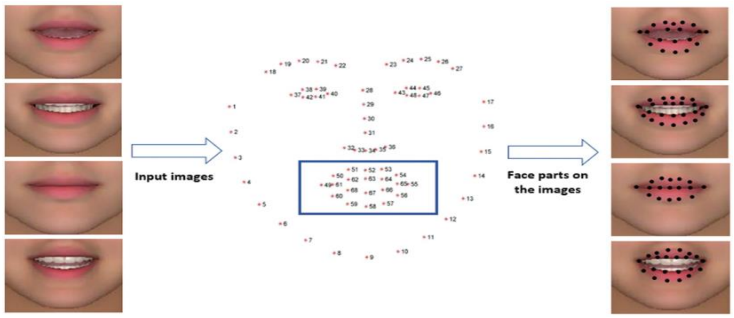


Figure 1: Lip Region Detection

2.4 Preprocessing

The cropped lip regions undergo a series of preprocessing steps to ensure data uniformity and reduce computational complexity. Each frame is converted to grayscale to eliminate color dependency and emphasize motion intensity. The frames are then resized to 140 × 46 pixels and normalized to a consistent scale. Preprocessing also includes data augmentation techniques, such as rotation, brightness adjustment, and flipping, to simulate variations in lighting, facial orientation, and head movement. This step enhances the model's ability to generalize across different speakers and recording conditions.

2.5 Feature Extraction

In this stage, visual features are extracted from the preprocessed video frames using a 3D Convolutional Neural Network (3D-CNN) integrated with EfficientNetB0. The 3D-CNN captures both spatial and temporal dependencies by applying convolutions across multiple consecutive frames, enabling the model to learn motion dynamics in lip movements. EfficientNetB0, known for its balanced scaling of depth, width, and resolution, enhances the feature extraction process by producing high-quality representations with optimized computational efficiency. This combination ensures that the model captures fine-grained details in the lip motion patterns while maintaining lightweight architecture.

2.6 Sequence Modelling

The extracted feature maps are passed into a Bidirectional Long Short-Term Memory (Bi-LSTM) network. The Bi-LSTM processes the features in both forward and backward temporal directions, effectively learning contextual relationships between preceding and succeeding frames. This step allows the model to recognize the sequential nature of speech and handle coarticulation effects, where adjacent phonemes influence each other's visual representation. The bidirectional architecture enables the system to understand complete sentence-level patterns, improving prediction consistency across varying lip movement speeds and styles.

2.7 Decoding and Classification

The sequential features obtained from the Bi-LSTM are mapped to textual representations using the Connectionist Temporal Classification (CTC) loss function. CTC is designed for sequence-to-sequence tasks where direct alignment between input and output is unknown. It enables the model to predict variable-length sequences and automatically align them to corresponding characters or words without manual segmentation.

This end-to-end training approach eliminates the need for frame-level labeling and significantly improves model efficiency and adaptability.

2.8 Output Generation

The final output of the system is textual transcription, representing the recognized words or sentences derived purely from visual cues. The predicted text is decoded from the CTC layer and displayed through a Streamlit-based interface, which provides a user-friendly platform for real-time testing and visualization. Users can upload silent videos, and the model will output the corresponding text, demonstrating its capability to perform speech recognition without any audio and input.

2.9 Model Training

The LipSyncNet model is trained using the **Adam optimizer** with a learning rate of **0.0001** for **50 epochs**. The training process minimizes CTC loss while optimizing model weights through backpropagation. During evaluation, metrics such as **accuracy** and **Word Error Rate (WER)** are used to assess performance. The model achieved an impressive **96.7% accuracy** and a **WER of 8.2%** on the GRID Corpus dataset, outperforming earlier lip-reading systems such as LipNet and LCANet. The results confirm the model's robustness, efficiency, and ability to generalize across different speakers and conditions.

2.10 System Architecture

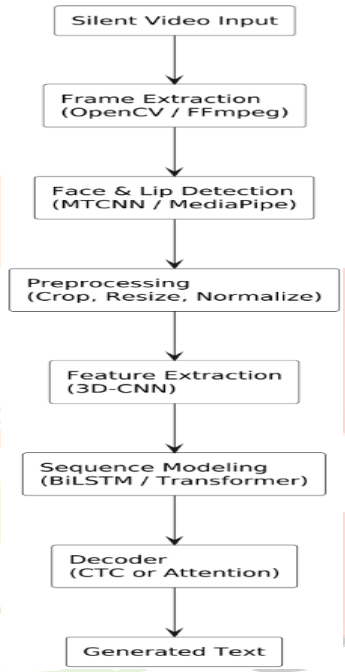


Figure 2: Proposed System Architecture

The architecture effectively combines spatial feature extraction and temporal sequence modeling to achieve accurate visual speech recognition.

2.11 Algorithm

1: Load silent video.
2: Extract frames from the video.
3: Detect face and locate mouth region.
4: Apply preprocessing operations.
5: Feed processed frames into 3D-CNN.
6: Extract deep visual features using EfficientNetB0.
7: Learn temporal dependencies using Bi-LSTM.
8: Decode output using CTC.
9: Generate text output.
10: Display predicted sentence.

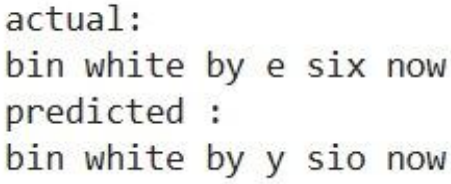
III. RESULTS AND DISCUSSION

The performance of the proposed LipSyncNet visual speech recognition system was evaluated using the widely adopted GRID Corpus, containing approximately 33,000 video samples recorded under controlled conditions. Each sample consists of a short, fixed-grammar spoken sentence, making the dataset highly suitable for benchmarking lip-reading models. Experiments were conducted on a GPU-enabled environment using PyTorch, with the model trained for 50 epochs using the Adam optimizer and the Connectionist Temporal Classification (CTC) loss function. To assess the effectiveness of the proposed architecture, we evaluated the model using standard sequence-level metrics including Character Error Rate (CER), Word Error Rate (WER), and Precision. These metrics provide a comprehensive understanding of the system's capability to accurately interpret silent lip movements. The model achieves strong quantitative performance, with a precision value of 93–95% on the test set. The overall accuracy of the model 96.7% . The best CER obtained was 6.8%, and the WER was 8.2%, indicating a small number of substitution, insertion, and deletion errors in predicted sequences. These results confirm that the integration of 3D CNN, EfficientNetB0, and Bi-LSTM effectively captures spatio-temporal lip dynamics and sequential dependencies. The EfficientNet encoder contributes significantly by extracting fine-grained lip shape and contour features, while the Bi-LSTM effectively models temporal transitions between visemes. The CTC layer further enhances accuracy by enabling flexible alignment between the variable-length input frames and character sequences without requiring manual annotation of frame boundaries.

The experimental results demonstrate that the proposed architecture successfully learns both spatial and temporal speech information. The combination of 3D-CNN and EfficientNetB0 provides effective visual feature extraction, while Bi-LSTM captures temporal speech patterns. The use of CTC Loss enables efficient sequence prediction without manual alignment.

Table 1. Performance Analysis

SN.	Actual Text	Predicted Text	Accuracy
1	bin white by e six now	bin white by y sio now	80.55
2	set red by a three please	set red by a three please	100
3	set green by g three please	set green by y three please	92.23
4	lay white with d seven soon	lay white with d seven soon	100
5	set blue at f five again	set blue at h five agin	83.33



IV. CONCLUSION

The LipSyncNet project represents an innovative step forward in the field of speech recognition and assistive communication. By utilizing deep learning and visual feature extraction techniques, it provides a reliable solution for understanding and generating speech purely from lip movements, even in noisy or audio-restricted environments. This makes the system especially useful in scenarios where traditional audio-

based speech recognition systems fail, such as crowded areas, industrial settings, or for individuals with hearing or speech impairments. The integration of convolutional and recurrent neural networks enables LipSyncNet to capture both spatial and temporal aspects of lip movements, ensuring higher accuracy and robustness. Its real-time processing capability enhances user interaction and offers potential applications in human-computer interaction, virtual communication, dubbing, and accessibility technologies. As the project evolves, it holds the potential to redefine how machines perceive and process human communication. By bridging the gap between visual and auditory speech understanding, LipSyncNet can contribute to more inclusive and intelligent communication systems, supporting advancements in AI-driven accessibility, education and media technologies.

V. REFERENCES

[1] L. Qu, C. Weber, and S. Wermter, "LipSound2: Self-Supervised Pre-Training for Lip-to-Speech Reconstruction and Lip Reading," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[2] S. Fenghour, D. Chen, K. Guo, B. Li, and P. Xiao, "Deep Learning-Based Automated Lip-Reading: A Survey," *IEEE Access*, vol. 9, pp. 121–144, 2021.

[3] N. Rathipriya and N. Maheswari, "A Comprehensive Review of Recent Advances in Deep Neural Networks for Lipreading With Sign Language Recognition," *IEEE Access*, vol. 11, pp. 78510–78525, 2023.

[4] D. Li, Y. Gao, C. Zhu, Q. Wang, and R. Wang, ''Improving speech recognition performance in noisy environments by enhancing lip reading accuracy,'' Sensors, vol. 23, no. 4, p. 2053, Feb. 2023, doi: 10.3390/s23042053. VOLUME 12, 2024 110903 S. A. A. Jeevakumari, K. Dey: LipSyncNet: A Novel Deep Learning Approach for VSR.

[5] S. Jeon and M. S. Kim, ''End-to-end sentence-level multi-view lipreading architecture with spatial attention module integrated multiple CNNs and cascaded local self-attention-CTC,'' Sensors, vol. 22, no. 9, p. 3597, May 2022, doi: 10.3390/s22093597.

[6] C. Sheng, X. Zhu, H. Xu, M. Pietikäinen, and L. Liu, ''Adaptive semantic-spatio-temporal graph convolutional network for lip reading,'' IEEE Trans. Multimedia, vol. 24, pp. 3545– 3557, 2022, doi: 10.1109/TMM.2021.3102433.

[7] H. Wang, G. Pu, and T. Chen, ''A lip reading method based on 3D convolutional vision transformer,'' IEEE Access, vol. 10, pp. 77205–77212, 2022, doi: 10.1109/ACCESS.2022.3193231.

[8] M. A. Haq, S.-J. Ruan, W.-J. Cai, and L. P. Li, ''Using lip reading recognition to predict daily Mandarin conversation,'' IEEE Access, vol. 10, pp. 53481–53489,2022,doi: 10.1109/ACCESS.2022.3175867.

[9] M. Miled, M. A. B. Messaoud, and A. Bouzid, ''Lip reading of words with lip segmentation and deep learning,'' Multimedia Tools Appl., vol. 82, no. 1, pp. 551–571, Jan. 2023, doi: 10.1007/s11042-022-13321-0.

[10] S. Fenghour, D. Chen, K. Guo, B. Li, and P. Xiao, ''Deep learningbased automated lipreading: A survey,'' IEEE Access, vol. 9, pp. 121184–121205, 2021, doi: 10.1109/ACCESS.2021.3107946.

[11] R. A. Ramadan, ''Detecting adversarial attacks on audio-visual speech recognition using deep learning method,'' Int. J. Speech Technol., vol. 25, pp. 625–631, Jun. 2021, doi: 10.1007/s10772-021-09859-3.