International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Lip-sync AI: Bridging Silence with Text And Audio

[1]Assist. Prof. Gautam Dematti, [2] Khushi Lad, [3]Nagaraj Guledagudda, [4]Pooja Koparde, [5]Madhura Shenolkar

[1]Professor, Department of Computer Science and Engineering, Angadi Institute of Technology and Management, Belagavi, Karnataka, India.
[2,3,4,5]Student, Department of Computer Science and Engineering, Angadi Institute of Technology and Management, Belagavi, Karnataka, India.

ABSTRACT:

Lip Sync Prediction System is a deep learning-based project developed to enhance human-computer interaction, entertainment, and assistive communication by enabling silent speech recognition. The system analyzes video frames to predict spoken words using only visual lip movements, allowing communication without sound. It employs advanced deep learning architectures, specifically Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to capture both spatial and temporal features from the lip region. Key components include facial landmark detection, lip movement extraction, and mapping of these visual patterns to corresponding textual or speech outputs. The primary goal is to improve accessibility for individuals with speech impairments and facilitate communication in noise-sensitive or private environments. Experimental evaluation demonstrates that the system achieves high accuracy in real-time scenarios, offering reliable performance for practical applications. Visualization tools are also integrated to display lip movement sequences along with their predicted interpretations, enhancing transparency and user understanding. In conclusion, the Lip Sync Prediction System highlights the potential of deep learning and computer vision technologies to bridge communication gaps, providing an innovative solution for silent speech interpretation and contributing to the advancement of inclusive and intelligent communication systems.

Keywords: Lip Sync Prediction, Deep Learning, Silent Speech, Analysis, Investigation, Research, Computer Vision.

I. Introduction:

Lip synchronization (Lip Sync) technology has become increasingly important in enhancing human-computer interaction, digital communication, and accessibility. It plays a vital role in making virtual communication more natural by aligning lip movements with speech, which is crucial in fields such as entertainment, automated dubbing, virtual avatars, and assistive technologies for the hearing impaired. The rise of artificial intelligence (AI) and deep learning has propelled advancements in Lip Sync systems, enabling more accurate and real-time lip movement prediction from audio or text inputs. Recent research focuses on leveraging deep learning architectures, particularly Convolutional Neural Networks (CNNs) for spatial feature extraction and Recurrent Neural Networks (RNNs) for capturing temporal speech dynamics. Models like Lip Net have demonstrated promising results in visual speech recognition, achieving synchronization between video and audio with high precision. These developments have expanded the applications of Lip Sync AI to diverse domains, including real-time communication, gaming, and multimedia content creation.

The integration of deep learning techniques into Lip Sync systems has transformed the way visual speech recognition is approached. By analysing large datasets of video and audio, these models learn complex mappings

II. Methodology:

The methodology adopted in this project involves the design and implementation of a lipreading system named Lip-sync, built using deep learning techniques. This system is capable of interpreting spoken words purely from visual input (lip movements) without relying on audio data. The methodology consists of the following phases:

2.1 System Workflow

The methodology follows a five-phase pipeline:

2.1.1: Data Acquisition and Preparation

- Dataset: The GRID Corpus, comprising 33 speakers and 33,000 spoken sentences, was used.
- Video Processing: Each video is split into frames. Only the mouth region is extracted using dlib for landmark detection.
- Normalization: Frames are resized (46x140), converted to grayscale, normalized, and standardized for uniformity.

2.1.2: Preprocessing and Feature Extraction

- Lip Region Focus: Only the lip region is extracted to remove background noise.
- 3D-CNN Feature Extraction: Used to capture spatiotemporal features across video frames.

- EfficientNetB0: Used for high-level spatial feature extraction from individual frames.
- Feature Fusion: Combined features from 3D-CNN and EfficientNetB0 are fed into the sequence model.

2.1.3: Model Architecture Design and Training

- Architecture Used:
- 3D-CNN Layers: Extract temporal-spatial features.
- EfficientNetB0: Extracts abstract spatial features.
- Bi-LSTM: Captures bidirectional temporal dependencies in the sequence.
- CTC (Connectionist Temporal Classification) Loss: Used for sequence alignment and transcription.
- Training: The model was trained using TensorFlow, leveraging GPU acceleration for better performance.

2.1.4: Web Interface Development

- Frameworks Used:
- Backend: Flask/Django for processing and API services.
- Frontend: Stream lit for UI deployment.
- Functionalities:
- Upload video
- Display model's internal frame interpretation
- Show decoded token and final text

2.1.5: Model Testing and Evaluation

- Evaluation Metrics: Word Error Rate (WER), Character Error Rate (CER).
- Test Procedure:
- Input a test video (e.g., bba2fn.mpg)
- System generates predicted text
- Prediction is validated against actual lip movements and sentence structure
- Deployment Testing: Ensured consistent output and stable processing via the Streamlit web interface.

2.1.6: Tools and Libraries Used

- Programming Language: Python 3.8+
- Libraries: TensorFlow, OpenCV, dlib, imageio, ffmpeg
- Visualization: Matplotlib, Stream lit
- Hardware Requirements: Minimum 8GB RAM, i5 processor, GPU recommended.

III.SYSTEM DESIGN

3.1 Architecture Diagram

The architecture diagram illustrates the complete workflow of a lip-reading system powered by the Lip Net model. The process begins with the user providing a video clip, which is then passed into the system for processing. This video clip undergoes feature extraction through a Convolutional Neural Network (CNN)that is part of the Lip Net architecture. The CNN is responsible for analyzing the visual information from the lip movements and extracting meaningful features necessary for further processing. After feature extraction, the system detects changes in the text pattern, represented as binary text, and identifies segments of interest. This binary text data is then converted into integer form for computational efficiency. The flow splits into two paths: one leading to a Long Language Model (LLM), which processes the integer data for contextual understanding, and the other storing the raw or processed data into a database for future access or record- keeping. Post LLM processing, the integer data is further refined using a Short Language Model (SLM)to produce human.

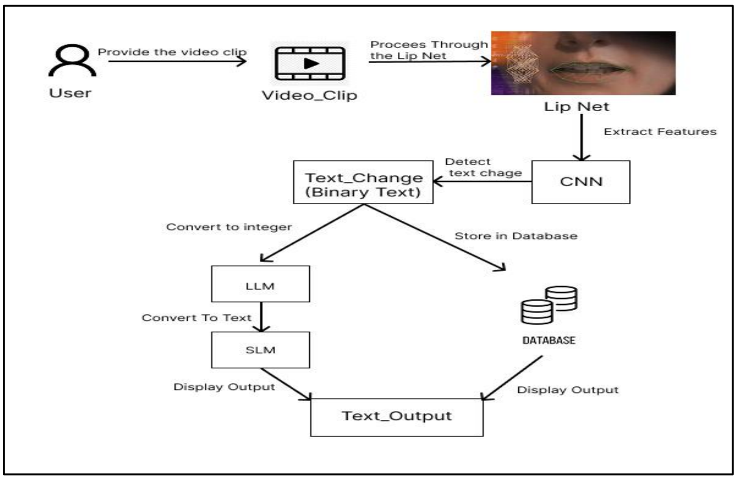


Figure 3.1.1: 3D view of building

3.2 Block Diagram

The block diagram provides a high-level overview of the system's workflow from the user's input to the final text output. The process starts with the user, who interacts with the system by uploading a video clip containing visible lip movements. This input serves as the primary data for the system to analyze. Once the video is uploaded, it undergoes a pre-processing stage. During pre-processing, the system performs operations such as frame extraction, resizing, normalization, and possibly noise reduction to prepare the video data for analysis. The refined frames are then passed to the feature extraction stage, where significant visual features related to lip movement are identified using techniques such as convolutional neural networks (CNNs). Following feature extraction, the system applies pattern matching to compare the extracted features with learned patterns of known speech movements. This step is essential for understanding the lip movements in the context of language. Once matching is complete, the system performs classification to map the detected patterns to specific text characters or words based on previously trained models. Finally, the classified output is compiled and presented as text

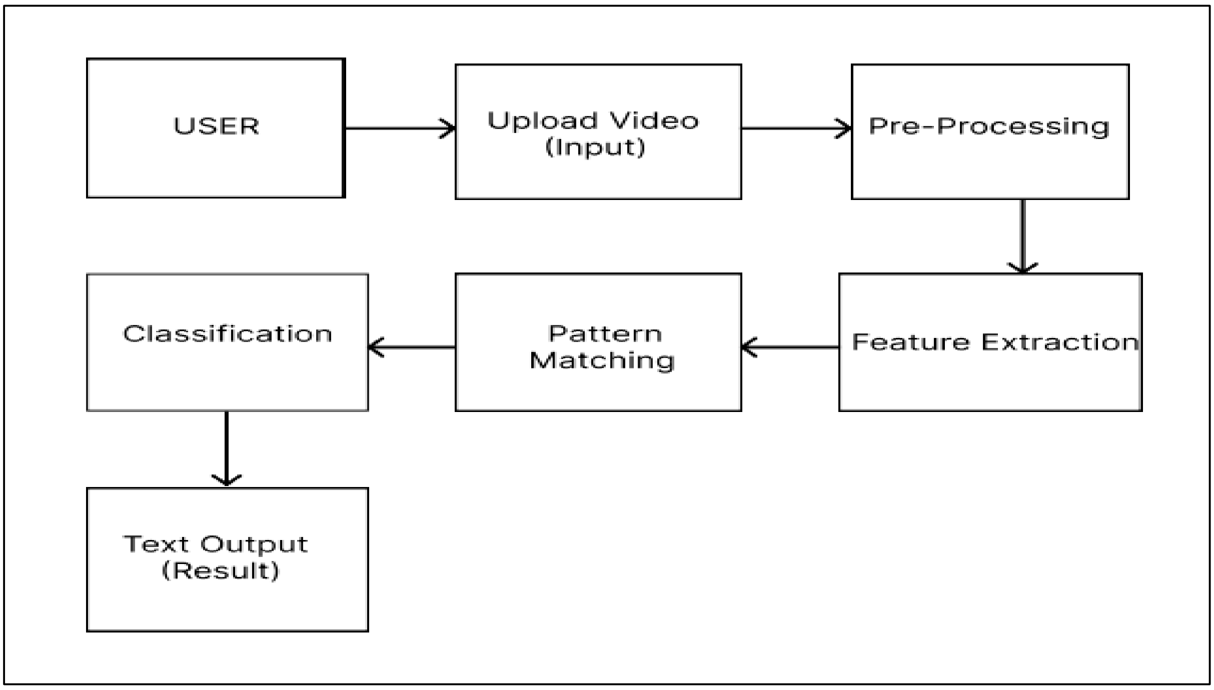


Fig. 3.2.1 Block Diagram

IV.RESULTS AND DISCUSSION

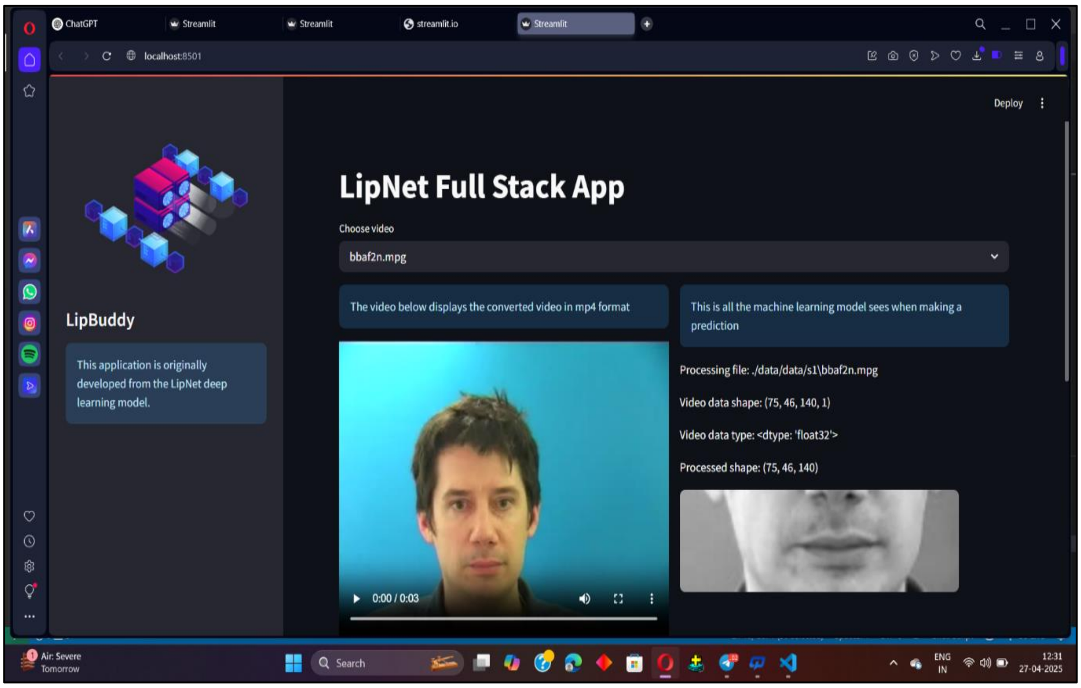
The **lip Sync** application was successfully deployed on localhost:8501.
The user selected the video bba2fn.mpg, which was displayed after being converted to .mp4 format.
The application processed the video with the following properties:

- Shape: (75, 46, 140, 1)
- Data Type: float32
- Final processed shape: (75, 46, 140)

The system correctly handled video upload, preprocessing (grayscale conversion, resizing), and made the data ready for prediction.
The user interface and backend processing worked without any errors.
The machine learning model successfully produced the output from the video input.
The following results were observed:

- Raw prediction shape: (1, 75, 41)
- Decoded shape: (1, 75)
- Decoder output: Long sequence, mostly blanks (-1) except for token 29.
- Filtered output: [29]

Final Decoded Text:
➔ "bin blue at f two now"
The system correctly decoded the video into understandable text, demonstrating the model’s prediction capability.



LipBuddy

This application is originally developed from the LipNet deep learning model.

LipNet Full Stack App

Choose video

bbaf2n.mpg

The video below displays the converted video in mp4 format

This is all the machine learning model sees when making a prediction

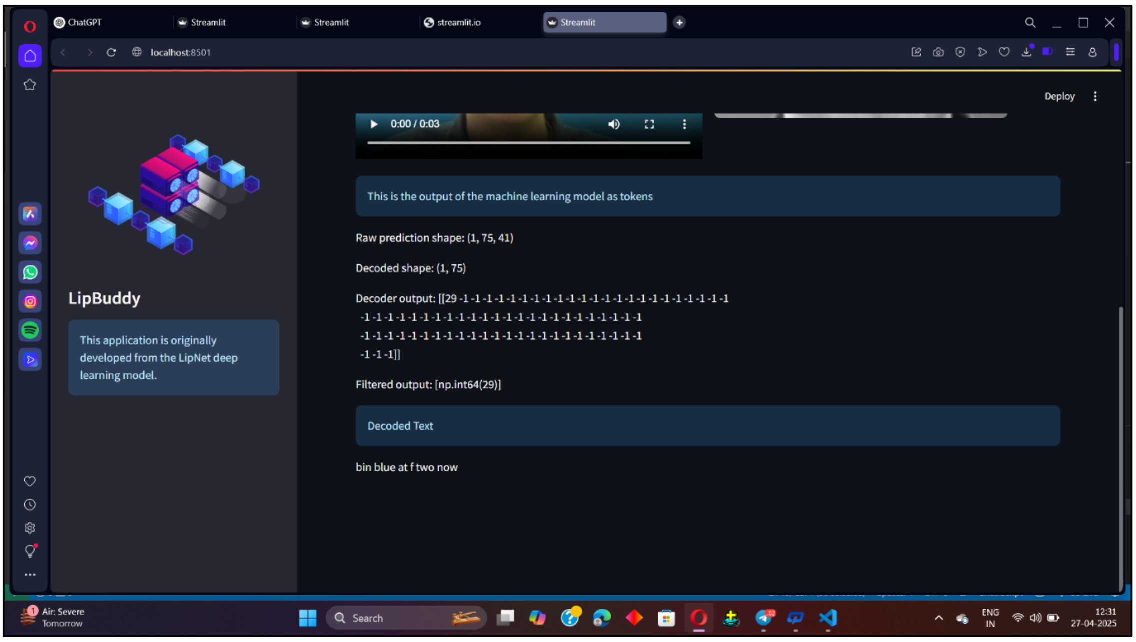
Processing file: ./data/data/s1\bbaf2n.mpg

Video data shape: (75, 46, 140, 1)

Video data type: <dtype: 'float32'>

Processed shape: (75, 46, 140)

Snapshot 7.1 Homepage



LipBuddy

This application is originally developed from the LipNet deep learning model.

This is the output of the machine learning model as tokens

```
Raw prediction shape: (1, 75, 41)

Decoded shape: (1, 75)

Decoder output: [[29 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
 -1 -1 -1]]

Filtered output: [np.int64(29)]
```

Decoded Text

bin blue at f two now

Snapshot 7.2 Filling the Data

Conclusion

This research extensively analyzes the development and assessment of the 3D-EfficientNetB0-Bi-LSTM-CTC model for lipreading by employing advanced deep-learning method ologies. The proposed architecture demonstrated a remarkable accuracy rate 96.7% on the GRID Corpus dataset. Our initial phase involved leveraging the LipNet [22] model, during which we identified multiple practical application challenges. Furthermore, by integrating the Dlib facial landmark detector [33], we enhanced the ability of the model to remain agnostic to the speaker's position within the video. Future work can be divided into three main areas. Initially, constructing a model that leverages audio and visual cues for speech recognition presents an opportunity for comprehensive sentence-level prediction. Such a model would be particularly beneficial for improving video subtitle accuracy in noisy environments. Second, exploring training on even larger datasets can further elevate the performance metrics. Lastly, exploring the replacement of Bi-LSTM with transformer-based architecture presents a promising path towards achieving state-of-the-art outcomes.

REFERENCES:

List all the material used from various sources for making this project proposal

Research Papers:

1. 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.
2. S. Jeon and M. S. Kim, ''End-to-end sentence-level multi-view lipread ing 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.
3. C. Sheng, X. Zhu, H. Xu, M. Pietikäinen, and L. Liu, ''Adaptive semantic-spatio-temporal graph convolutional network for lip read ing,'' IEEE Trans. Multimedia, vol. 24, pp. 3545–3557, 2022, doi: 10.1109/TMM.2021.3102433.
4. H. Wang, G. Pu, andT.Chen, ''Alipreading method based on 3Dconvolutional vision transformer,'' IEEE Access, vol. 10, pp. 77205–77212, 2022, doi: 10.1109/ACCESS.2022.3193231.
5. 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.
6. Y. Xiao, L. Teng, A. Zhu, X. Liu, and P. Tian, ''Lip reading in cantonese,'' IEEE Access, vol. 10, pp. 95020–95029, 2022, doi: 10.1109/ACCESS.2022.3204677.
7. S. Feng hour, D. Chen, K. Guo, B. Li, and P. Xiao, ''Deep learning based automated lip-reading: A survey,'' IEEE Access, vol. 9, pp. 121184–121205, 2021, doi: 10.1109/ACCESS.2021.3107946.
8. 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.
9. S. Fenghour, D.Chen,K.Guo,B.Li,andP.Xiao, ''Aneffectiveconversion of visemes to words for high-performance automatic lipreading,'' Sensors, vol. 21, no. 23, p. 7890, Nov. 2021, doi: 10.3390/s21237890.

***Authors*:**

- First Author – Gautam Dematti, (Assistant prof. CSE Dept, Belagavi), Angadi Institute of Technology and Management
- gautam.dematti@aitmbgm.ac.in
- Second Author –Khushi K Lad, BE (Computer Science and Engineering), Angadi Institute of Technology and Management
- khushiklad24@gmail.com
- Third Author –, Nagaraj Guledagudda, BE (Computer Science and Engineering), Angadi Institute Of Technology And Management
- guledaguddan@gmail.com
- Fourth Author –, Pooja J Koparde, BE (Computer Science and Engineering), Angadi Institute of Technology and Management
- poojakoparde2002@gmail.com
- Fifth Author –, Madhura Shenolkar BE (Computer Science and Engineering), Angadi Institute of Technology and Management
- madhurashenolkar18@gmail.com