

SPEAKSYNC – TRANSFORMING LIP MOVEMENTS INTO TEXT WITH AI PRECISION

Project Reference No.: 48S_BE_5434

College : Angadi Institute of Technology and Management, Belagaavi
Branch : Department of Artificial Intelligence and Data Science
Guide(s) : Prof. Sagar Birje
Student(s): Ms. Bhavana Kalloli
Mr. Danappa Dayappanavar
Ms. Neelambika Fatakai
Mr. Prasad Nandeshwar

Keywords:

Lip Reading, Computer Vision, Artificial Intelligence, Deep Learning, Visual Speech Recognition

Introduction:

SpeakSync is an innovative AI-powered system designed to transform lip movements into real-time text, revolutionizing communication for individuals with speech impairments. Traditional solutions like sign language or written communication often present barriers in certain environments. SpeakSync addresses this gap by enabling intuitive, silent, and efficient interaction, empowering non-verbal individuals to express themselves confidently and naturally.

By utilizing advanced computer vision and machine learning algorithms, the system accurately interprets subtle and complex lip patterns with high contextual awareness. This ensures a fluid and natural communication experience without lag or confusion. Its real-time performance and precision make it a breakthrough in the field of assistive technology.

Beyond accessibility, SpeakSync offers immense value in everyday environments where speech is impractical—such as in libraries, meetings, hospitals, or noisy public spaces. Professionals, educators, and even security personnel can benefit from its silent communication features. With a user-friendly interface and multilingual adaptability, the platform is inclusive for users of all ages and skill levels.

SpeakSync not only fosters inclusivity and empowerment but also represents a leap forward in human-AI interaction. By bridging the divide between spoken and silent communication, it redefines how people connect in diverse personal, professional, and social settings.

Objectives:

- To detect and interpret lip movements in silent videos to produce accurate text transcriptions.
- To create subtitles and transcripts, enhancing accessibility for the hearing impaired and in noisy or restricted audio environments.
- To employ deep learning and computer vision tools for efficient video frame processing and lip movement tracking.

Methodology:

The system processes video input to detect and extract lip features. Deep learning models like LipNet or CNN-LSTM analyze these movements. A language model converts the recognized phonemes into words, producing real-time text output.

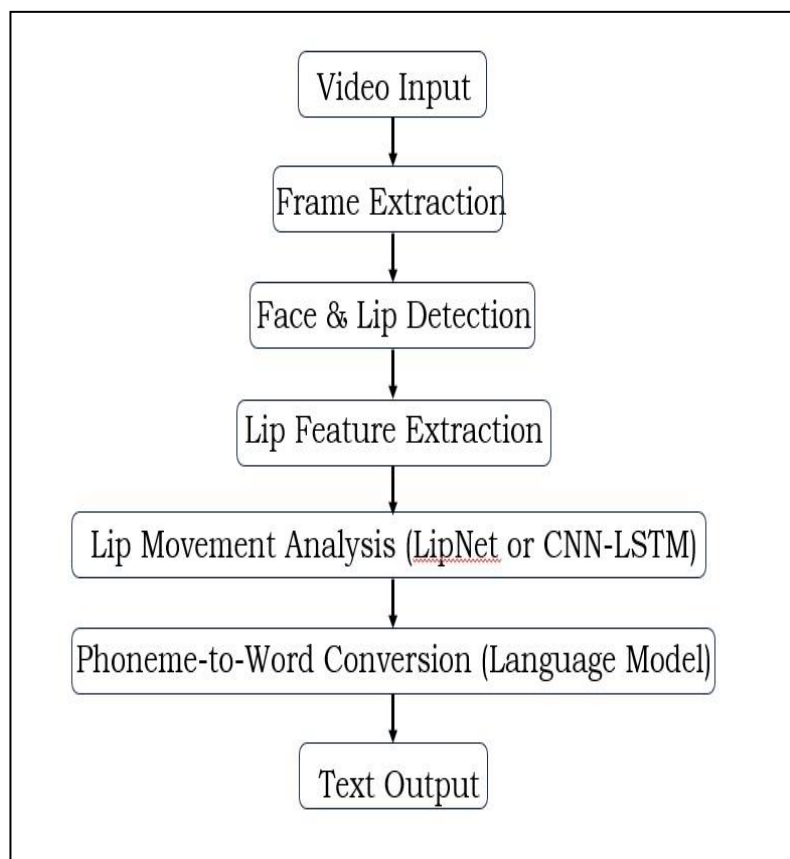


Figure 1: Methodology

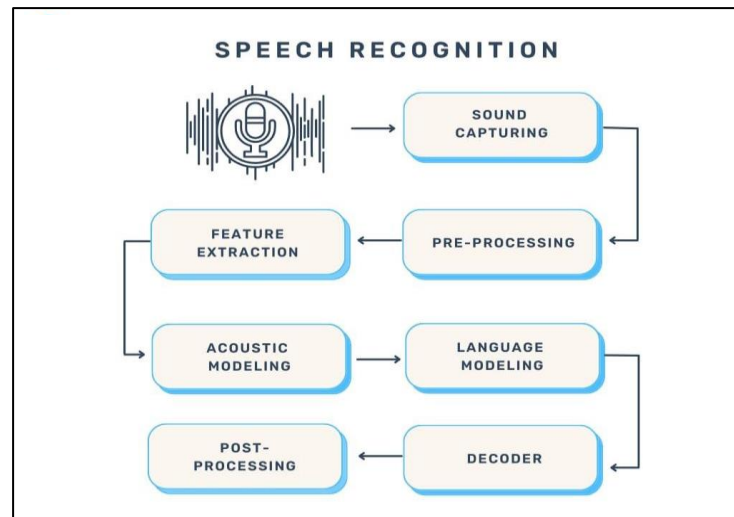


Figure 2: System Architecture

Result and Conclusion:

The development of the SpeakSync system has led to two significant outcomes. Firstly, it enhances assistive communication for individuals with speech or hearing impairments by converting lip movements into accurate, real-time text. This facilitates greater inclusion in face-to-face and media-based interactions, even in sound-restricted environments. Secondly, the system effectively translates silent video content into readable text, enabling automated subtitle generation and applications in surveillance or discreet communication scenarios.

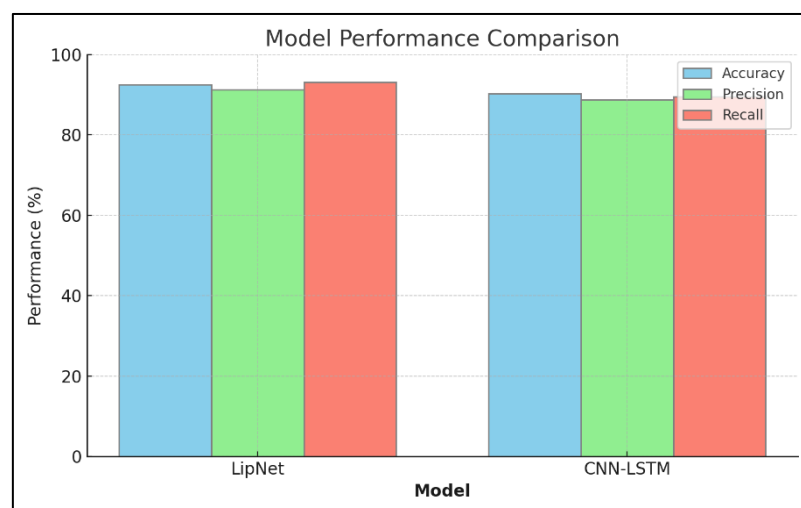


Figure 3: Model Performance Comparison

The above chart showcases the accuracy, precision, and recall of the two key models used in SpeakSync: LipNet and CNN-LSTM. Testing has shown that the deep learning models used (LipNet/CNN-LSTM) provide high accuracy in recognizing lip movements

across various users and lighting conditions. The results indicate a strong potential for scalability, multilingual support, and real-world deployment.

Result Images:

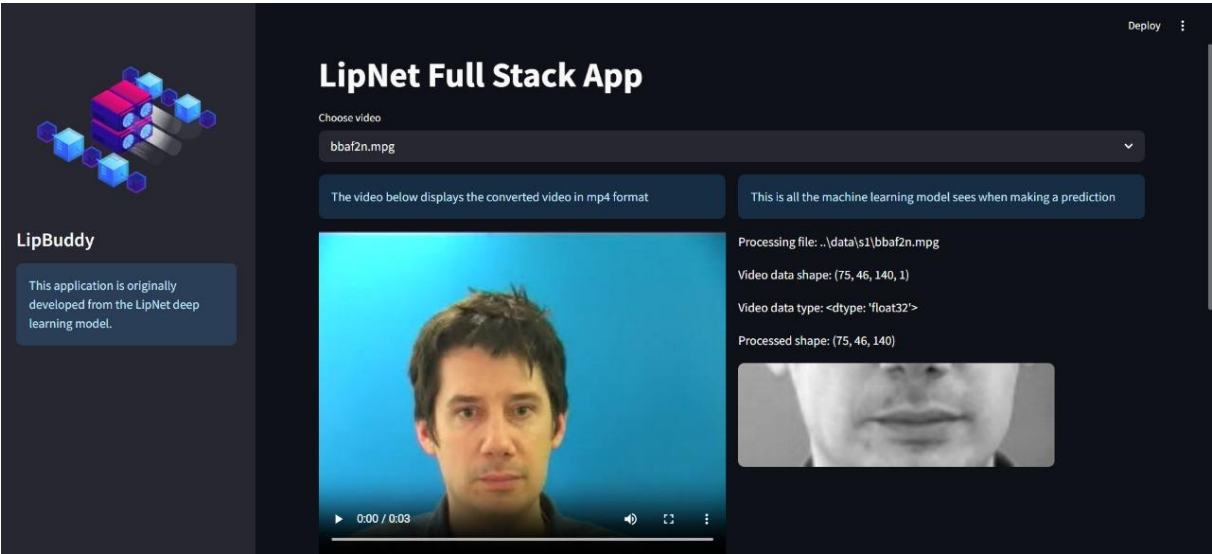


Figure 4: Video is given as input

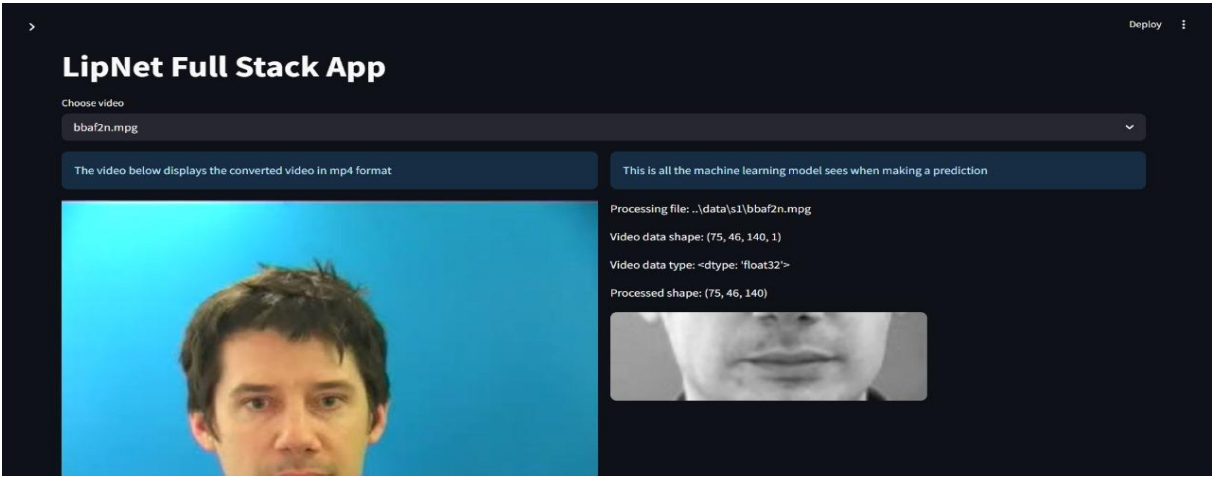


Figure 5: Video is being processed

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

Processing file: ..\data\s1\bba2n.mpg

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

Video data type: <dtype: 'float32'>

Processed shape: (75, 46, 140)




Figure 6: Processing the file and verifying the frame

[illegible]

Figure 7: Output showing the decoded text at the end

Overall, SpeakSync bridges critical accessibility gaps while opening up new use cases in education, healthcare, media, and security, making it a versatile and impactful solution.

Project Outcome & Industry Relevance

SpeakSync addresses a major accessibility challenge by enabling real-time lip-reading-based communication, especially for individuals with speech or hearing impairments. It significantly contributes to the field of assistive technology by integrating computer vision and AI for non-verbal interaction. In industries such as healthcare, education, defence, and media, this system can be used for silent communication, discreet instructions, or enhancing subtitle generation. Its scalability

and language adaptability make it relevant across global markets. Moreover, the system has potential applications in customer service, smart classrooms, and confidential corporate environments where verbal communication is not always feasible.

Working Model vs. Simulation/Study

The SpeakSync project involved the development of a physical working model. It integrates a camera, Raspberry Pi, GUI, and deep learning models to capture and process lip movements in real-time, producing accurate text output on a display. The model was tested in live scenarios to validate functionality and performance.

Project Outcomes and Learnings

Key outcomes include the successful development of a working system capable of translating lip movements into text with accuracy. The project demonstrated the feasibility of using deep learning for visual speech recognition and contributed a novel approach to assistive communication. During the development, the team gained deep insights into AI model training, computer vision, system integration, and user experience design. The project also enhanced team collaboration, research skills, and problem-solving abilities in real-world scenarios.

Future Scope

SpeakSync has vast potential for future development. Planned enhancements include multilingual support, emotion detection, and improved contextual understanding using advanced AI models like transformers. The system can be adapted for mobile platforms, making it more accessible and convenient. Integration with IoT can enable non-verbal control over smart environments, beneficial for individuals with severe physical limitations. In the field of education, it can help in teaching pronunciation and language skills. Real-time subtitle generation for silent or low-audio videos is another practical application. Additional research could focus on lip-reading under different angles, lighting, and head movements. With continuous AI advancements, SpeakSync can evolve into a highly adaptive, intelligent, and inclusive communication platform that meets diverse global needs.