WWW.IJRPR.COM

International Journal of Research Publication and Reviews

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

LipSense: Decoding Speech with Deep Learning

Mrs. Sarala Devi Ande[1], Meghana Bhaskara[2], Srinidhi Vabuloju[3], Vijay Kumar Yedelli[4], Nithish Jadala[5]

[1] Assistant Professor, Dept. of CSE-Data Science,ACE Engineering College, India
[2345] B.Tech CSE-Data Science, ACE Engineering College, India
Emails: saraladevia14@gmail.com, meghana.bhaskaraa@gmail.com, srinidhivabuloju@gmail.com, yedellikittu@gmail.com, mailnithish2004@gmail.com

A B S T R A C T

LipSense is a deep learning-based system for visual speech recognition, commonly known as lip reading. This project aims to assist individuals with hearing impairments and provide alternative speech recognition where audio signals are unavailable or distorted. The system employs Convolutional Neural Networks (CNNs) for spatial feature extraction and Recurrent Neural Networks (RNNs) for sequential decoding of lip movements from video frames. Implemented using TensorFlow and OpenCV, LipSense allows users to upload videos and generates subtitles in real time. With applications in accessibility, surveillance, education, and human-computer interaction, LipSense is a step toward making technology more inclusive and intelligent.

Keywords: Lip Reading, Deep Learning, CNN, RNN, TensorFlow, Speech Recognition, Visual Input, Accessibility.

1. Introduction

Effective communication is crucial in daily life, yet individuals with hearing impairments often face significant challenges, particularly in environments where audio communication is unreliable or unavailable. Traditional speech recognition systems rely heavily on sound, making them ineffective in noisy, silent, or distorted audio settings. These limitations reduce accessibility and hinder inclusive technological adoption.

To address this issue, we developed LipSense, a deep learning-based system that performs speech recognition purely through lip movement analysis.By eliminating the need for audio input,the system provides a fully visual method for decoding speech. Users can upload a video and the system processes the visual data through advanced neural network models built using TensorFlow and OpenCV. Once the lip movements are analyzed, the corresponding text is displayed as subtitles.Designed for accessibility, LipSense benefits individuals with hearing impairments and is also applicable in surveillance, silent communication, and human-computer interaction, offering a smart and inclusive alternative to traditional audio-based systems.

2. Literature Review

Lip reading has evolved from manual, human-based techniques to advanced AI-driven systems that aim to interpret speech purely through visual cues. Traditional approaches relied on handcrafted features such as geometric lip contours, edge detection, and motion estimation. These methods lacked robustness and failed to perform consistently in real-world scenarios due to their sensitivity to speaker variability, lighting conditions, and head movements.

A major breakthrough in the field came with the introduction of LipNet by Assael et al. (2016) [1], which presented the first end-to-end deep learning model for sentence-level lip reading. LipNet combined Convolutional Neural Networks (CNNs) for spatial feature extraction with Recurrent Neural Networks (RNNs) for temporal sequence modeling. It demonstrated high accuracy on the GRID dataset, achieving 95.2% performance, and showcased the potential of deep learning in this domain.

Further improvements were made by Chung et al. (2017) through their work, *Lip Reading Sentences in the Wild* [2]. Their research focused on developing lip-reading systems capable of operating in uncontrolled, real-world video conditions. By introducing large-scale datasets and leveraging deep architectures, they addressed challenges such as varying speaker identities, background clutter, and lighting inconsistencies.

These two studies laid the groundwork for real-time, accurate lip reading using deep learning. Building on these foundations, **LipSense** proposes a practical system that integrates these core ideas into a usable application, enabling users to upload videos and receive accurate speech transcription in the form of text, without relying on audio input.

3. Methodology

Fig. 1. LIPSENSE System Methodology

The proposed system, **LipSense**, leverages deep learning techniques to decode speech by analyzing visual lip movements captured from video input. The methodology involves a structured pipeline that processes silent video frames to generate accurate text-based predictions, as illustrated in Fig. 1.

The process begins at the **User Interface**, where users upload video. This video input is then passed through the **Face Localization Module**, which detects and isolates the facial region in each frame. Following this, the system performs **Lip Localization**, focusing specifically on the mouth area to extract a sequence of lip movements.

Once the lip region is isolated, the video frames are passed through the **Feature Extraction Module**, where Convolutional Neural Networks (CNNs) analyze spatial features such as lip shape, contour, and movement patterns. These extracted features are then fed into the **Sequence Modeling and Classification Layer**, which utilizes Recurrent Neural Networks (RNNs) or Bi-directional LSTMs to capture temporal dependencies across frames and map them to corresponding speech units.

Finally, the model output is decoded into readable text using CTC (Connectionist Temporal Classification) decoding and displayed on-screen as **predicted speech** or subtitles. All steps are processed locally, ensuring fast and efficient performance without reliance on external servers.

This methodology provides a fully contactless and audio-free speech recognition solution, making it especially useful for users with hearing impairments or in environments where audio signals are unavailable, thereby promoting inclusivity and advancing human-computer interaction.

3.1 System Architecture

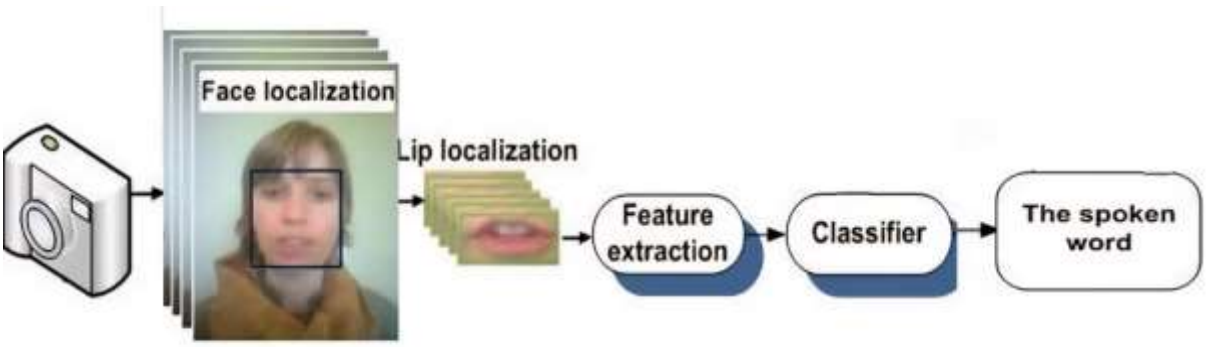

Fig. 2. LIPSENSE System Architecture

The architecture of the proposed **LipSense system** consists of a structured, end-to-end pipeline that performs real-time lip reading directly on the client side. The process begins at the **User Interface**, where users can upload a pre-recorded video. This interface is built to be lightweight, intuitive, and accessible for real-time interaction.

Once a video is submitted, the **Face and Lip Detection Module** is triggered. Using OpenCV and facial landmark detection, the system isolates the lip region from each frame. These cropped lip frames are then passed into the **Feature Extraction Module**, where a Convolutional Neural Network (CNN) processes the visual information and extracts spatial patterns of lip movements.

The extracted features are sequentially passed to the **Sequence Modeling and Classification Layer**, which utilizes Recurrent Neural Networks (RNNs), such as Bi-directional LSTMs, to learn temporal dependencies in the visual speech signal. The model then predicts the character or word sequences using a decoding mechanism such as **CTC (Connectionist Temporal Classification)**.

Finally, the **Speech Prediction Output** is displayed on the screen in the form of subtitles. Since the system is designed for local execution, all operations — from preprocessing to prediction — are performed on the user's machine, ensuring simplicity, speed, and ease of deployment without requiring internet connectivity or server-side processing.

4. Result and Discussion:

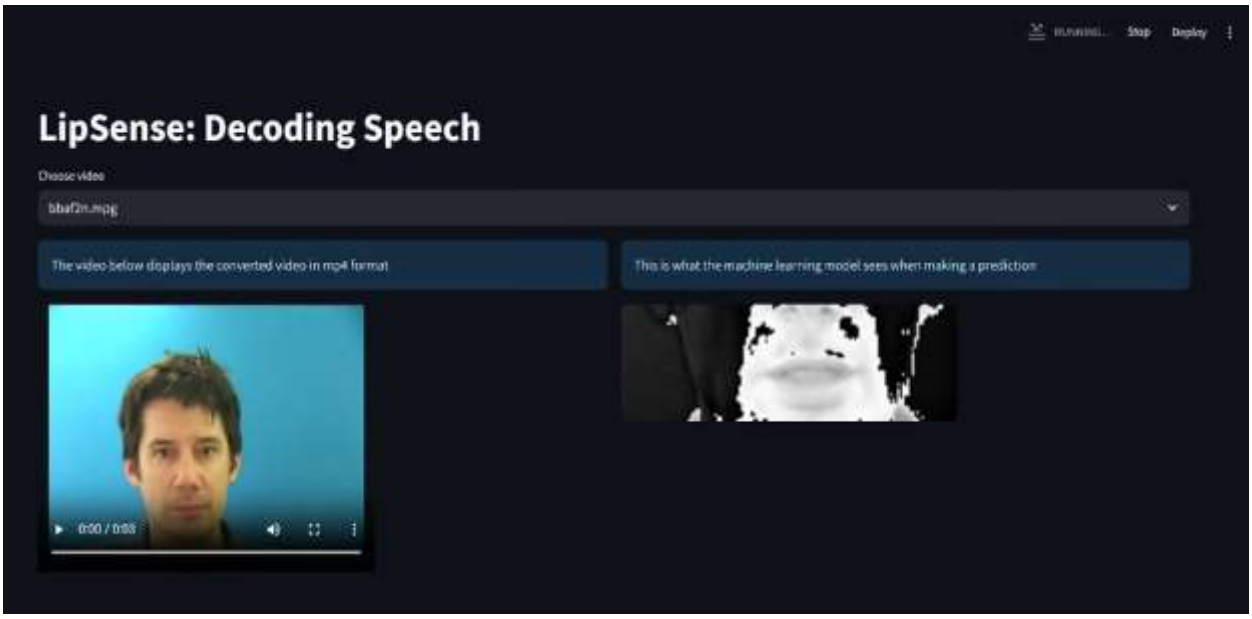

Fig 3.User Interface

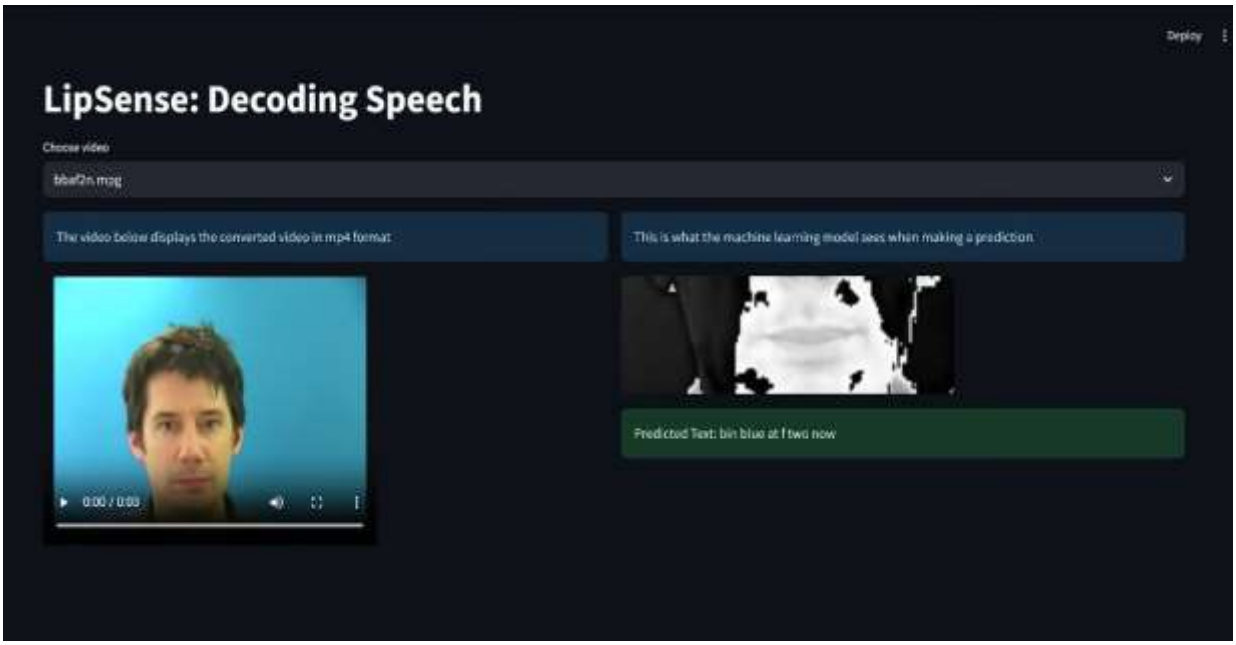

Fig 4.Predicted Output

5. Work Flow:

The **LipSense** system is designed to make silent speech recognition as simple and accessible as possible. From uploading a video to receiving the final text output, the workflow is intuitive and fully automated. Here's how it works:

Step 1: Opening the Application

Users begin by launching the LipSense interface, which runs as a web-based platform. The interface is built using Streamlit, offering a clean and user-friendly environment to interact with the system.

Step 2: Providing a Video

The user is prompted to upload a pre-recorded video file.This video should clearly show the speaker's face, especially the lips, for best results.

Step 3: Detecting the Face and Lips

After the video is submitted, the system automatically identifies the face in each frame and then zooms in on the lip region. This step is handled using computer vision techniques with OpenCV and facial landmark detection.

Step 4: Extracting Lip Features

The isolated lip regions are analyzed using a Convolutional Neural Network (CNN), which picks up important visual details like lip shape and motion. These visual cues help the system understand what is being said.

Step 5: Understanding the Sequence

Next, the system looks at the entire sequence of lip movements using a Recurrent Neural Network (RNN), such as BiLSTM. This allows it to make sense of the changes over time and interpret the flow of speech.

Step 6: Converting to Text

Using a technique called Connectionist Temporal Classification (CTC), the model translates its understanding of the visual data into actual words and sentences. The result is a readable text output.

Step 7: Showing the Result

Finally, the predicted text is displayed on the screen, giving users a clear subtitle-style output of what was said—completely without sound.

This workflow allows LipSense to turn silent videos into readable speech, making it a helpful tool for accessibility, quiet environments, and innovative communication.

6. Conclusion and Future Works:

The **LipSense** project presents a deep learning-based lip reading system designed to convert silent video inputs into accurate text using visual speech recognition. By leveraging Convolutional and Recurrent Neural Networks with tools like TensorFlow and OpenCV, LipSense provides a contactless and audio-free communication alternative. It achieves:

*A practical, real-time lip reading solution using visual input without relying on sound.

*Seamless subtitle generation from recorded video for improved accessibility.

*Enhanced communication support for hearing-impaired individuals in everyday and educational settings.

*Applicability in non-audio environments such as surveillance, industrial zones, and silent human-computer interaction.

*Robust integration of modules such as video processing, lip detection, feature extraction, sequence modeling, and text output in a single pipeline.

*A step forward toward inclusive technology that bridges communication gaps in diverse contexts.

Future Scope:

1.Multilingual Lip Reading:

Extend support to multiple languages and dialects, enabling wider global usage.

2.Real-Time Implementation:

Deploy the system on edge devices or web-based platforms to process real-time streaming video for live subtitle generation.

3.Emotion & Context Awareness:

Integrate facial emotion recognition and contextual cues to improve prediction quality in expressive or conversational scenarios.

4.Speaker Independence:

Train the model on a more diverse dataset to enhance generalization across speakers with different facial structures, accents, or speech patterns.

This conclusion and future direction firmly position **LipSense** as a scalable, intelligent, and human-centric lip reading solution, suitable for accessibility tools, smart interfaces, and security applications in the modern world.

7.Acknowledgement:

We express our sincere gratitude to all who supported us throughout the development of this project.We are especially thankful to Prof. Y V Gopala Krishna Murthy, General Secretary, and Mrs. M Padmavathi, Joint Secretary, for providing us the opportunity and environment to carry out this work.Our heartfelt thanks to Dr. P Chiranjeevi, Head of the Department, for his guidance and encouragement.We are deeply grateful to our internal guide Mr. G Parwateeswar, and project coordinator Mrs. B Saritha, for their consistent support, valuable feedback, and motivation throughout the project.

Lastly, we thank all the faculty members, staff for their constant encouragement and support.

8.References:

[1] Assael, Y. M., Shillingford, B., Whiteson, S., & de Freitas, N. (2016).

LipNet: End-to-End Sentence-Level Lipreading.

University of Oxford, DeepMind. Published on arXiv.

[https://doi.org/10.48550/arXiv.1611.01599]

[2] Chung, J. S., Senior, A. W., Vinyals, O., & Zisserman, A. (2017).

Lip Reading Sentences in the Wild.

University of Oxford, Google DeepMind. Published in IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[https://doi.org/10.48550/arXiv.1611.05358]