WWW.IJRPR.COM

International Journal of Research Publication and Reviews

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

Silent Sound Technology: A Deep Learning Approach to Visual Speech Recognition.

Mr. Mahendra S. M*[*1], Ms. Deepika Basanagouda Naik[#2], Ms. Pallavi Y T [#3]

[*1]Assistant Professor, Department of Electronics and Communication Engineering, Coorg Institute of Technology, Ponnampet, Karnataka, India
[#2, #3] 8th Semester, Department of Electronics and Communication Engineering, Coorg Institute of Technology, Ponnampet, Karnataka, India

ABSTRACT

The Silent Sound Technology enables the communication without audible speech by interpreting facial and lip movements using machine learning. This paper presents a system that uses deep learning, particularly Convolutional Neural Networks (CNNs), to identify and interpret silent speech. The system is capable of operating in real time and has applications in noisy environments, for individuals with speech impairments, and in secure communication. A prototype was implemented with an integrated web interface, demonstrating the accurate performance in controlled settings.

I.INTRODUCTION

Speech serves as the fundamental medium for a human interaction. However, conventional speech recognition systems fail in environments with background noise or for users with speech disabilities. Silent Sound Technology bridges this gap by analysing visual speech clues such as lip and jaw movements. Deep learning models have shown superior performance in image recognition, and their application to lip-reading offers a scalable solution for silent communication. The aim of this research is to implement a reliable, real-time visual speech recognition system that is user-friendly and effective.

II. SYSTEM ARCHITECTURE

The system follows a multi-stage pipeline:

1. Video Acquisition: Captures live video feed using a webcam.

2. Face and Mouth Detection: Applies HAAR cascade classifiers to detect the mouth region.

3. Pre-processing: Extracts the region of interest (ROI), converts frames to grayscale, and resizes them to uniform dimensions.

4. Feature Extraction: Uses CNNs to extract spatial patterns and lip movement dynamics.

5. Classification: Applies fully connected layers followed by softmax activation to classify input into one of the predefined classes (e.g., common words).

III. HARDWARE COMPONENTS

Webcam: Captures live facial input.

Processing Unit: Laptop or desktop with a modern CPU/GPU to run the deep learning model efficiently.

Display Device: Monitor or screen for user interface interaction.

Optional: Microphone for hybrid systems integrating both audio and visual clues.

IV.SYSTEM MODEL

The proposed technique works based on Convolutional Neural Network which is a part of Deep Learning. Deep learning is a branch of machine learning which is completely based on artificial neural networks, as neural network is going to mimic the human brain so deep learning is also a kind of mimic of human brain. In deep learning, we don't need to explicitly program everything. The concept of deep learning is not new. It has been around for a couple of years now. It's on hype nowadays because earlier we did not have that much processing power and a lot of data.

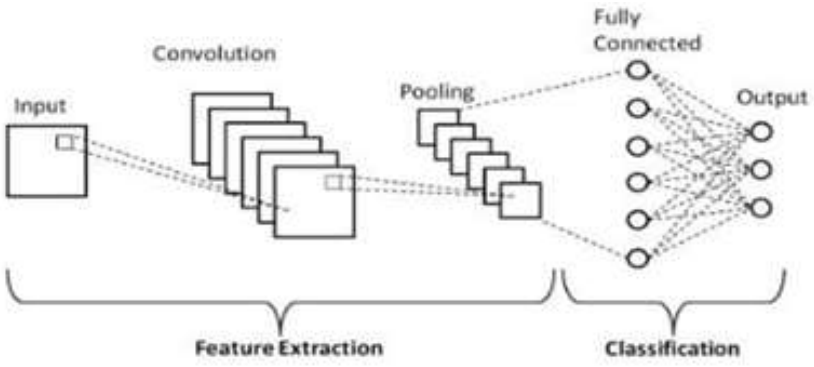

Fig. 1 Working of the CNN model

V. IMPLEMENTATION

The parameters from the input data. The activation function is a non-linear transformation function defines the output of one node which is then the input for the next layer of neurons. In the amount of parameters and computation in the network is reduced to control of overfitting by decreasing the spatial size of the network. Fully connected layer takes the input volume from the convolutional layer or pooling layer pooling layer to transform the result from the feature learning part to output.

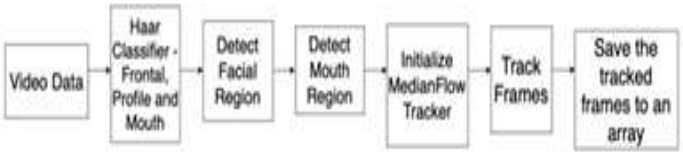

Fig. 2 Lip Detection Process

In this proposed system Our aim is to design an autonomous Silent Sound Technology system to translate lip movements in real-time to coherent sentences. We will use deep learning to classify lip movements in the form of video frames to phonemes. Afterward, we stitch the phonemes into words and combine these words into sentences.

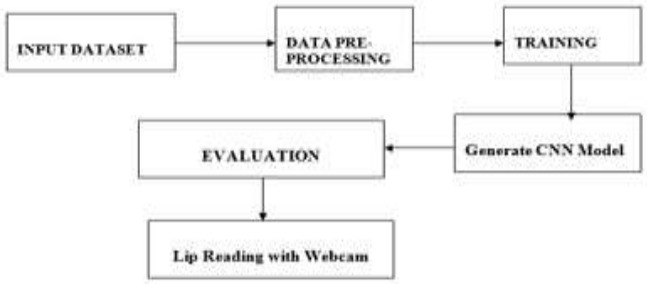

Fig. 3 Data Flow Diagram of System

This overall lip detection process is used as datasets and to train the Deep Learning Model. Later the system works as shown in Data Flow Diagram (DFD).

VI. RESULTS

1. The CNN model achieved high accuracy (above 95%) on the test dataset.

2. Real-time performance was validated using live webcam inputs.

3. The system responded within ~1 second latency, making it viable for real-time applications.

4. Misclassifications occurred primarily in visually similar words like “mat” and “bat”.

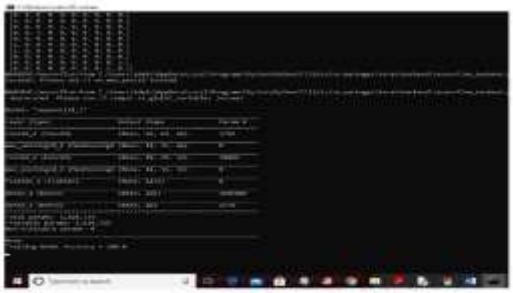

Fig. 4 CNN model training is done on dataset and we got 100% accuracy

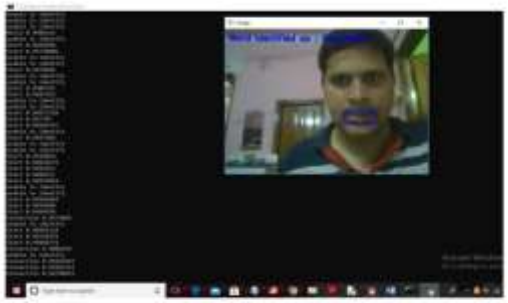

Fig. 5 Image which shows our Trained model is identifying the words

VI. CONCLUSION

This Paper successfully demonstrates a silent speech interface using visual input and deep learning. By leveraging CNNs and video processing tools, the system can predict words from silent lip movements with high accuracy.

VII. REFERENCES

1.Chollet, F. (2015). Keras: The Python Deep Learning library.

2.Assael, Y. M., et al. (2016). LipNet: End-to-End Sentence-level Lipreading.

3.Chung, J. S., & Zisserman, A. (2016). Lip Reading in the Wild.

4. Abadi, M., et al. (2016). TensorFlow: A for system large-scale mach.