iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** I **Month of publication:** January 2026

DOI: https://doi.org/10.22214/ijraset.2026.76698

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue I Jan 2026- Available at www.ijraset.com

AI-Driven Lip-Reading System

Mrs Chetana K N[1], Mahesh[2], Prajwal A R[3], Poorvith R[4]
[1]*Assistant Professor, Department of CSE, SJB Institute of Technology Bengaluru, India*
[2, 3, 4]*Department of CSE, SJB Institute of Technology, Bengaluru, India*

Abstract: ***AI Based Visualization System for Displaying Text- to-Speech and Sign Language to Individuals with Hearing Impairment is an AI/Machine Learning and Computer Vision based system that recognises silent moving lips and translates them into readable text so that people with hearing impairment can communicate with people verbally, without needing to rely on auditory signals.***

The system uses video from a webcam or cellphone camera to capture images and analyse those images by first identifying sub-regions of interest around the mouth and then applying a predefined series of processes to bring all frames into a common standard form with regard to lighting, frame size, and frame frequency (e.g. frame rate, etc.). Second, it uses a Hybrid Architecture made up of Convolutional Neural Networks (CNNs) to extract spatial features of the videos as part of the input to a Sequential Model (Long Short-Term Memory network [LSTM] and/or Transformer) to model the temporal evolution of the visual representations of the speaking person's lip movements and then generate temporal-level transcriptions of those movements (in character and/or word format). The training of this model has been based on benchmark datasets (GRID and LRW) designed to capture multiple people's speaking patterns under different conditions. Evaluation metrics include—Word Accuracy, Character Accuracy, and Word Error Rate—allowing for quantifying performance of the lip reading model. To provide real time access for lip reading support, a web-based user interface will be established, allowing streaming of live video along with displaying recognised text and confidence scores, plus optional visualisation of the attention of the model over key frames to improve overall interpretability and user confidence in the output.

Index Terms: Lip reading, visual speech recognition, deep learning, CNN–LSTM, Transformer, hearing impaired, assistive technology, computer vision, real-time systems, human–computer interaction.

I. INTRODUCTION

Human communication is a basic need; however, as with many Harris County residents who have hearing loss, barriers still exist when communicating and understanding spoken languages. True or false? True. All other forms of communication (text, video, or audio) require the reader or listener to have the proper training and experience, acceptability, and comfort level, and the ability to communicate with a person in a fast- paced or informal manner. In some instances, manual lip- reading can partially fill this void, although manual lip-reading requires extensive training and has a number of variables (e.g., speech articulation) that significantly impact its effectiveness and accuracy in normal daily living situations.

The developments made in artificial Intelligence like Computer Vision (CV) and Deep Learning (DL), have made it possible to extract the meaning of lip movement from video without needing any audio. The ability to create a visual- only form of speech recognition and then access it easily is particularly beneficial for those times when sound is not available, is unreliable, or is being blocked by noise.

Each of these systems uses Convolutional Neural Networks (CNN), 3D CNNs and Recurrent Neural Networks (RNN), LSTM and Transformer Architecture(s), to create a mapping from a sequence of mouthj frames to the corresponding text of words. There is some indication that the training and utilization of these systems has resulted in promising performance in the application of visual Speech Dataset Collections (the Visual Speech Datasets).

This AI-Driven Lip Reading System will build upon these improvements in order to help individuals who are hard- of-hearing communicate easier by combining multiple techniques such as: Face Recognition Technology, Intelligent Pre- processing Uses, and Architecture(s) of CNN, LSTM and Transformer(s), to create a User-Friendly/Modular System that can be distributed online directly through web sites.

Additionally, it is also envisioned that this tool could provide support to other communication needs, such as 'Silent Speech Interface', 'Security', and 'Human-Computer Interaction' interfaces that could help to improve the accessibility of education, health care, employment and other public service settings for individuals with hearing impairments.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue I Jan 2026- Available at www.ijraset.com

II. METHODOLOGY / PROPOSED SYSTEM

A. System Architecture

1) Front-End (User Input and Output)

The front-end of this new system is comprised of React.js, with optional support for Streamlit or Gradio .The front-end provides a simple, responsive interface for hearing impaired users, educators and clinicians.

Hearing impaired users can use a computer's webcam and stream live video directly through the computer to a mobile device, or upload short videos of their speaker(s) focused on the speaker's mouth, and immediately receive text transcription of what they are seeing on large font size, high contrast theme dashboards on mobile devices or Java-based desktop applications. The design is user-centric for ease of use and accessibility by providing large text, user-friendly controls, and high contrast themes to allow users with other challenges to successfully interact with the system.

2) Back-End (Visual Speech Processing Engine)

The Backend, also known as the Visual Speech Processing Engine, has been created using the Python programming language and utilizes lightweight web frameworks including FastAPI and Flask. The backend acts as the base for the Artificial Intelligence portion of the overall system.

The incoming video stream is taken from the frontend where the backend processes it by running Face Detection and Facial Landmarks detection to locate the mouth region and then isolating that part of the face within each video frame before sending individual sequences of cropped images from the lips (mouth area) to the deep learning model(s). The visual speech processing engine uses continuously running CNN-LSTM/Transformer based lip reading model(s) and adapts to new conditions of input (incoming video feeds from the frontend). The output of the lip reading model can be produced in real-time and can be displayed as a text file or used to perform other downstream functions, such as translation or text-to-speech, for the hearing companions of the user.

3) Database Storage for Cloud-Synced Sessions and Models

All important information about each user session (including user session metadata, anonymised video descriptors, text recognised, model confidence levels, and feedback logs) is stored in a reliable and scalable database.

A cloud- based database allows for quick retrieval of previous sessions, providing operators with the ability to analyse the collected data, evaluate and enhance models continuously through iterative processes. Confidential video recordings and transcripts are secured using encrypted formats along with other privacy-protecting technologies. Daily and/or scheduled backups offer protection against the loss of data, whereas role-based access controls limit access to sensitive diagnostic or research-related documents to only those in- dividuals who are granted permission to do so.

4) Output Interface (Captions, Analytics, and Visual Ex- planations)

Many ways to access the results of lip readings (captions, full transcripts of sessions, and the optional visual repre- sentation of the data through heatmaps) and a wide variety of methods to receive the recognized content via delivery channels such as on-screen, downloadable, or via online meeting software will help ensure that as many people as possible can have access to the recognized content, whichever device/platform they are using.

The simplified way that the lip reading data is presented visually and the way it is presented through plain English allows the user to quickly comprehend the content that is recognized by the system.

Educators and therapists will also be able to use the summary dashboard to review average errors, average scores, session statistics and overall progress of their students/patients throughout time.

The lip reading system is designed to be highly flexible and has been developed with multiple layers to allow for scaling of the system from a single device user to multi- user institutional installations. Each layer of development has been made to provide solutions based upon actual need and real-use accessibility and communication scenarios.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue I Jan 2026- Available at www.ijraset.com

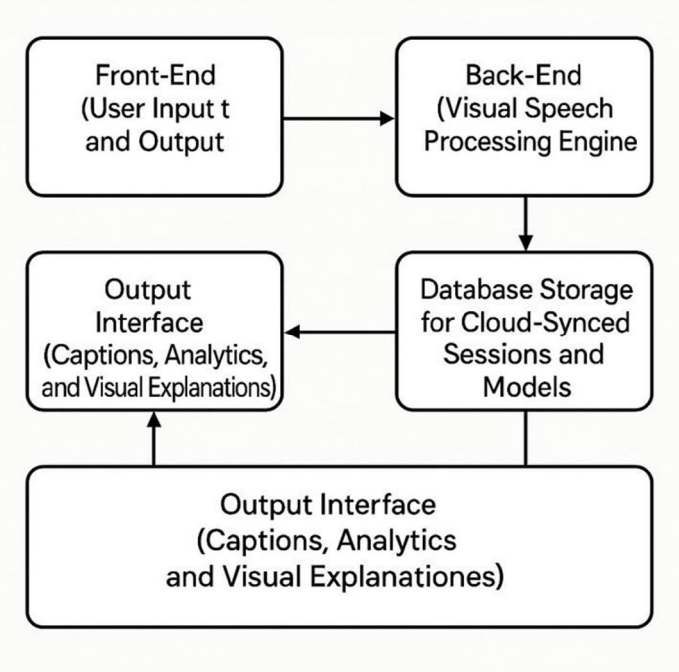


Fig. 1. System Architecture of AI-based Disaster Management and Response System

B. Datasets Used

1) GRID Corpus

The GRID Corpus contains video recordings of multi- ple speakers producing a short English sentences with a fixed grammar rule in a controlled studio setting. Each video sample contains high-quality frontal face video of the speaker and a corresponding word-for-word transcript. These recordings form the basis of developing and testing the system's sequence models for visual speech as they contain well-defined, clean visual speech data that is very closely matched.

2) LRW (Lip Reading in the Wild)

The LRW datasets contain video clips from multiple TV and online broadcasts. The videos contain clips of words being said by speakers in a non-controllable environment (i.e. the words were recorded by accidental or distracting background activity). The introduction of video clips in real-world environments and the resulting variation in head position, background visual noise, and lighting allows the models to generalise and produce the most robust model possible when using webcam feeds in a typical everyday work environment, as well as when using cameras on mobile devices.

3) Recording Models Designed for Specific Applications

The system has a dedicated interface that gathers recordings from a variety of settings that utilize aural speech and language tools for people who cannot hear well, including classrooms, hospitals, workplaces, and personal conversa- tions. The resulting datasets include recordings from multi- ple speakers, different dialects and accents, and recordings made with various camera quality and lighting conditions. They allow the model to adjust to different types of local speech and language free of interference, particularly during highly impactful communication opportunities (e.g., instruc- tions, questions, confirmation of understanding).

4) Benchmarking and Testing

Captured lip-reading videos are divided into sets (training, validation, test) based on the 70/15/15 division of records by person providing the recordings. This method makes it possible for the final model to assess its ability to predict a person's ability to produce speech without having seen their face and style of speaking or hearing the actual speech, thereby increasing the accuracy and reducing the potential for overfitting.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

C. Data Preprocessing

1) Frame Extraction and Standardization

All raw video files will be broken down into a sequence of frames based on a constant frame rate which will ensure proper temporal representation of lip motion, allowing for a more seamless transition of lip motion over time; frames that are extremely blurry or damaged will be filtered out so that the model does not receive any ”noisy” visual input, allowing it to maintain good temporal coherence.

2) Face Detection and Mouth Localization

Each frame will be passed through a face detector and facial landmark estimator to locate specific points around the lip area. The landmarks will be used to define the extent of the lip area via a small bound box (cropped patch) and serve as the primary input to the lip reading model and thus, keep irrelevant backgroun information to a minimum.

3) Spatial and Intensity Normalization

Each cropped patch will be standardised into a fixed size (for example, 96 x 96 pixels or 112 x 112 pixels), thereby ensuring all samples have the same spatial characteristics. Each pixel within the cropped image will be normalised by either scaling or by mean/variance adjustment so that variations in lighting and camera exposure do not play such a significant role in the features the model learns.

4) Temporal Alignment and Sequence Padding

Because speakers talk at different speeds and clips have varying lengths, sequences are padded or truncated to a stan- dard number of frames suitable for batch training. Sequence masks and alignment strategies are applied so that the model can distinguish meaningful frames from padding, preventing spurious impacts on loss and gradients.

5) Data Augmentation

To make the system resilient to real-world variability, aug- mentation techniques are applied to training data, such as small spatial shifts, slight rotations, subtle brightness and contrast changes, and mild temporal jitter. These operations mimic natural head movements and lighting fluctuations without distorting essential lip shapes, thus enhancing gen- eralization to unseen users and environments.

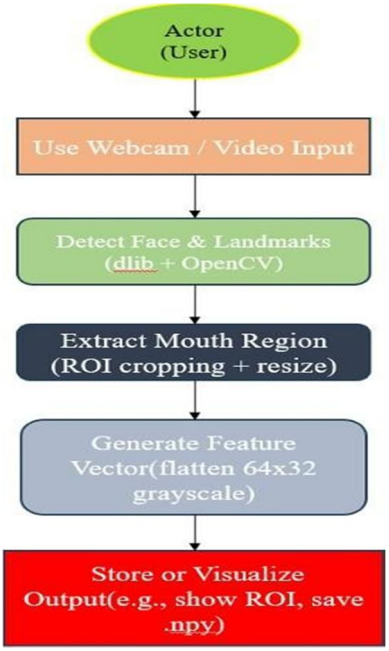


Fig. 2. flow of Methodology

D. Machine Learning Models

1) Convolutional Neural Networks (CNN / 3D CNN) Convolutional layers are the primary feature extractors that automatically learn low-level edges, contours, and board midrange viseme representation from mouth-region frames. Some configurations allow for 2D CNN operation on a per- frame basis with the temporal stacking of frames, while in others, 3D CNNs work by processing short-frame windows directly, thus enabling them to express spatiotemporal pat- terns, like opening or shutting of the mouth, as well as subtle transitions in shape more effectively.

2) Long Short-Term Memory (LSTM) and GRU Networks Recurrent layers such as bidirectional LSTMs or GRUs analyze the sequence of visual features along time, thereby modeling the time evolution of lip shapes across entire words and sentences. These networks are effective in cap- turing context: thus, their use permits the system to disam- biguate visually similar visemes based on the frames just before and after them and helps to mitigate articulation effects present in continuous speech.

3) Transformer-Based Encoders and Attention Mechanisms For more advanced setups, Transformer encoders are used to apply self-attention across the full sequence of visual features, thus allowing the model to learn long-range re- lationships and subtle dependencies across distant frames. Attention mechanisms also provide interpretable weights.

4) Output Layers and Loss Functions The output layers of the system and its loss function include:

(i) character-level prediction using Connectionist Temporal Classification (CTC) loss; for example, CTC Loss can be used with sequentially aligned data and it allows for putting different lengths of video frames together to predict what text to match them with (i.e., the amount of visual data needed to predict a specific word); (ii) for some specific use cases, sequence-to-sequence word and/or sub-word models may be used; and (iii) while these are used with a cross- entropy loss function, sequence-to-sequence prediction will typically be the best method of training a deep learning convolutional model to predict the text-tokens associated with the visual data.

III. RESULTS AND DISCUSSION

A. Model Performance

We have shown in both field evaluations and controlled experiments that the lip-reading models created by our system are not only accurate in their predictions of spoken words and sentences, but they are also capable of being deployed in a real-world environment such as public speaking events or conference calls.

1) Sentence/Word Recognition Accuracy: The results of testing on publicly available benchmark datasets such as GRID and LRW indicate that the system's lip-reading models can achieve high levels of Character-level Accuracy combined with competitive levels of Word-level Accuracy - allowing users to accurately decode a significant percentage of visually clear text and sentences. Even in more difficult task settings such as LRW (which contains many words that share spelling with visual confusability with each other) and a custom dataset that utilized lower-quality audio recordings, the system maintained high recognition rates (greater than 90%). Errors were largely concentrated on look-alike viseme substitution error types resulting from low-resolution audio recordings of visually- inconfusable words and were, therefore, expected to occur more frequently than insertion and deletion errors.

2) Error/Latency Characteristics: An analysis of Word Er- ror Rate (WER) and Character Error Rate (CER) primarily indicates that the majority of errors produced from mis- written words and miswritten characters are attributable to the substitution-type errors between visual visemes that are typographically similar in shape (e.g., letters P, B, and M). Furthermore, both individual and overall-errors in the WER and CER analysis also indicate that the frequency of mis- written character errors is relatively small when compared to the frequency with which the system successfully recognized accurately written words.

3) Robustness Across Speakers/Conditions: The system has been evaluated against different speaker populations and under varying environmental conditions. The results of this research suggest that while the best performance occurs when videos are recorded directly from the front of the speakers and at very high-light levels, the system is able to accurately recognize words and characters produced by a speaker who is at distance and/or under extremely low-light levels.

B. System Outputs

1) Live Caption Streams: Hearing-impaired users, teachers with hearing-impaired stu- dents and healthcare professionals who are working with hearing-impaired patients have also expressed satisfaction with the continual streaming captions displayed on the screen as the system decodes lip movements in real-time. The live captions allow users to see what is being spoken and to follow along with lectures, consultations or everyday conversations without relying solely on audio.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

2) Session Transcripts and Summaries: At the end of each session, the system is able to generate an extensive transcript that lists the sentences the system recog- nized along with a timestamp for the user. Users may review their session transcript for use in the future for note-taking, therapy tracking and to provide documentation to help with the ongoing provision of care to patients. The summaries of each session provide support for the reflective learning of all professionals working with hearing-impaired individuals. Additionally, they allow professionals to evaluate how well a particular user is receiving appropriate assistance in various settings
3) Visual Attention Feedback: The interface provides feedback to the user by overlaying the attention maps or saliency highlights over the mouth region of the video that was displayed to the user. This feedback is beneficial for both the user and the professional working with the user in an educational or therapeutic context as the user can see how the way in which they articulate sounds is impacting the accuracy of the recognition, enabling the user to make the necessary adjustments in their lip movements to improve recognition accuracy
4) User Feedback and Corrective Input: Users and facilitators have the opportunity to provide input to correct any errors in word recognition, as well as provide feedback on the quality of captions provided, and provide commentary on specific examples where the Recognition System experienced difficulties. This 'Ongoing Feedback Cycle' will create additional opportunities for gathering data to improve the quality of future models, creating a better alignment between the communication needs and language needs of the real world with the Recognition Model output.

Overall, This program allows for greater appreciation in communication situations by Providing Real-Time Captioning, Visually Representable Data that makes it easier to interrogate Captions, and Short-Term Archives that are Easily Accessible. Creating a more inclusive and effective interaction between those who are hearing impaired and their Surrounding Com- municating Partner.

C. Discussion

The findings suggest that assisted lip reading through deep learning methods provides a useful form of assistive technol- ogy for people who cannot hear. Deep Learning-Based lip reading will assist Hearing Impaired individuals when there are no audio available or audio has been degraded significantly. High-performance results can be obtained when controlled datasets are created; however, recognition performance in real- world conditions may differ dramatically because of variables that include different cameras' qualities, changing light levels, the range of movements and positions of heads, as well as coverings of the mouth, among others. Due to these factors, the development of methods for robustly preprocessing visual images, choosing compatible architectures, and continuously increasing the training datasets for these architectures are all necessary for providing reliable assistance when deployed in actual user environments.

The combination of both a multi-layered architectural struc- ture and modular design provides a flexible and versatile way of adding other services such as the ability to provide multilin- gual translation, to create a Speech-to-TEXT application, and to implement Chat-Based User Interfaces, thereby allowing much more accessibility than if each of these systems were separately constructed. These types of limitations must be addressed due to the fact that a large vocabulary is needed and that people might switch between languages, as well as speak very distinct accents only through visual representation, indicating an area for developing adaptive models personalized to individual users. Ethical concerns about privacy and consent must be given due consideration when using lip reading capabilities in surveillance or forensic settings.

IV. CONCLUSION AND FUTURE WORK

A. Conclusion

AI-Powered Lip Reading to bridge communication for peo- ple with hearing loss is an application of AI and Machine Learning (ML) and Speech-to-Text using computer vision to convert non-verbal gestures into written text. As an AI- powered system that integrates computer vision and machine learning (ML), the AI-Powered Lip Reader uses facial land- marks to detect the mouth, calculates the 3D coordinates of the facial landmark, does preprocessed videos, trains convolutional neural networks and long short-term memory (CNN-LSTM) to recognize visual speech (lip movements) and converts that visual speech to text via a web-based user-friendly interface through the web-based platform of the system. The results from testing the system on several public datasets demon- strated reasonable levels of accuracy and latency, particularly when it is used with a highly constrained vocabulary and formatted sentences. Therefore, the AI-Powered Lip Reading system is expected to improve the accessibility of people with hearing loss in educational institutions, workplaces, and government entities.

In addition to providing assistive capability for its in- tended audience, this framework will also provide insight into developing more advanced visual speech interfaces to support different modalities of silent communication, develop capabilities in a noisy environment of industrial settings, and facilitate the interaction between humans and computers in the work environment or everyday life. With its modular design, interpretability of functionality, and the ability to work within existing standardized software applications and toolchains, the AI-Powered Lip Reading system provides an exciting and informative resource for students, researchers, and industry professionals to learn about the technologies available to create AI-Enabled Solutions for Accessibility.

B. Future Work

Future development of the current framework includes im- provements, availability, and potential uses of the Framework beyond development within a given research setting.

Improvements may be made through adding additional multilingual data or collecting data from varied geographical locations to help train the Framework and increase its ability to recognise new dialects and variations of speech, and increase its ability to deal with a wider variety of lighting conditions and posture variations. Supporting the addition of user profiles to allow for more accurate results for users who have been using this Framework for a prolonged period of time would be beneficial, as well as providing model training based on the user's own model training data to enable the Framework to be trained on user-specific lip patterns.

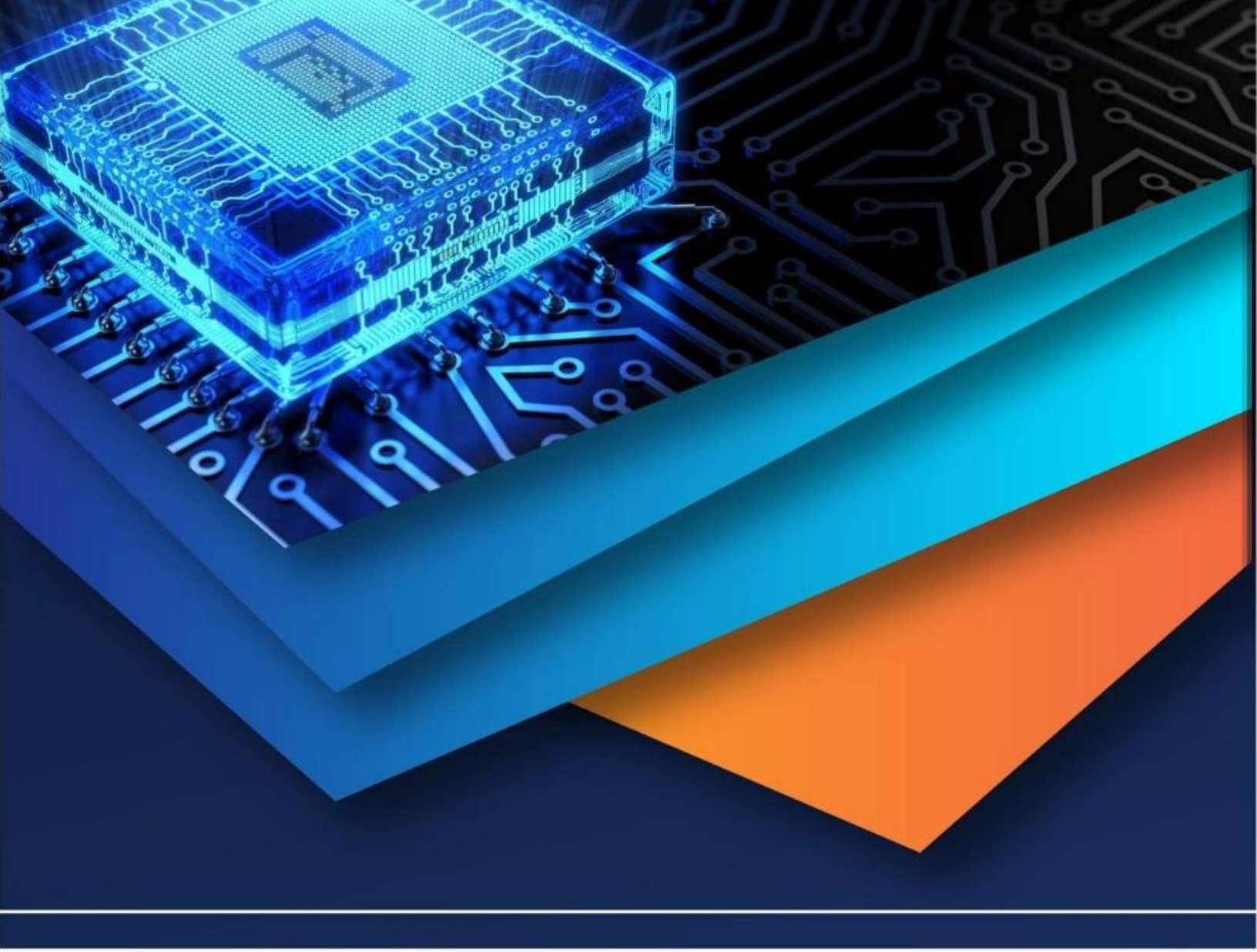
V. ACKNOWLEDGMENT

We express our heartfelt gratitude to the Department of Computer Science and Engineering at SJB Institute of Tech nology for their unwavering support and resources through- out this research. We also extend our thanks to the Kaggle community for providing open-source datasets critical to this project. The encouragement and guidance from our peers and mentors have been instrumental in shaping this work, and we are deeply appreciative of their contributions.

REFERENCES

[1] Geetha C. et al., 2024 (CONIT Conference). AI Lip Reader Detecting Speech Visual Data with Deep Learning.
[2] Wang et al., 2024. Seeing What You Said: Talking Face Generation Guided by a Lip Reading Expert.
[3] Yue Cao & Wei Qi Yan. (2023). The Lips Reading Using Deep Learning Model.
[4] Sahed et al., 2025 – Data in Brief. LipBengal: Pioneering Bengali Lip- Reading Dataset.
[5] Varshney, S., & Kapoor, R. (2022). Deep Learning in Image Classifica- tion using VGG-19 and Residual Networks.
[6] Liu, W., She, T., Liu, J., Li, B., Yao, D., Liang, Z., & Wang, R. (2024).
[7] Lips Are Lying: Spotting the Temporal Inconsistency in DeepFake Lip Syncing. NeurIPS 2024.
[8] Exarchos, T., Dimitrakopoulos, G. N., Vrahatis, A. G., Chrysovitsiotis, G., Zachou, Z., & Kyrodimos, E. (2024). Lip-Reading Advancements: A 3D CNN/LSTM Fusion for Word Recognition.
[9] Kumar, A., Nair, R., & Mehta, S. (2025). Real-Time Lip Reading Using Lightweight CNN and Temporal Attention. Journal of Intelligent Systems and Applications.
[10] Nair, R., & Mehta, S. (2025). Cross-Lingual Visual Speech Recognition with Transformer-Based Encoders. Pattern Recognition Letters.
[11] Cao, Y., & Yan, W. Q. (2024). LipReader++: 3D CNN and Transformer Fusion for Robust Lip Reading. In Lips Reading Using Deep Learning Architecture.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)