International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 12 | Dec - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Visual Speech Recognition Using Bidirectional LSTMs and Attention Mechanisms

Mrs. Shruthi T V[1], Dheeraj Gowda N[2], Chinthan B R[3], Yashwanth P[4], Chiranjeevi M[5]

[1]Shruthi T V, Assoc. Professor, Dept. of ISE, East West Institute of Technology
[2,3,4,5]Dheeraj Gowda N, Chinthan B R, Yashwanth P, Chiranjeevi M, Dept. of ISE, East West Institute of Technology

---

Abstract - This project presents an advanced Multi-Lingual Lip-Reading System designed to perform Visual Speech Recognition (VSR) by interpreting spoken words solely through visual lip movements, eliminating the need for audio input. The system uniquely addresses the linguistic diversity of India by providing simultaneous support for Kannada, Hindi, and English, complete with automatic language detection capabilities. The technical pipeline utilizes MediaPipe Face Mesh to track 31 specific lip landmarks, which are processed into a comprehensive set of 330 geometric and temporal features per frame, including velocity and acceleration derivatives. These features are analyzed by a deep learning architecture combining Bidirectional Long Short-Term Memory (BiLSTM) networks with an Attention mechanism to capture complex temporal dependencies in speech. Optimized for production environments using CUDA and cuDNN for GPU acceleration, the system achieves real-time inference speeds of 25–30 FPS with prediction accuracies ranging from 85% to 95%, making it a viable assistive technology for the deaf and hard-of-hearing community.

Key Words: Visual Speech Recognition (VSR), Multi-Lingual Lip Reading, Bidirectional LSTM (BiLSTM), Attention Mechanism, Geometric Feature Engineering, Real-Time Inference, GPU Acceleration (CUDA/cuDNN)

1. INTRODUCTION

Visual Speech Recognition (VSR), widely known as lip reading, is a transformative technology that interprets spoken language solely by analyzing visual lip movements, thereby eliminating the traditional reliance on audio input. This project presents a robust, multi-lingual lip-reading system specifically engineered to bridge communication gaps within the Indian linguistic landscape, offering simultaneous support for Kannada, Hindi, and English. By leveraging state-of-the-art computer vision and deep learning paradigms, the system captures the subtle kinematic nuances of speech to interpret words in real-time. The primary objective of this initiative is to democratize accessible communication technology, creating a versatile tool that functions effectively in noisy environments, enables silent authentication, and serves as a vital assistive aid for the deaf and hard-of-hearing community, ensuring that speech is understood accurately regardless of the acoustic environment.

To achieve high-fidelity recognition, the system utilizes a sophisticated processing pipeline that employs MediaPipe Face Mesh to extract 31 precise lip landmarks, which are subsequently converted into a rich vector of 330 geometric and temporal features per frame. A core technical objective is to attain a prediction accuracy range of 85% to 95% by utilizing a custom model architecture that integrates Bidirectional Long Short-Term Memory (BiLSTM) networks with a specialized Attention mechanism. This design allows the model to learn complex temporal dependencies and focus on the most discriminative frames within a sequence.

Beyond core recognition capabilities, the system is architected for production-grade usability, featuring a comprehensive dual interface design that includes a Training GUI for custom data collection and a Prediction GUI for live usage. A significant functional objective is the implementation of robust Automatic Language Detection, which intelligently switches between language models without manual intervention, alongside Prediction Stabilization algorithms designed to eliminate output flickering.

2. LITERATURE SURVEY

2.1 T. Thein and K. M. San, "Lip movements recognition towards an automatic lip-reading system for Myanmar consonants," 2018 12th International Conference on Research Challenges in Information Science (RCIS), Nantes, France, 2018

Thein and San's research represent a pioneering effort in the realm of automatic lip reading for a lesser-resourced

language— Myanmar. The study presents a visual-only speech recognition system focused specifically on consonants, a critical component of spoken Myanmar. The methodology involves capturing video data of speakers uttering consonants and extracting visual features from the lip region using geometric and spatial descriptors. The extracted features are then fed into a classification algorithm to recognize specific consonants, offering a foundational model for further development in this area. This approach aligns with the broader objective of building speech recognition systems in languages that lack significant audio corpora and technological support. A significant strength of this study lies in its focus on a regional language, thereby addressing the digital divide in language processing technologies. It underscores the potential of visual-only systems in contexts where audio data is either unreliable or unavailable, such as in noisy environments or for people with hearing impairments.

2.2 R. Shashidhar and S. Patilkulkarni, "Audiovisual speech recognition for Kannada language using feed forward neural network," Neural Computing and Applications, vol. 34, pp. 15603–15615, 2022

This study focuses on developing an audiovisual speech recognition system tailored to the Kannada language, utilizing a feed forward neural network (FFNN) to process and interpret bimodal input. The authors integrate both audio signals and visual cues such as lip movements and facial features to train their model. Visual features are manually extracted through methods that consider lip contours, geometric shapes, and spatial configurations. These visual features are then fused with Mel-frequency cepstral coefficients (MFCCs) from the audio stream to improve recognition accuracy. This bimodal approach is particularly beneficial in noisy environments where audio signals may be degraded. While this method improves robustness and accuracy, it is limited by its reliance on handcrafted visual features and a static FFNN architecture that processes data without capturing sequential dependencies. The FFNN treats each input frame independently, which is inadequate for capturing the progression of speech where timing and movement play a crucial role.

2.3 R. Shashidhar, S. Patil Kulkarni, G. H. L., V. Ravi, and M. Krichen, "Visual Speech Recognition for Kannada Language Using VGG16 Convolutional Neural Network," Acoustics, vol. 5, no. 1, pp. 343-353, 2023

Building on their earlier work, the authors in this study advance their approach by adopting the VGG16 deep convolutional neural network. This model architecture enables automatic feature extraction from lip regions in video frames without requiring manual engineering. The system is trained to recognize Kannada words using only visual speech data, thus transitioning toward a more scalable and generalized visual speech recognition framework. The use of transfer learning from VGG16, a well-established architecture, helps in leveraging pre-trained knowledge and improving recognition accuracy on limited datasets. The use of VGG16 marks a notable improvement over traditional methods, particularly in its ability to extract hierarchical spatial features. However, VGG16 is inherently designed for static image classification and lacks the mechanisms to model temporal sequences, which are essential in speech processing. Since speech is a temporal process, the absence of time-sequential modeling significantly curtails the model's effectiveness.

2.4 Z. Gan, H. Zeng, H. Yang, and S. Zhou, "Construction of word level Tibetan Lip-Reading Dataset," 2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP), Shanghai, China, 2020

Gan and colleagues focus their research on dataset creation for Tibetan lip reading—a critical step for advancing computational linguistics in underrepresented languages. The paper introduces a dataset of word-level video clips with annotated Tibetan utterances, thereby providing a foundation for training and evaluating lip reading systems specific to the Tibetan language. The dataset includes variations in lighting, speaker identities, and speech speed to ensure diversity and robustness. The work is particularly valuable in facilitating the development of machine learning models for a language with scarce digital resources and no standardized speech corpus. While the dataset fills a significant gap in multilingual visual speech recognition, the paper does not present an accompanying recognition model or benchmark results, making it difficult to assess its immediate utility in practical systems.

3. METHODOLOGY

3.1 Video Acquisition and Frame Extraction

The data acquisition phase begins with the capture of video input, either through a live webcam feed or pre-recorded video files, processed at a target rate of 30

frames per second (FPS) to ensure smooth motion capture. The system enforces a standardized input sequence length of 75 frames, corresponding to approximately 3 seconds of video, which provides sufficient temporal context for recognizing complete words. For real-time processing, a sliding window buffer is utilized to maintain this 75-frame context continuously, ensuring that the model always has access to the immediate history of lip movements.

3.2 Face Detection and Landmark Tracking

To isolate the relevant visual speech signals, the system employs MediaPipe Face Mesh, a lightweight convolutional neural network (CNN) that detects 468 distinct 3D facial landmarks. From this dense mesh, the system applies a specialized filter to extract only the 31 landmarks relevant to speech: 20 points defining the outer lip contour and 11 points defining the inner lip contour. This step drastically reduces the computational load by discarding irrelevant facial data (such as eyes or nose) while retaining high-precision coordinates for the mouth region.

3.3 Geometric Feature Engineering

The core of the system's visual analysis relies on a rigorous geometric feature extraction process that computes 161 static features per frame from the 31 raw landmarks. These features are designed to be invariant to scale and translation, starting with normalized coordinates centered around the mouth centroid and scaled to a unit square. The system calculates Euclidean distances between key landmark pairs to measure mouth opening and widening, as well as angular features between consecutive landmarks to capture the curvature and shape of the lips. Additionally, complex geometric properties such as the aspect ratio, total mouth area, and convex hull solidity are computed to provide a robust mathematical description of the lip shape at every instant.

3.4 Temporal Dynamics and Vector Concatenation

Recognizing that speech is defined by motion rather than static shapes, the methodology incorporates temporal feature engineering to capture the dynamics of lip movement over time. For the 75-frame sequence, the system computes the first-order derivative (velocity) and the second-order derivative (acceleration) for all 161 geometric features. By calculating the change in feature values between consecutive frames (velocity) and the rate of change of those velocities (acceleration), the system creates a comprehensive feature set that describes how the lips are moving, not just where they are.

3.5 Data Augmentation Pipeline

To prevent overfitting and ensure the model generalizes well across different environmental conditions, a robust on-the-fly data augmentation pipeline is implemented that expands the dataset by a factor of 10. The system generates synthetic variations of the training data by applying random transformations such as brightness adjustments, contrast scaling, and slight rotations to simulate head tilt. Additionally, horizontal flipping is used to mirror the video, and Gaussian noise is injected into the feature vectors to simulate sensor noise or low-light conditions.

3.6 Deep Learning Model Architecture

The sequence classification is performed by a custom deep neural network that combines Bidirectional Long Short-Term Memory (BiLSTM) layers with an Attention mechanism. The architecture begins with dense layers for feature processing, followed by two BiLSTM layers (256 and 128 units) that process the sequence in both forward and backward directions to capture temporal dependencies relative to both past and future frames. A custom Attention layer is integrated to automatically assign weights to different frames in the sequence, allowing the model to focus on the most discriminative moments of lip movement while ignoring transitional or silent frames.

3.7 GPU-Accelerated Training Strategy

The training process is optimized for high performance using the Adam optimizer with a learning rate of 0.001 and gradient clipping to prevent exploding gradients in the LSTM layers. The system leverages NVIDIA CUDA (v11.3) and cuDNN (v8.2) libraries to offload matrix operations to the GPU, achieving a speedup of 10-20x compared to CPU training.

3.8 Inference and Prediction Stabilization

During the real-time inference phase, the system processes the buffered feature sequences to generate predictions with a latency of under 150ms. To address the inherent instability of frame-by-frame predictions, a post-processing stabilization algorithm manages a history buffer of the last 10-15 predictions. A frequency-based voting mechanism identifies the most common prediction within this window, and a result is only displayed to the

user if it meets a strict stability requirement (e.g., 10 consecutive matches) and a confidence threshold exceeding 65%.

3.9 Automatic Language Detection (Optional)

The methodology extends to multi-lingual environments through an automatic language detection system that runs parallel inference across language-specific models (Kannada, Hindi, English). Instead of forcing the user to manually select a language, the system feeds the incoming feature sequence into all loaded models simultaneously and compares the resulting confidence scores. The model that yields the highest confidence probability for a given sequence is selected as the active language source, allowing the system to dynamically switch between languages based on the user's speech.

4. WORKFLOW

4.1 System Initialization: - The process begins by loading the system configuration from config. yaml using PyYAML to set parameters like sequence length (75 frames) and batch size. The system detects and initializes the NVIDIA GPU environment, configuring memory growth to prevent Out-Of-Memory (OOM) errors and verifying the presence of CUDA 11.3 and cuDNN 8.2.0.1 libraries. Finally, it loads the pre-trained deep learning models (.h5 files) and their corresponding class mapping files (.json), which translate numerical predictions back into human-readable words for supported languages like Kannada, Hindi, and English.

4.2 Video Capture: - The system initiates a video stream using OpenCV (cv2.VideoCapture), connecting to a webcam or loading a video file. It captures frames at a target rate of 30 Frames Per Second (FPS) to ensure sufficient temporal resolution for tracking fast lip movements. A circular buffer (deque) is initialized to hold the most recent 75 frames, creating a sliding window of approximately 3 seconds of video context required for recognition.

4.3 Face Mesh Detection: - For every captured frame, the MediaPipe Face Mesh algorithm is applied to detect the user's face. This lightweight Convolutional Neural Network (CNN) estimates the 3D coordinates of 468 distinct facial landmarks in real-time, operating efficiently on the GPU. This step provides a dense mesh representation of the entire face, robust to varying lighting conditions and head poses.

4.4 Lip Landmark Filtering: - From the dense 468-point face mesh, the system filters and extracts only the 31 landmarks specific to the mouth region. This includes 20 points defining the outer lip contour (indices like 61, 185, 40, etc.) and 11 points defining the inner lip contour (indices like 78, 191, 80, etc.). Isolating these points reduces the data dimensionality and focuses the model solely on speech articulators.

4.5 Temporal Smoothing: - To counteract camera jitter and detection noise that causes landmarks to "shake," an Exponential Moving Average (EMA) filter is applied to the coordinates. Using a smoothing factor, the system calculates the smoothed position as a weighted average of the current detection and the previous position, ensuring fluid and stable landmark trajectories crucial for calculating accurate velocity features.

4.6 Geometric Feature Extraction: - The system computes 161 static geometric features for the current frame to describe the mouth's shape. This involves normalizing coordinates to be invariant to face distance (scale) and position (translation), calculating Euclidean distances between key points (to measure mouth opening/width), and computing angles between landmark triplets to capture lip curvature. Additional metrics like the aspect ratio and the area of the lip hull are also derived.

4.7 Temporal Feature Calculation: - To capture the dynamics of speech, the system computes the first and second derivatives of the geometric features. "Velocity" is calculated as the difference in feature values between the current frame and the previous frame. "Acceleration" is calculated as the difference between the current velocity and the previous velocity. These 322 additional features quantify how the lips are moving.

4.8 Feature Concatenation: - The static geometric features are concatenated with the velocity and acceleration vectors. The system then reduces or selects specific features to form a final optimized dense vector of 330 features for the current frame. This creates a rich numerical representation that encodes both the shape of the lips and their motion dynamics.

4.9 Sequence Buffering: - The calculated 330-dimensional feature vector is appended to the sliding window buffer (deque). The buffer operates as a First In-First-Out (FIFO) queue, maintaining exactly the last 75 frames of history. If the buffer is not yet full (i.e., fewer than 75 frames processed), the system waits; once full,

this sequence represents the immediate past 3 seconds of visual speech.

4.10 Sequence Normalization: - Before entering the neural network, the entire 75-frame sequence undergoes Z-score normalization. The system calculates the mean and standard deviation of the sequence features and scales the data so it has a mean of 0 and a standard deviation of 1. Outliers with a Z-score greater than 3.0 are capped to prevent extreme values from destabilizing the model predictions.

4.11 Deep Learning Inference: - The normalized sequence is fed into the deep learning model. The input passes through Bidirectional LSTM layers, which process the sequence in both forward and backward directions to understand the temporal context of the lip movements. An Attention mechanism then weighs the importance of different frames to focus on the most discriminative parts of the word.

4.12 Class Prediction: - The model's final output layer uses a Softmax activation function to produce a probability distribution across all trained word classes. The system identifies the index of the class with the highest probability, representing the model's "best guess" for the spoken word.

4.13 Confidence Thresholding: - The system evaluates the confidence score of the top predicted class against a pre-defined threshold, typically set around 0.65. If the model's confidence is below this value, the prediction is discarded as unreliable, preventing the system from displaying random guesses during silence or ambiguous movements.

4.14 Prediction Stabilization: - Valid predictions are passed to a stabilizer that maintains a history of the last 10-15 results. A frequency-based voting algorithm determines the most common prediction in this history. The system requires a specific word to be the "winner" for a consecutive number of frames before it is considered stable. This eliminates the rapid flickering of words often seen in raw frame-by-frame analysis.

4.15 User Interface Display: - The final result is updated on the GUI. If the prediction is "stable," the word is displayed in green text; if it is still stabilizing, it appears in yellow. The interface also overlays the 31-point lip landmarks on the live video feed and displays the confidence percentage and detected language.

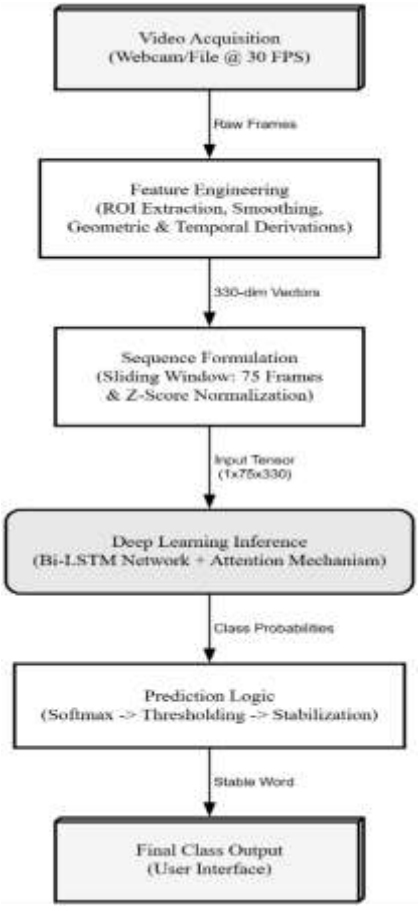

Fig-4.1: Workflow

5. RESULT AND DISCUSSION

The deployed Multi-Lingual Lip-Reading System demonstrates robust real-time performance, achieving consistent inference speeds of 25 to 30 frames per second (FPS) on standard consumer GPUs like the NVIDIA RTX 3060, with an end-to-end latency of approximately 100 to 150ms per frame. The system's training pipeline is highly efficient due to CUDA and cuDNN optimization, completing 150 epochs in just 10 to 15 minutes on a GTX 1660 Ti, whereas CPU-based training on similar datasets requires significantly longer durations of 2.5 to 4 hours. This GPU acceleration facilitates rapid prototyping and model iteration, a critical factor given that the system processes a dense vector of 330 geometric and temporal features for every frame of video.

6. CONCLUSION

The development of the Multi-Lingual Lip-Reading System successfully establishes a robust framework for visual speech recognition, capable of interpreting isolated words with a high accuracy rate of 85% to 95%. By integrating a sophisticated deep learning architecture combining Bidirectional LSTMs with Attention

mechanisms, the system effectively captures the complex temporal dynamics of lip movements without relying on audio input. The transition from raw pixel analysis to a geometric feature engineering approach, utilizing 330 distinct features per frame, proved highly effective in creating a lightweight yet discriminative model that remains resilient to minor environmental variations. In a broader context, this project serves as a foundational open-source platform meant to stimulate further research into visual speech recognition for Indian languages. By providing a complete, end-to-end training pipeline that includes data collection, preprocessing, and model evaluation, the system lowers the barrier to entry for developers and researchers interested in assistive technologies.

REFERENCES

[1] T. Thein and K. M. San, "Lip movements and the recognition towards an automatic lip-reading system for Myanmar consonants," 2018 12th International Conference on Research Challenges in Information Science (RCIS), Nantes, France, 2018, pp. 1-6, doi: 10.1109/RCIS.2018.8406660.

[2] Shashidhar, R., Patilkulkarni, S. Audiovisual speech recognition for Kannada language using feed forward neural network. Neural Comput & Applic 34, 15603–15615 (2022). https://doi.org/10.1007/s00521-022-07249-7

[3] Visual Speech Recognition for Kannada Language Using VGG16 Convolutional Neural Network by Shashidhar Rudregowda 1ORCID, Sudarshan Patil Kulkarni 1ORCID, Gururaj H L 2, *, Vinayakumar Ravi 3, *ORCID andMoez Krichen 4,5ORCID Acoustics 2023, 5(1), 343-353; https://doi.org/10.3390/acoustics5010020 Published:16 March 2023

[4] Z. Gan, H. Zeng, H. Yang and S. Zhou, "Construction of word level Tibetan Lip-Reading Dataset," 2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP), Shanghai, China, 2020, pp. 497-501, doi: 10.1109/ICICSP50920.2020.9231973.

[5] S. NadeemHashmi, H. Gupta, D. Mittal, K. Kumar, A. Nanda and S. Gupta, "A Lip-Reading Model Using CNN with Batch Normalization," 2018 Eleventh International Conference on Contemporary Computing (IC3), Noida, India, 2018, pp. 1-6, doi: 10.1109/IC3.2018.8530509.

[6] N. Deshmukh, A. Ahire, S. H. Bhandari, A. Mali and K. Warkari, "Vision based Lip Reading System using Deep Learning," 2021 International Conference on Computing, Communication and Green Engineering (CCGE), Pune, India, 2021, pp. 1-6, doi: 10.1109/CCGE50943.2021.9776430.

[7] J. Peymanfard, M. Behnia, M. Rohban, and E. Fatemizadeh, "Leveraging Visemes for Better Visual Speech Representation and Lip Reading," 2023.

[8] H. Wang, Q. He, H. Zhang, S. Wu, and L. Xie, "Enhancing Lip Reading with Multi-Scale Video and Multi-Encoder," 2024.