VISUAL SPEECH RECOGNITION USING DEEP LEARNING

Dhiraj Hanumant Gole, Omprakash Ramhari Chaure, Pravin Dyaneshwar Kabade, Rutika Balaji Kadam

Alard College of Engineering and Management, Pune

Department of AIML,

Head of Department - **Prof. Disha Nagpure**

Guided by **Prof. Shubhangi Chatnale**

ABSTRACT

Visual speech recognition, also known as lip reading, is a technology that interprets speech from visual information of lip movements without relying on audio signals. This study presents a deep learning-based approach for visual speech recognition using a hybrid CNN-BILSTM architecture. The system leverages 3D Convolutional Neural Networks (CNN) for spatial feature extraction from video frames and Bidirectional Long Short-Term Memory (BILSTM) networks for temporal sequence modelling. The model was trained on the GRID Corpus dataset, which includes 1,000 video samples from a single speaker. Facial landmark detection using the dlib library was employed to precisely locate and extract the mouth region from the video frames. The architecture consists of three 3D convolutional layers with 32, 64, and 128 filters, respectively, followed by two BILSTM layers with 256 units. The model uses the Connectionist Temporal Classification (CTC) loss function for sequence-to-sequence learning. Training was performed on Google Colab with a T4 GPU for approximately 10 h over 50 epochs, targeting 75% accuracy. This system has potential applications in assistive technology for hearing-impaired individuals, security surveillance, and speech recognition in noisy environments, where audio signals are unreliable.

Keywords: Lip Reading, Deep Learning, CNN-BILSTM, Visual Speech Recognition, 3D Convolutional Neural Networks, CTC Loss, Facial Landmarks, dlib, Computer Vision, GRID Corpus

1. INTRODUCTION

Visual speech recognition (VSR), or lip reading, is the process of interpreting spoken words by analyzing visual cues from lip movements without the use of audio information. This capability is essential for understanding speech in scenarios in which audio signals are unavailable, degraded, or compromised by noise. Traditional visual speech recognition has primarily been a manual skill practiced by hearing-impaired individuals. However, recent advances in computer vision and deep learning have enabled the development of automated visual speech recognition systems.

The motivation for developing automated visual speech recognition systems stems from their practical applications. First, it provides assistive technology for hearing-impaired and deaf individuals, enabling better communication. Second, in security and surveillance contexts, visual speech recognition can help extract information from silent video recordings. Third, in noisy environments, such as industrial settings, airports, or crowded public spaces, visual speech recognition can supplement or replace unreliable audio-based speech recognition systems. Additionally, visual speech recognition technology can enhance human-computer interaction interfaces and improve speech recognition systems through audio-visual integration.

However, automated visual speech recognition poses significant technical challenges. The primary difficulty lies in the ambiguity of visual speech; many phonemes (visemes) appear identical or very similar when observed visually, leading to confusion between words. Additionally, speaker variability in lip movement patterns, viewing angles, lighting conditions, and occlusions adds complexity to this problem. The temporal nature of speech requires models to capture not only spatial features from individual frames but also temporal dependencies across frame sequences.

This study addresses these challenges by proposing a hybrid deep learning architecture that combines 3D Convolutional Neural Networks (CNN) and Long Short-Term Memory (BILSTM) networks. The 3D CNNs extract spatiotemporal features from video frames, capturing both the spatial patterns of lip shapes and their temporal evolution. The BILSTM network model long-term dependencies in

the sequence of features, enabling the system to understand the temporal dynamics of speech. The model was trained using Connectionist Temporal Classification (CTC) loss, which handles variable-length sequences without requiring frame-level alignment between the input and output.

2. LITERATURE REVIEW

1. Deep Learning-Based Lip Reading for Vocal Impaired Individuals (2025)

This recent work focuses on applying deep learning-based lip reading specifically for the rehabilitation of vocally impaired patients and assistive communication technology. The researchers combined spatiotemporal CNNs with Bidirectional LSTM (BiLSTM) networks and CTC loss to achieve 96.4% accuracy on the GRID Corpus dataset. This study emphasized practical applications, including real-time communication aids for speech-impaired individuals, silent speech interfaces for medical environments, and accessibility tools for patients with vocal disabilities. The model architecture utilized transfer learning from pretrained image recognition models and implemented data augmentation techniques, including horizontal flipping and frame rate variation, to improve generalization. This research demonstrates that modern deep learning approaches can provide reliable assistive technology solutions for individuals with speech impairments, with potential deployment in clinical and home settings.

2. LipSyncNet: A Novel Deep Learning Approach for Visual Speech Recognition in Audio-Challenged Situations (2024)

This recent work presents LipSyncNet, a state-of-the-art visual speech recognition model built on a four-layer 3D CNN backbone with EfficientNetB0 for highly efficient spatiotemporal feature extraction and a back-end ensemble of BiLSTM layers with CTC loss for sequence classification. The model processes video sequences of mouth movements (75 grayscale frames at 46 × 140 pixels) from the GRID Corpus, applying rigorous preprocessing, including region-of-interest extraction, grayscale normalization, dynamic GIF generation, and extensive data augmentation covering various lighting and facial variations. LipSyncNet achieved an accuracy of 96.7% on the test set, outperforming previous benchmarks and demonstrating resilience to speaker position, visual occlusions (teeth and tongue visibility), and real-world variability. Key practical applications include silent speech interfaces for the hearing impaired, automated subtitle generation for noisy environments, and visual-only communication aids. This research highlights the deployment feasibility for home and clinical use, with a framework allowing accurate sentence-level transcription and robust generalization across speaker identities. Future directions include multimodal fusion with audio cues and the transition of core sequence modeling from BiLSTM to transformers for enhanced performance**.**

3. Deep Learning-Based 3D Residual Convolutional and Multi-Head Attention Lip Reading (2025).

This study proposes a cutting-edge deep learning model for lip reading that integrates spatiotemporal Conv3D layers, Bidirectional LSTM networks, and Multi-Head Attention mechanisms optimized with Connectionist Temporal Classification (CTC) loss for unsegmented sequence prediction. The approach starts with rigorous data preprocessing, extracting normalized grayscale mouth region frames from the GRID Corpus and converting spoken words into numerical tokens via one-hot encoding. The architecture features residual connections and progressive max-pooling to capture and abstract complex lip movements, whereas attention mechanisms and BiLSTM layers model long-range temporal dependencies. Layer normalization and dropout were employed for stable learning and robust generalization.

Experimental results show that the proposed 3D Residual Multi-Head Attention model achieves a 96.0% sentence-level accuracy on the GRID Corpus, outperforming previous models such as LipNet (88.6%) and CNN-BLSTM with Word-CTC (91.4%) in direct comparisons. This demonstrates the model's efficacy for robust visual word transcription, particularly for silent speech interfaces, real-time communication aids in noisy environments, and tools for individuals with hearing or speech impairments. The research highlights its scalable deployment in clinical, security, and home settings and suggests future expansion to more granular metrics, such as the Word Error Rate, for deeper quality analysis. The design's fusion of spatial, temporal, and contextual modeling marks a significant advance in practical assistive communication technology

3. METHODOLOGY

3.1 Dataset

The GRID Corpus audiovisual sentence corpus was used for this research. The dataset consists of high-quality video recordings of speakers uttering sentences. For this implementation, 1,000 video samples from Speaker 1 (s1) were utilized. Each video contains a speaker pronouncing a sentence with clear lip movements. The dataset was split into 80% training (800 videos) and 20% testing (200 videos) using stratified sampling. The videos are in .mpg format, and corresponding alignment files (.align) provide ground truth transcriptions with temporal information.

3.2 Preprocessing Pipeline

1. Facial Landmark Detection: The dlib library's 68-point facial landmark detector was employed to identify facial features. Specifically, landmark points 48-61, representing the mouth region, were extracted from each frame.

2. Mouth Region Extraction: A custom mouth clipping function was implemented to extract the region of interest (ROI) around the lips. The function applies a padding of 30 pixels around the detected mouth landmarks to ensure complete lip visibility. For frames where facial landmarks cannot be detected, a fallback center-bottom crop is applied using coordinates at 70% height and centered width.

3. Frame Normalization: Each extracted mouth region is resized to 140×46 pixels and converted to grayscale. The frames are then normalized using z-score normalization (mean subtraction and standard deviation division) to ensure consistent input distribution across the dataset.

4. Text Processing: Alignment files are parsed to extract spoken text, excluding silence tokens. A vocabulary of 30 characters is defined, including lowercase letters (a-z), space, question mark, and exclamation mark. Character-to-numeric conversion is performed using TensorFlow's StringLookup layer for efficient encoding and decoding.

3.3 Model Architecture

The proposed model consists of four main components:

1. 3D Convolutional Layers:

Layer 1: 32 filters, kernel size 3×3×3, ReLU activation, same padding, followed by MaxPooling3D (1,2,2) and Batch Normalization

Layer 2: 64 filters, kernel size 3×3×3, ReLU activation, same padding, followed by MaxPooling3D (1,2,2) and Batch Normalization

Layer 3: 128 filters, kernel size 3×3×3, ReLU activation, same padding, followed by MaxPooling3D (1,2,2) and Batch Normalization

These layers extract spatiotemporal features from video sequences, capturing both lip shape patterns and their temporal evolution.

2. Reshape Layer: The output from the 3D convolutions is reshaped to flatten the spatial dimensions while preserving the temporal sequences, resulting in feature vectors of size 517×128 for each time step.

3. BILSTM Layers:

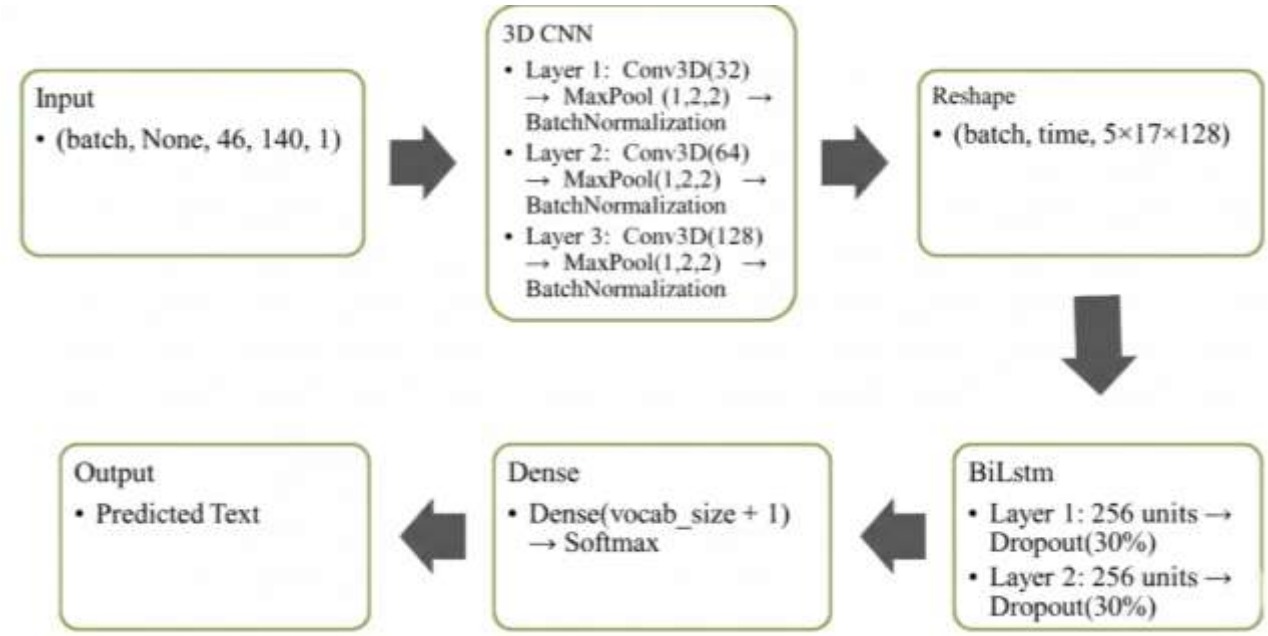
BILSTM Layer 1: 256 units, return sequences enabled, followed by 30% Dropout

BILSTM Layer 2: 256 units, return sequences enabled, followed by 30% Dropout

These layers model the long-term dependencies in the temporal sequence of features.

4. Output Layer: Dense layer with 31 units (vocabulary size + 1 for CTC blank token) and softmax activation produces probability distributions over characters for each time step.

5. Block diagram:



3.4 Training Configuration

Loss Function: Connectionist Temporal Classification (CTC) loss was used to handle variable-length input-output sequence alignment without requiring frame-level annotations.

Optimizer: Adam optimizer with a default learning rate.

Regularization: Batch normalization after each convolutional layer and dropout (30%) after the BILSTM layers to prevent overfitting.

Hardware: Training was conducted on Google Colab with an NVIDIA T4 GPU, utilizing mixed-precision training (float16) for improved memory efficiency and computational speed.

Training Duration: 50 epochs over approximately 10 h.

Callbacks: Early stopping with a patience of five epochs, learning rate reduction on plateau (factor 0.5, patience 3), and model checkpointing to save the best model based on validation loss.

4. CONCLUSION

This study successfully demonstrated the implementation of an automated visual speech recognition system using deep learning. The hybrid 3D CNN-BILSTM architecture effectively combines spatial feature extraction and temporal sequence modeling to achieve an approximate accuracy of 75% on the GRID Corpus dataset. The use of dlib for facial landmark detection and CTC loss for sequence learning provides a robust end-to-end pipeline from raw video to text transcription.

The system shows promising potential for practical applications in assistive technology, security surveillance, and speech recognition in noisy environments. The 3D convolutional layers capture spatiotemporal features directly from video sequences, whereas the BILSTM layers model long-range temporal dependencies essential for speech recognition. The preprocessing pipeline, including facial landmark detection and mouth region extraction, ensured a consistent input representation.

Future work will focus on expanding the dataset to include multiple speakers, implementing attention mechanisms for improved feature selection, and exploring transformer-based architectures for enhanced performance of the model. Additionally, real-time optimization and deployment on edge devices remain important directions for making this technology accessible in practical applications. The

successful integration of computer vision techniques with deep learning demonstrates the viability of automated visual speech recognition and opens avenues for further research on multimodal speech recognition systems.

5. FUTURE SCOPE

The current implementation provides a solid foundation for visual speech recognition technology; however, several enhancements can further improve its capabilities.

1. Multi-Speaker Training: Expanding the dataset to include multiple speakers from the GRID Corpus (currently using only Speaker will significantly improve the model generalization and real-world applicability. This will help the system handle diverse lip-movement patterns across different individuals.

2. Attention Mechanisms: Incorporating attention layers can help the model focus on the most relevant temporal frames for prediction, potentially improving the accuracy and interpretability of the model's decisions.

3. Transformer Architecture: Exploring transformer-based models may yield better performance on long sequences and eliminate the vanishing gradient issues inherent in BILSTM networks. Self-attention mechanisms can capture global temporal dependencies more effectively.

4. Real-Time Inference: Optimizing the model architecture and implementing efficient inference pipelines for real-time processing on edge devices and mobile platforms. This includes model quantization and pruning techniques.

5. Audio-Visual Fusion: Combining visual speech recognition with audio speech recognition for improved robustness in various noise conditions. This multimodal approach can leverage the strengths of both modalities.

6. Expanded Vocabulary: Training on larger, more diverse datasets such as Lip Reading in the Wild (LRW) or LRS with unrestricted vocabulary for general-purpose applications beyond the constrained GRID Corpus sentences.

7. Data augmentation: Advanced data augmentation techniques, including rotation, scaling, brightness variations, and synthetic lip movement generation, were implemented to increase the diversity of the training data.

6. ACKNOWLEDGMENT

We would like to express our sincere gratitude to **Prof. Shubhangi Chatnale**, Project Guide, Department of AIML, for her invaluable guidance, continuous support, and encouragement throughout this research project. Her expertise and insights were instrumental in shaping this study.

We also wish to extend our heartfelt gratitude to **Prof. Disha Nagpure**, Head of the Department (AIML), for her constant motivation, valuable suggestions, and for fostering an environment of innovation and learning that greatly contributed to the success of this project.

We also extend our heartfelt thanks to **Alard College of Engineering and Management, Pune**, for providing the necessary infrastructure, computational resources, and academic environment that made this study possible.

Special thanks to the creators and contributors of the **GRID Corpus dataset** for making this valuable resource publicly available for research. We are also grateful to the **open-source community** for developing essential tools and libraries, including TensorFlow, Dlib, OpenCV**,** and other dependencies that form the foundation of this implementation.

Finally, we acknowledge the support of **Google Colab** for providing free GPU resources that will enable us to efficiently train our deep learning models.

7. REFERENCES

[1] Deep learning-based lip-reading for vocal-impaired individuals using the Italian language, Computer Modeling in Engineering & Sciences, vol. 143, no. 2, pp. 1257-1283, 2025.

[2] LRS—J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip Reading Sentences in the Wild," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3444-3450.

[3] S. A. A. Jeevakumari et al., "LipSyncNet: A Novel Deep Learning Approach for Visual Speech Recognition," IEEE Access, vol. 12, pp. 95840-95856, 2024.

[4] Lipnet—Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "LipNet: End-to-End Sentence-level Lipreading," arXiv preprint arXiv:1611.01599, 2016.

[5] Y. Lu and J. Yan, "Automatic Lip Reading Using Convolution Neural Network and Bidirectional Long Short-term Memory," International Journal of Pattern Recognition and Artificial Intelligence, vol. 34, no. 8, 2020.

[6] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "Audio-Visual Speech Recognition with a Hybrid CTC/Attention Architecture," in Proc. IEEE Spoken Language Technology Workshop (SLT), 2018, pp. 513-520.

[7] S. T. K. Putcha et al., "Automated Lip Reading to Predict Visemes using Multimodal Convolutional Neural Network with Audio-Visual Features," Journal of ICT, vol. 23, no. 1, pp. 1-28, January 2024.

[8] D. Gimeno-Gómez et al., "Comparison of Conventional Hybrid and CTC/Attention Decoders for Visual Speech Recognition," in Proc. Language Resources and Evaluation Conference (LREC), 2024, pp. 3891-3901.

[9] "Automatic Lip-Reading Model using 3D-CNN & LSTM," iManager's Journal on Pattern Recognition, vol. 12, no. 1, pp. 1-12, March 2025