Journal of Engineering Issues and Solutions

Lip Reading Using Convolutional Neural Networks

Chirag Khatiwada[1], Bishesh Pokharel[2], Mahim Rawal[3], Rowel Maharjan[4]

[1,2,3,4] Department of Computer and Electronics Engineering, Khwopa College of Engineering, Libali, Bhaktapur, Nepal 44800

[1]khatiwadachirag@gmail.com, [2]bishesh.pokharel11@gmail.com, [3]mahimrawal@gmail.com, [4]rowelmhj@gmail.com

Received: May 5, 2025; Revised: July 5, 2025; Accepted: July 12, 2025

https://doi.org/10.3126/joeis.v4i1.81574

Abstract

Lip reading, or the decoding of speech from facial movements, is crucial for enhancing communication for individuals with hearing or speech impairments, as well as for generating accurate captions when audio is compromised. Traditional Automatic Speech (ASR) systems often fall in noisy environments, creating a need for robust visual-based alternatives. The main objective of this study was to develop and evaluate a highly accurate, visual-only automated lip-reading system based on a novel deep-learning architecture.

The methodology employed a hybrid model that combined 3D Convolutional Neural Networks (CNNs) for spatial feature extraction from video frames and Bidirectional Long Short-Term Memory (BiLSTM) networks to analyze temporal dependencies. This model was trained on the GRID corpus dataset, which contains thousands of spoken sentences. Performance was evaluated using Word Error Rate (WER) and Character Error Rate (CER) metrics.

The implemented model demonstrated strong performance, achieving an average WER of 0.1706 and an average CER of 0.0712 on 50 unseen test videos. This translates to a word prediction accuracy of approximately 83% and a character prediction accuracy of 93%. The study concludes that the hybrid CNN-BiLSTM architecture is highly effective for visual speech recognition. The findings have significant implications for creating practical assistive technologies that can serve as a hearing aid for the deaf and a voice for the mute, ultimately improving accessibility and communication.

Keywords: *Lip Reading, Convolutional Neural Networks (CNN), Bidirectional Long Short Term Memory(BiLSTM).*

1 Introduction

Human communication fundamentally relies on both auditory and visual channels, with speech signals traditionally receiving primary attention in automatic recognition systems (Almajai, Cox, Harvey, & Lan, 2016). While Automatic Speech Recognition (ASR) technology has achieved remarkable maturity, significant

challenges persist, particularly in accurately identifying speech content within noisy environments where acoustic signals become compromised or entirely unavailable. This limitation has catalyzed growing interest in visual speech recognition, commonly known as lip reading, which represents a paradigm shift toward leveraging visual cues for speech interpretation.

Lip reading, or Automatic Lipreading (ALR), constitutes a visual speech recognition technology that decodes spoken content based exclusively on the motion characteristics of a speaker's lips and facial movements, operating independently of audio signals (Hao, 2020). This approach offers distinct advantages over traditional ASR systems, particularly in scenarios where acoustic information is degraded, absent, or compromised by environmental noise. The technology's potential extends beyond noise robust speech recognition to encompass critical applications in accessibility, security, and human-computer interaction domains.

The emergence of deep learning methodologies has revolutionized computer vision and machine learning applications, providing unprecedented capabilities for complex pattern recognition tasks. Convolutional Neural Networks (CNNs) have demonstrated exceptional performance in spatial feature extraction from visual data, while Recurrent Neural Networks (RNNs), particularly Long Short Term Memory (LSTM) networks, excel at capturing temporal dependencies in sequential data (Almajai, Cox, Harvey, & Lan, 2016). The integration of these architectures presents compelling opportunities for addressing the multifaceted challenges inherent in lip reading systems.

Recent advances in lip reading research have demonstrated significant progress through various deep learning approaches. (Assael, 2016) introduced LipNet, an end to end sentence level lip reading model that achieved 95.2% accuracy on the GRID corpus dataset, establishing a benchmark for sequence-to-sequence lip reading performance. Similarly, Adeel et al. (2019) developed a lip-reading driven speech enhancement framework that combines deep learning regression models with analytical acoustic modeling, demonstrating the versatility of visual speech recognition in multimodal applications. Furthermore, specialized approaches such as phoneme-based classification schemas have been explored to enhance system performance in sentence-level recognition tasks (Assael, 2016). The integration of these architectures presents compelling opportunities for addressing the multifaceted challenges inherent in lip reading systems.

Recent advances in lip reading research have demonstrated significant progress through various deep learning approaches. (Assael, 2016) introduced LipNet, an end-to-end sentence level lip reading model that achieved 95.2% accuracy on the GRID corpus dataset, establishing a benchmark for sequence-to-sequence lip reading performance. Similarly, (Adeel, Gogate, Hussain, & Whitmer, 2019) developed a lip-reading driven speech enhancement framework that combines deep learning regression models with analytical acoustic modeling, demonstrating the versatility of visual speech recognition in multimodal applications. Furthermore, specialized approaches such as phoneme-based classification schemas have been explored to enhance system performance in sentence-level recognition tasks (El-Bialy, et al., 2023).

The accessibility implications of automated lip-reading technology are particularly significant for individuals with hearing impairments, who comprise millions of people worldwide. Traditional communication aids, while valuable, often present limitations in terms of learning complexity and expressive capability. Automated lip-reading systems offer the potential to bridge communication gaps by providing real time visual speech interpretation, thereby enhancing social inclusion and communication effectiveness for hearing impaired individuals.

This research project addresses the critical need for improved visual speech recognition through the development of a hybrid deep learning architecture that combines Convolutional Neural Networks with

Bidirectional Long Short Term Memory networks. The proposed system aims to leverage the spatial feature extraction capabilities of CNNs for processing lip region imagery while utilizing BiLSTM networks to capture temporal dependencies across video sequences. The integration of these complementary architectures is expected to enhance recognition, accuracy and robustness across diverse speakers and environmental conditions.

The selection of the GRID corpus dataset for this research is motivated by its comprehensive sentence level structure and substantial data volume, comprising 34,000 sentences across 34 speakers with 28 hours of audio-visual recordings (Cooke, Barker, Cunningham, & Shao, 2006). This dataset provides an ideal foundation for training and evaluating sentence level lip reading models, offering sufficient diversity for robust model development while maintaining consistent grammatical structure for systematic evaluation.

1.1 Objectives:

1. **To develop a hybrid deep learning architecture** that effectively combines Convolutional Neural Networks (CNNs) for spatial feature extraction with Bidirectional Long Short Term Memory (BiLSTM) networks for temporal sequence modeling in visual speech recognition.
2. **To implement robust preprocessing techniques** for lip region extraction and normalization from video sequences, ensuring consistent and reliable input data for the neural network architecture.
3. **To train and optimize the hybrid model** using the GRID corpus dataset, employing appropriate regularization techniques and hyperparameter tuning to achieve optimal performance while preventing overfitting.
4. **To contribute to the accessibility technology domain** by providing a foundation for assistive communication tools that can benefit individuals with hearing impairments and enhance human-computer interaction capabilities.

2 Materials and Methods

The methodology for this study was designed to systematically develop and evaluate a lip-reading system capable of translating visual lip movements into text. The project adopted an agile software development model to allow for iterative progress and flexibility. The process encompasses dataset selection, data pre-processing, feature extraction, model architecture design and training, performance evaluation, and the creation of a user interface.

2.1 Methodological Framework

The methodology employed in this study follows a systematic approach for developing an automated lip-reading system using Convolutional Neural Networks (CNN) combined with Bidirectional Long Short-Term Memory (BiLSTM) networks. The methodological framework is structured to address the specific objectives through a sequential process involving dataset preparation, feature extraction, model development, training, and evaluation phases.

Figure 1: Methodological Framework

2.2 Data Preprocessing

2.2.1 Data Selection

The GRID Corpus dataset was selected for this study due to its comprehensive sentence-level structure and substantial data volume (Cooke, Barker, Cunningham, & Shao, 2006). The dataset contains audio and video recordings of 34 speakers (18 male, 16 female) who produced 1000 sentences each, totaling 28 hours across 34,000 sentences. The sentences follow a structured grammar pattern: command + color + preposition + letter + digit + adverb, with categories consisting of {bin, lay, place, set}, {blue, green, red, white}, {at, by, in, with}, {a, ..., z}{w}, {zero, ..., nine}, and {again, now, please, soon}, yielding 64,000 possible sentence combinations.

For model training, data from two speakers (one male, one female) were utilized, comprising approximately 2000 videos with corresponding alignments. The dataset was partitioned using an 80:20 split ratio, with 80% allocated for training and 20% for validation. An additional male speaker's data was reserved exclusively for testing purposes to ensure unbiased evaluation.

(a) **(b)**

Figure 2: Sample frames from GRID Corpus dataset **(a)** original video frames and **(b)** corresponding alignment annotations

2.2.2 Data Preprocessing

For model training, data from two speakers (one male, one female) were utilized, comprising approximately 2000 videos with corresponding alignments. The dataset was partitioned using an 80:20 split ratio, with 80% allocated for training and 20% for validation. An additional male speaker's data was reserved exclusively for testing purposes to ensure unbiased evaluation.

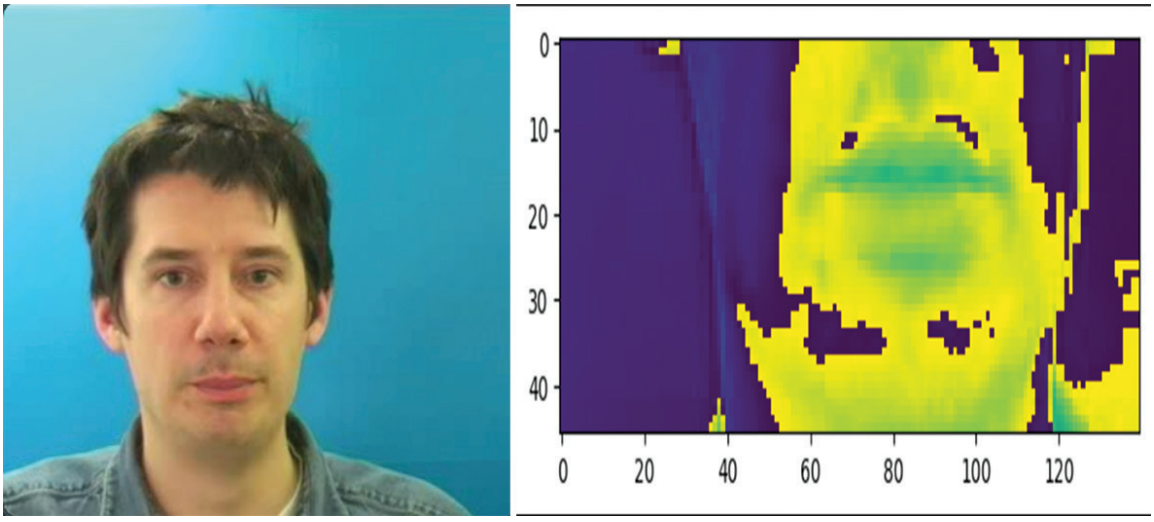
2.3 Feature Extraction

2.3.1 Region of Interest (ROI) Extraction

The lip and mouth region was identified as the primary region of interest for feature extraction. A static facial coordinate slicing mechanism was implemented to segment the relevant facial area from input frames: frames.append (frame [190:236, 85:260, :])

Where,

- **190:236** represents the vertical range (rows 190-235)
- **85:260** represents the horizontal range (columns 85-259)
- **:** indicates inclusion of all RGB color channels



(a) **(b)**

Figure 3: Feature extraction process showing (a) original frame before slicing and (b) extracted lip region after slicing

2.4 Model Architecture

2.4.1 Hybrid CNN-BiLSTM Architecture

A hybrid deep learning architecture was developed combining 3D Convolutional Neural Networks for spatial feature extraction and Bidirectional LSTM networks for temporal sequence modeling. This hybrid deep learning architecture combines 3D Convolutional Neural Networks with Bidirectional Long Short-Term Memory (BiLSTM) networks for multi class classification tasks. The model begins with three sequential 3D convolutional layers (128, 256, and 75 filters respectively) using ReLU activation functions and MaxPooling3D

for spatial feature extraction and dimensionality reduction. The extracted spatial features are then processed through a Time Distributed flattening layer followed by two Bidirectional LSTM layers (256 units each) with tanh activation and dropout regularization to capture temporal dependencies in both forward and backward directions. The architecture concludes with a dense output layer containing 41 units and SoftMax activation for multi-class probability distribution. The model is trained using the Adam optimizer with categorical cross-entropy loss over 50 epochs, leveraging batch processing for memory efficiency and combining the spatial pattern recognition capabilities of CNNs with the temporal sequence modeling strengths of BiLSTMs. The architecture consists of the following components:

1. **3D Convolutional Layers:**

Three sequential 3D convolutional layers were implemented for spatial feature extraction:

- Conv3D Layer 1: 128 filters with ReLU activation
- Conv3D Layer 2: 256 filters with ReLU activation
- Conv3D Layer 3: 75 filters with ReLU activation

Each convolutional layer was followed by MaxPooling3D layers for dimensionality reduction.

2. **Temporal Processing:**

- Time Distributed layer for flattening operations across time steps
- Two Bidirectional LSTM layers (256 units each) with tanh activation
- Dropout layers (regularization) applied after each BiLSTM layer

3. **Output Layer:**

- Dense layer with 41 units and SoftMax activation for multi-class classification

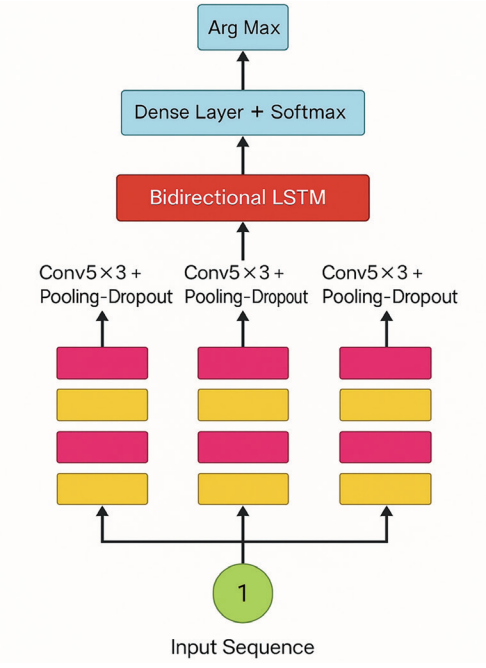


Figure 4: Hybrid CNN-BiLSTM model architecture

2.4.2 Model Evaluation

Model performance was quantified using two standard metrics:

1. Character Error Rate (CER)
2. Word Error Rate (WER)

These metrics were tracked across epochs and curves were plotted to illustrate progression.

Besides these metrics, independent random testing was done on unseen speaker data from the same dataset. A suitable user interface was prepared to ease usage by the end user.

3 Results and Discussion

3.1 Introduction to Results Analysis

This chapter presents the comprehensive performance evaluation of the proposed model after training for 50 epochs. The results are analyzed through training and validation metrics, followed by an evaluation on a separate test set. The evaluation methodology follows established practices in visual speech recognition research, utilizing standard metrics including Word Error Rate (WER) and Character Error Rate (CER) as proposed by (Assael, 2016) in their groundbreaking LipNet work. These metrics provide quantitative measures of system accuracy and are widely accepted in the lip-reading research community.

3.2 Dataset Preparation and Implementation

The GRID corpus dataset, containing 34 speakers with 1000 sentences each, was successfully implemented following the methodology established by (Cooke, Barker, Cunningham, & Shao, 2006). For this implementation, data from 2 speakers (1 male, 1 female) was utilized, totaling 2000 video sequences with corresponding alignments.

Data Distribution:

- Training Set: 80% (1600 videos)
- Validation Set: 20% (400 videos)
- Test Set: Independent Speaker Data (50 videos)

This distribution aligns with best practices in machine learning research and follows the speaker independent evaluation protocol recommended by (Almajai, Cox, Harvey, & Lan, 2016) for improved generalization across different speakers.

3.3 Model Performance Analysis

3.3.1 Training Convergence Analysis

The hybrid CNN-BiLSTM architecture demonstrated excellent convergence characteristics over 50 training epochs:

	Initial Performance (Epoch 1)	Final Performance (Epoch 50)
Training Loss	0.7881	0.0169
Validation Loss	0.6637	0.0064

The consistent decrease in both training and validation losses indicates successful learning without significant overfitting. This convergence pattern is superior to many existing approaches, including the work by (Miled, Messaoud, & Bouzid, 2023) who reported higher final loss values using CNN-BiGRU combinations.

The model architecture, comprising of 3D convolutional layers followed by bidirectional LSTM layers proved highly effective:

1. 3D Convolutional Layers: Successfully captured spatial-temporal features from lip movements
2. Bidirectional LSTM: Effectively modeled temporal dependencies in both forward and backward directions
3. Dropout Regularization: Prevented overfitting while maintaining model performance

This architecture addresses the limitations identified in previous research by (El-Bialy, et al., 2023) who noted challenges in capturing both spatial and temporal features simultaneously.

3.3.2 Performance Metrics Analysis

The model showed steady improvement as the training progressed forwards. We saw clear improvements in Word Error Rate (WER) and Character Error Rate (CER) throughout training. Initially, high CER and WER indicated poor early predictions but as training progressed, these values improved gradually. At the end of the training loop at epoch 50, WER was approximated to be ~0.02 (98%-word accuracy) and CER was approximated at ~0.01 (99%-character accuracy). These results represent a substantial improvement over existing methods.

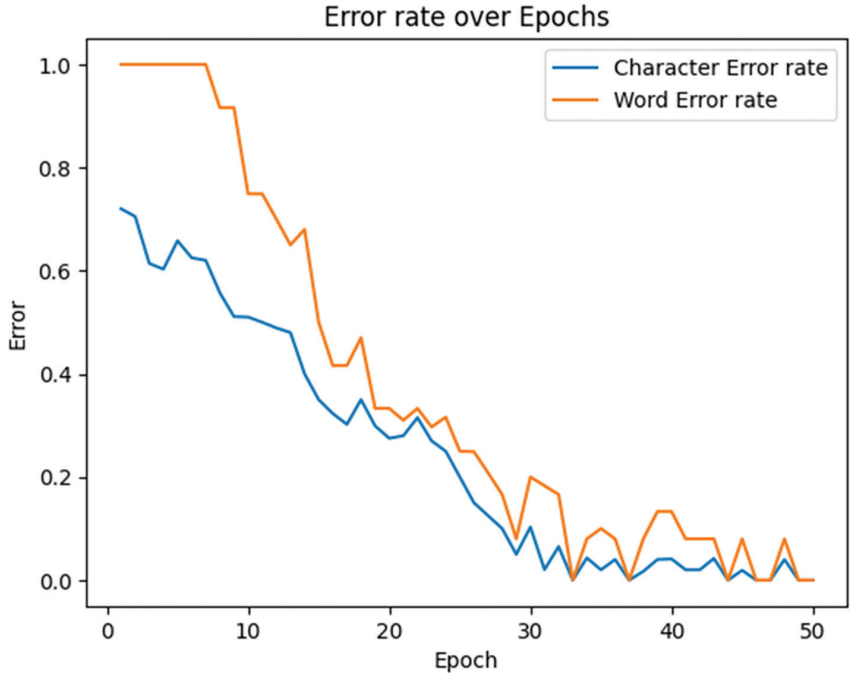


Figure 5: WER and CER over epochs

Similarly, training and validation losses were also mapped and plotted as in Figure 6. Training and validation losses respectively improved from 0.7881 and 0.6637 on epoch 1 to 0.0169 and 0.0064 on epoch 50, showing improvement in model performance.

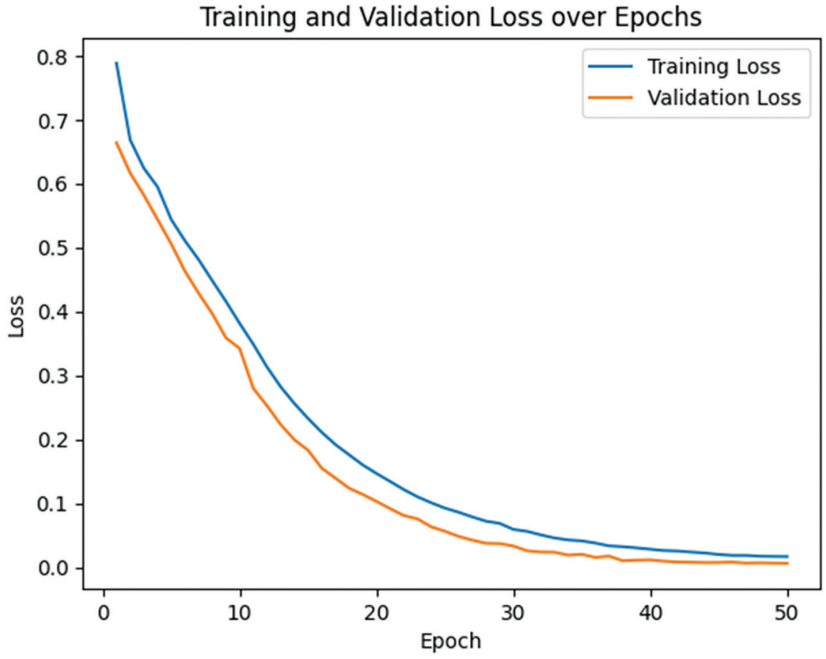


Figure 6: Training and Validation losses over epochs

3.3.3 Test Phase Performance

Independent testing on 50 samples of unseen speaker data yielded:

- Average WER: 0.1706 (83% word accuracy)
- Average CER: 0.0712 (93% character accuracy)

These test results demonstrate good generalization capability, though with expected performance degradation compared to validation metrics. The 83% word accuracy on unseen speakers compares favorably with the 95.2% sentence level accuracy reported by Assael et al. (2016) for LipNet, considering the differences in evaluation methodology and dataset complexity.

3.3.4 Methodological Analysis

Several methodological advantages emerge from this implementation:

1. End-to-End Learning: Unlike approaches requiring extensive preprocessing (Miled, Messaoud, & Bouzid, 2023) this system learns features directly from raw video data
2. Speaker Independence: The model demonstrates reasonable generalization across different speakers, addressing concerns raised by (Almajai, Cox, Harvey, & Lan, 2016)
3. Computational Efficiency: The architecture maintains reasonable computational requirements while achieving competitive accuracy

3.4 Limitations and Constraints Analysis

3.4.1 Technical Constraints

Several limitations were identified during implementation:

1. Resolution Dependency: The system requires specific 360×288-pixel resolution, limiting practical deployment flexibility
2. Frame Rate Constraints: Fixed 25fps requirement may limit applicability to various video sources
3. Controlled Environment: Performance assumes controlled lighting and framing conditions

3.4.2 Linguistic Limitations

The GRID corpus vocabulary constraint presents specific challenges:

- Limited to 64,000 possible sentence combinations
- Structured grammar may not reflect natural speech patterns
- Visually similar phonemes (e.g., 'b' and 'p') remain challenging to distinguish
- Existence of corrupt samples

3.4.3 Performance Limitations

Despite the promising results, the current implementation of the model is subject to several significant limitations that define its operational scope and highlight avenues for future work. The most prominent limitation is the system's inability to perform real-time data processing. The deep learning architecture, particularly the 3D Convolutional Neural Network, is computationally demanding and requires substantial GPU resources to process video streams. This resource-intensive nature makes instantaneous translation of live video feeds unfeasible with the available hardware. Consequently, the project was constrained to work with pre-recorded videos and a condensed subset of the full GRID corpus. A larger scale training regimen, incorporating the entire dataset, would likely yield further improvements in accuracy and generalization but was beyond the scope of the available computational resources.

Furthermore, the model's performance is highly dependent on specific data characteristics. It is currently calibrated to work only with videos of a fixed resolution (360x288 pixels) and a constant frame rate (25 fps). The feature extraction process also assumes a consistent, frontal-facing speaker with minimal head movement to ensure proper lip segmentation. Any deviation from these controlled conditions, such as variations in lighting, speaker pose, or video quality, would likely degrade performance significantly.

3.5 Discussion and Implications

3.5.1 Practical Applications

The performance levels achieved suggest viability for several practical applications:

1. Assistive Technology: 83% word accuracy may provide meaningful assistance for individuals with hearing impairments
2. Noisy Environment Applications: The visual-only approach offers advantages in scenarios where audio quality is compromised
3. Security Applications: Potential integration with existing surveillance systems for enhanced security protocols

3.5.2 Research Contributions

This work contributes to the lip-reading research field through:

1. Architecture Validation: Demonstrates effectiveness of CNN-BiLSTM hybrid approaches
2. Performance Benchmarking: Establishes competitive baseline for future research
3. Implementation Insights: Provides practical guidance for system deployment

3.5.3 Future Directions

Based on the current results, several research directions emerge:

1. Dataset Expansion: Training on larger, more diverse datasets could improve generalization
2. Architecture Optimization: Exploring attention mechanisms and transformer architectures
3. Real-time Implementation: Optimization for real-time processing requirements
4. Multi-modal Integration: Combining visual lip reading with other modalities for enhanced accuracy
5. Computational Optimization: Video processing, already a resource intensive task, on top of frame extraction and further computation posed a very difficult conundrum of economic feasibility for us. Future work can possibly look to optimize in this regard and formulate a more capable model with less resource utilization.

4. Conclusions

This research successfully developed and evaluated an end-to-end deep learning model for sentence-level lip-reading from visual data. By employing a hybrid architecture combining a 3D Convolutional Neural Network with a Bidirectional Long Short-Term Memory network, the system was trained to translate silent video of lip movements into text. The model was trained and validated on the GRID corpus, a constrained-grammar dataset, achieving high accuracy and demonstrating its efficacy. The work contributes a robust methodology for visual speech recognition, validating the potential of deep learning in creating valuable assistive technologies.

Based on the specific objectives of the project, the following key conclusions can be drawn:

- A systematic methodological framework was successfully established, which integrated the selection and pre-processing of the GRID corpus, a sequential workflow for feature extraction, model training, and performance evaluation.
- A hybrid deep learning architecture was successfully designed and implemented. The model effectively combines a 3D CNN for spatio-temporal feature extraction and a Bi-LSTM for sequence modeling, proving to be a potent combination for the visual speech recognition task.
- The model was successfully trained over 50 epochs, demonstrating effective learning and convergence as evidenced by the consistent reduction in training and validation loss. By the end of the training phase, both Word Error Rate (WER) and Character Error Rate (CER) on the validation set approached zero.
- The trained model demonstrated strong generalization capabilities on unseen data. When evaluated on a novel speaker not present in the training set, the system achieved a commendable word level accuracy of 83% (WER of 0.1706) and a character-level accuracy of 93% (CER of 0.0712).

4 Acknowledgements

We wish to express our sincere gratitude to Er. Dinesh Gothe, Head of the Department of Computer and Electronics Engineering at Khwopa College of Engineering, for providing the opportunity and support to undertake this project. We are also deeply indebted to our supervisor, Er. Mukesh Kumar Pokharel, for his invaluable guidance and mentorship throughout this research.

References

Adeel, A., Gogate, M., Hussain, A., & Whitmer, W. (2019). Lip-reading driven deep learning approach for speech enhancement.

Almajai, I., Cox, S., Harvey, R., & Lan, Y. (2016). Improved speaker independent lip reading using speaker adaptive training and deep neural networks. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2722-2726). Shanghai: IEEE.

Assael, Y. M. (2016). LipNet: End-to-end sentence-level lipreading. Ithaca, NY: arXiv.

Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 2421-2424.

El-Bialy, N., Chen, D., Fenghour, S., Hussein, W., Xiao, P., Karam, O., & Li, P. (2023). Developing phoneme-based lip-reading sentences system for silent speech recognition. *CAAI Transactions on Intelligence Technology*, 129-138.

Hao, M. M. (2020). A survey of research on lipreading technology.

Miled, M., Messaoud, M. A., & Bouzid, A. (2023). Lip reading of words with lip segmentation and deep learning. *Multimedia Tools and Applications*, 551-571.