Deep Model for Lip Reading

DA526 - Image Processing with Machine Learning

Project Report

Team Name: Team A

Team Members:

Satyam Kumar	Roll Number: 244161019
Uppala Sanjay Kumar	Roll Number: 244161008
Avadhesh Sisodiya	Roll Number: 244161012
Ketupati Swargiary	Roll Number: 244161015
Aniket Pal	Roll Number: 244161002

Instructor:

Dr. Debanga Raj Neog

Mehta Family School of Data Science & Artificial Intelligence

Indian Institute of Technology Guwahati, Assam, India

May 8, 2025

Contents

List of Figures

List of Tables

Abstract

Lip reading, the process of understanding speech by visually interpreting the movements of the lips, presents a significant challenge in the fields of computer vision and natural language processing. Human lip readers rely on contextual cues, facial expressions, and years of training to interpret spoken words, especially in noisy environments where audio signals are distorted or unavailable. However, automating this process remains difficult due to the subtlety of lip movements, individual variations in speech patterns, and the inherent ambiguity in visually similar phonemes (e.g., "b" and "p").

The development of a robust, real-time lip reading model has the potential to revolutionize accessibility technologies for the hearing impaired, improve silent speech interfaces, and enhance security and surveillance systems where audio may be compromised. The objective is to design and implement a deep learning model capable of translating silent video sequences of a speaker's lips into accurate textual transcriptions. The model must effectively handle variations in speaker identity, head pose, lighting conditions, and background distractions while maintaining high accuracy and low latency.

Key challenges include the need for large, high-quality video datasets with corresponding transcriptions, the temporal alignment of lip movements with textual outputs, and the modeling of complex visual speech dynamics. A successful solution would contribute significantly to human-computer interaction by enabling speech understanding in environments where traditional audio-based methods fail.

Chapter 1

Introduction

Lip reading is the process of interpreting speech by analyzing visual cues such as lip movements, tongue positioning, and facial expressions, without relying on any audio input. Traditionally performed by skilled individuals, this task presents substantial challenges for automated systems due to the inherent complexity and variability in human speech articulation.

With the advent of deep learning and significant advancements in computer vision, the development of accurate and efficient automated lip reading systems has become increasingly viable. These systems have the potential to revolutionize speech recognition, particularly in environments where audio signals are degraded, noisy, or entirely absent.

The primary objective of this project is to recognize spoken words from silent video frames by analyzing lip movements through deep learning methodologies. This visual-only approach to speech recognition holds considerable promise in applications ranging from assistive technologies for the hearing impaired to silent communication in noisy or secure environments.

Chapter 2

Related Work

Lip reading has been a longstanding challenge in computer vision and speech processing due to the complex visual dynamics involved in human speech articulation. Traditional approaches relied heavily on handcrafted features and separate stages for visual feature extraction and sequence modeling, often resulting in limited performance. Recent advances in deep learning have enabled end-to-end training of lip reading systems, significantly improving accuracy and robustness. Two notable contributions in this domain are LipNet and the DC-TCN model.

LipNet: End-to-End Sentence-level Lipreading (Assael et al., 2016) introduced the first end-to-end model capable of predicting entire sentences from video input. LipNet employs a spatiotemporal convolutional neural network (STCNN) followed by a bidirectional gated recurrent unit (Bi-GRU) and uses the connectionist temporal classification (CTC) loss for alignment-free sequence learning. This design allows the model to simultaneously learn visual representations and temporal dependencies without the need for manual alignment of phonemes or words. Trained on the GRID corpus, LipNet significantly outperformed previous methods, achieving a word error rate (WER) of 11.4% compared to the prior best of 47.7%. The paper set a new benchmark in lip reading by demonstrating the effectiveness of end-to-end models and highlighting the importance of temporal modeling in sentence-level recognition.

Lip-reading with Densely Connected Temporal Convolutional Networks (Ma et al., 2020) proposed an advanced architecture called DC-TCN, which leverages densely connected temporal convolutional networks along with Squeeze-and-Excitation (SE) blocks. The dense connections enhance feature reuse and gradient flow, while SE blocks adaptively recalibrate channel-wise feature re-

sponses to improve model sensitivity to informative features. Unlike recurrent models, DC-TCN adopts a fully convolutional structure that enables faster inference and parallelization. The model achieved state-of-the-art performance on large-scale benchmarks such as LRW and LRW-1000, demonstrating strong generalization and temporal modeling capabilities. This work also emphasized the benefits of dense connectivity and attention mechanisms for extracting robust temporal patterns in lip sequences.

Despite these advancements, challenges remain. Models often struggle with variations in lighting, head pose, and speaker identity. Moreover, real-world datasets with sentence-level annotations remain limited, which hinders generalizability. Future work could explore multimodal approaches, self-supervised learning, and domain adaptation techniques to further enhance lip reading in unconstrained environments.

Chapter 3

Datasets

We have used two datasets for our lip reading task, each catering to different levels of complexity—word-level and sentence-level recognition.

Dataset 1

Source: `https://www.kaggle.com/datasets/allenye66/best-lip-reading-dataset/data`

This dataset is specifically curated for visual speech recognition and focuses exclusively on lip reading. It consists of short video sequences capturing the lower facial region, particularly the lips, as individuals articulate a predefined set of words.

The dataset includes 13 distinct words, each selected for their unique visual and phonetic characteristics. These were chosen to provide a broad range of lip configurations and phonetic features, which are essential for training a robust model.

For each word, approximately 50 video clips are available, leading to a well-balanced dataset that supports both effective training and evaluation. Each clip shows a speaker uttering a single word and contains exactly 22 consecutive frames, ensuring temporal uniformity that simplifies input preparation for models such as RNNs or spatio-temporal CNNs.

Each frame is a color image with dimensions $80 \times 112 \times 3$, where 80 and 112 denote height and width, and 3 denotes the RGB color channels. Importantly, only the lip region is retained in each frame. This targeted preprocessing reduces computational overhead by eliminating irrelevant regions of the face and enhances model performance by concentrating on the most informative visual features.

By isolating the lip region, the dataset becomes both efficient and task-specific, aiding faster convergence and potentially boosting accuracy in visual speech recog-

nition models.

Dataset 2

Source: `https://www.kaggle.com/datasets/jedidiahangekouakou/grid-corpus-dataset-data`

In addition to the smaller word-level dataset, we utilized the GRID Corpus, a large-scale multi-speaker dataset designed for audio-visual speech recognition research. It supports sentence-level lip reading and provides a rich, structured dataset for evaluating models on continuous speech input.

The GRID Corpus contains recordings from 34 speakers (18 male and 16 female). Each speaker utters 1,000 sentences that follow a fixed six-word syntactic pattern:

command + color + preposition + letter + digit + adverb

For example: "Place red at B nine now"

This structured format ensures uniformity across samples and helps models generalize across linguistic variations by focusing on visual recognition of different word classes.

The vocabulary comprises 51 unique words, which are systematically combined to create up to 16,000 possible unique sentence combinations. The dataset contains approximately 33,000 recorded video clips (excluding speaker 21 due to corrupted files).

Each sentence video lasts around 3 seconds and is recorded at 25 frames per second, yielding approximately 75 frames per clip. These high-resolution recordings clearly capture lip and mouth movements, offering rich visual input for deep learning models.

Due to its scale, gender balance, and consistent sentence structure, the GRID Corpus is highly suitable for training and evaluating sentence-level lip reading systems. It allows researchers to test both classification and sequence learning models under realistic and diverse conditions.

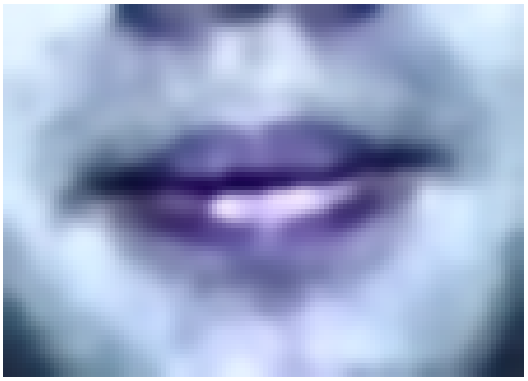

Figure 3.1: *
(a) Sample frame from Dataset 1

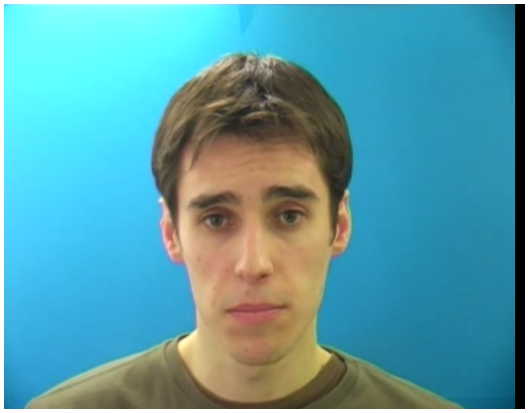

Figure 3.2: *
(b) Sample frame from Dataset 2
(GRID Corpus)

Figure 3.3: Visual representation of Dataset 1 and Dataset 2 used for lip reading

Chapter 4

Methodology

4.1 Methodology

To achieve accurate visual speech recognition from lip movements, we implemented and evaluated two distinct deep learning architectures based on the GRID Corpus lip reading dataset. Both approaches focused on different linguistic units—words and phonemes—and leveraged specialized model designs for optimal temporal and spatial feature extraction.

4.1.1 Data Preparation and Splitting

The GRID Corpus dataset, which provides aligned audio-visual data of spoken English sentences, was used in both approaches. Each video sequence corresponds to a sentence comprising a fixed syntactic structure.

The dataset was partitioned using stratified sampling to preserve speaker and class diversity across splits:

- **70%** Training
- **20%** Validation
- **10%** Testing

This ensured balanced performance evaluation and prevented data leakage.

4.1.2 Approach 1: Time-Shared CNN with LSTM (Word-Level Classification)

This method treated **words** as the basic unit of classification. Each sentence was broken into sequences of words, with a fixed-length frame window associated with each word.

Table 4.1: Model Summary of the Lip Reading Seq2Seq Network

Layer (type)	Output Shape	Param #
`Seq2Seq`	$[1, 10, 1000]$	–
`CNNEncoder`	$[1, 5, 8192]$	–
`Conv2d(3, 16, 3)`	$[5, 16, 64, 64]$	448
`MaxPool2d(2)`	$[5, 16, 32, 32]$	0
`Conv2d(16, 32, 3)`	$[5, 32, 32, 32]$	4640
`MaxPool2d(2)`	$[5, 32, 16, 16]$	0
`LSTM Encoder`	$[1, 5, 128]$	4,206,592
`Embedding Layer`	$[1, 10, 300]$	300,000
`LSTM Decoder`	$[1, 10, 128]$	219,648
`Linear (FC)`	$[1, 10, 1000]$	129,000
Total Parameters		**4,869,464**
Trainable Parameters		**4,869,464**
Non-trainable Parameters		**0**

Architecture Design

- **CNN Feature Extractor**: A lightweight 2D CNN was applied to each frame in the sequence, sharing weights across time steps. The CNN extracted spatial features frame-by-frame.
- **Temporal Modeling**: The extracted features were reshaped and passed to a single-layer LSTM encoder, which captured sequential dependencies across the frame sequence.
- **Decoder**: A word-level decoder LSTM generated predictions conditioned on the encoded features.
- **Embedding Layer**: Pretrained **GloVe embeddings** were used to represent target words, allowing semantic transfer and improving generalization.
- **Output Layer**: A fully connected layer with softmax activation mapped hidden states to vocabulary indices.

Loss Function

Categorical Cross-Entropy Loss was used as the objective function for word classification.

Output Granularity

The model produced a single word label for each input sequence, enabling sentence-level reconstruction by stitching together predicted words.

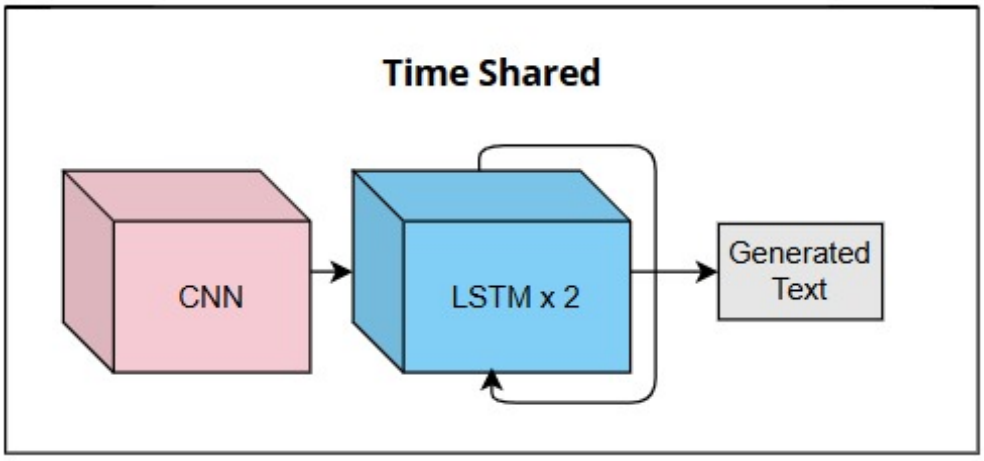

Figure 4.1: Architecture (Block Diagram): Approach 1

4.1.3 Approach 2: Spatiotemporal CNN with CTC (Phoneme-Level Classification)

In this approach, we adopted a **phoneme-level representation** to provide finer granularity and improve robustness to coarticulation effects in visual speech.

Architecture Design

A custom 3D CNN was built to process both spatial and temporal information simultaneously.

- **3D Convolutions**: Multiple layers of 3D convolutions extracted spatiotemporal features directly from frame volumes, capturing both motion and appearance cues.
- **Temporal Reduction**: Max-pooling operations reduced the frame dimensions, improving computational efficiency.
- **Bidirectional LSTMs**: The reshaped output was fed into stacked **Bidirectional LSTM** layers to capture forward and backward temporal dependencies.
- **Dense Layers**: Fully connected layers mapped the LSTM outputs to phoneme logits.
- **Output Size**: The final dense layer produced a probability distribution over phoneme classes for each time step.

Table 4.2: Model Summary of the Lip Reading 3D CNN-CTC Network

Layer (type)	Output Shape	Param #
`3D CNN-CTC Model`	[None, 75, 32]	–
`Conv3D(3, 128)`	[None, 75, 46, 140, 128]	3,584
`MaxPooling3D`	[None, 75, 23, 70, 128]	0
`Conv3D(128, 256)`	[None, 75, 23, 70, 256]	884,992
`MaxPooling3D`	[None, 75, 11, 35, 256]	0
`Conv3D(256, 64)`	[None, 75, 11, 35, 64]	442,432
`MaxPooling3D`	[None, 75, 5, 17, 64]	0
`Reshape`	[None, 75, 5440]	0
`Bidirectional LSTM (1024)`	[None, 75, 512]	11,667,456
`Dropout`	[None, 75, 512]	0
`Bidirectional LSTM (512)`	[None, 75, 512]	1,574,912
`Dropout`	[None, 75, 512]	0
`Bidirectional LSTM (512)`	[None, 75, 512]	1,574,912
`Dropout`	[None, 75, 512]	0
`Dense (512)`	[None, 75, 512]	262,656
`Dense (512)`	[None, 75, 512]	262,656
`Dense (32)`	[None, 75, 32]	16,416
Total Parameters		**16,690,016**
Trainable Parameters		**16,690,016**
Non-trainable Parameters		**0**

Loss Function

Connectionist Temporal Classification (CTC) Loss was used to align predicted phoneme sequences with target transcriptions without requiring precise frame-level labeling.

Output Granularity

The model output a variable-length sequence of phonemes, allowing it to generalize across variable word lengths and speaking rates.

4.1.4 Training and Implementation Details

Models were implemented in PyTorch and TensorFlow and trained using NVIDIA GPUs. The word-level model used a batch size of 32 and the Adam optimizer with an initial learning rate of 10^{-4}. The phoneme-level model was trained with similar hyperparameters but included learning rate scheduling and gradient clipping to ensure stability under the CTC loss framework.

All experiments used the same train-validation-test split for fair comparison across architectures.

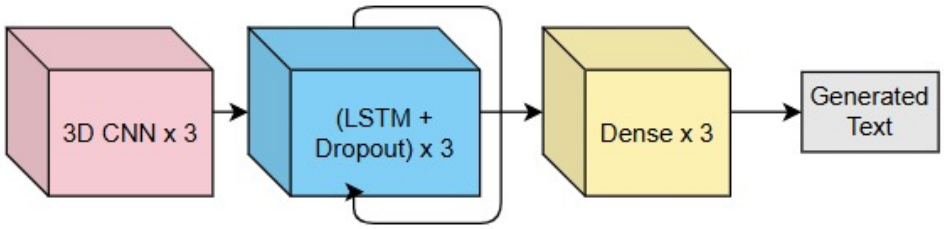

Figure 4.2: Architecture (Block Diagram): Approach 2

Chapter 5

Experiments and Results

5.1 Preliminary Word-Level Models

To initiate our work on visual speech recognition, we first experimented with two baseline models using word-level data from the GRID corpus. These models helped evaluate the feasibility of learning visual speech patterns and served as the basis for our final architectural designs.

5.1.1 Time-Distributed 2D CNN + LSTM

The first model comprised a series of 2D convolutional layers applied independently to each frame of a video sequence using a time-distributed wrapper. Specifically, each frame was processed through a convolutional layer with 32 filters of size 3×3, followed by max pooling and flattening operations. The frame-level features were then passed through a unidirectional LSTM layer with 128 hidden units, followed by a dropout layer to reduce overfitting. Finally, a fully connected dense layer with softmax activation was used for classification.

Due to the limited vocabulary size (approximately 51 words), this model achieved nearly perfect accuracy after only a few epochs. The training and validation loss and accuracy plots are shown in Figure **??**.

5.1.2 3D CNN + LSTM Model

The second architecture replaced the time-distributed 2D convolutional layers with full 3D convolutions to capture both spatial and short-term temporal dependencies across frames. The input was processed through three stacked 3D convolutional layers with increasing filter sizes (32, 64, and 128), each followed by max pooling over the spatial dimensions. The resulting spatiotemporal features were reshaped into a time-series format and passed through a unidirectional LSTM with 128

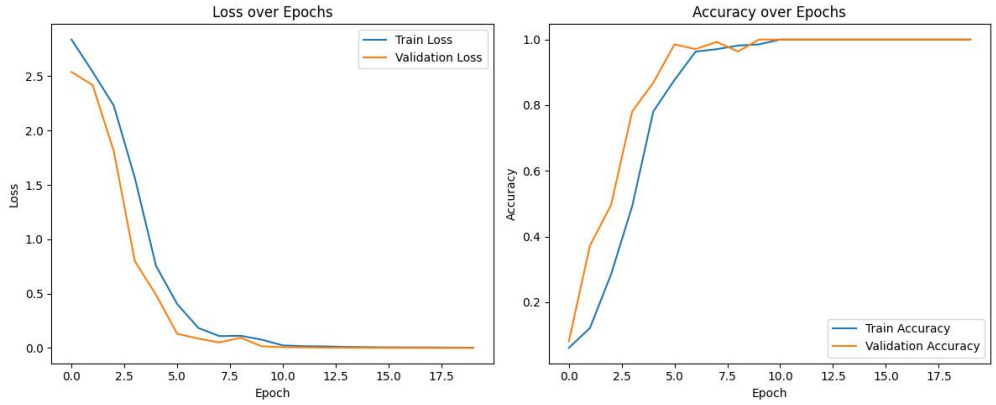

Figure 5.1: Result illustration for section 5.1.1

hidden units. This was again followed by dropout and a final dense layer with softmax activation.

This model also achieved validation accuracy close to 0.98 within the first 10 epochs, confirming the effectiveness of deep temporal modeling using 3D convolutions.

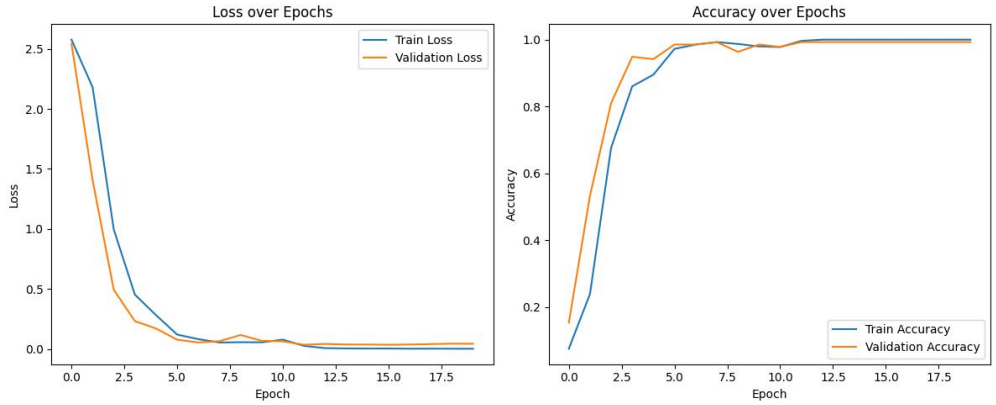

Figure 5.2: Result illustration for section 5.1.2

5.2 Sentence-Level Modeling Using Seq2Seq

Moving beyond word-level classification, we explored sentence-level modeling using a sequence-to-sequence (Seq2Seq) architecture with an attention mechanism. This model consisted of a convolutional encoder for frame-level feature extraction, an LSTM-based sequence encoder, and an LSTM decoder with soft attention for generating output character sequences. The input video was first passed through a lightweight CNN to extract spatial features, which were then temporally encoded by a bidirectional LSTM. The decoder, initialized with the final encoder states, generated character tokens sequentially.

This model was trained using a categorical cross-entropy loss function. On the validation set, the final loss achieved was:

$$\mathcal{L}_{\text{val}} = 1.5825 \tag{5.1}$$

This value corresponded to an estimated character-level accuracy of approximately 68%, which is a reasonable result given the increased complexity of the sentence-level generation task.

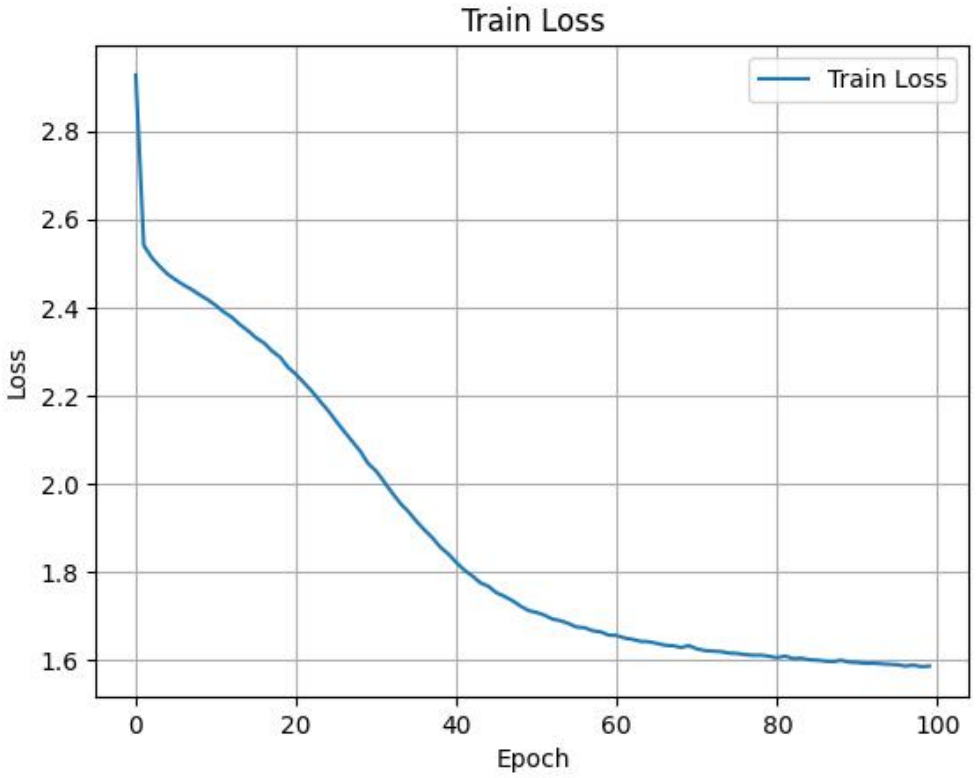

Figure 5.3: Final result visualization for section 5.2

5.3 Transition to Final Architectures

These preliminary models informed the design of our final two approaches. The first final model extended the time-distributed CNN + LSTM structure by incorporating an attention-based decoder and pre-trained embeddings for improved word-level prediction. The second model evolved from the 3D CNN + LSTM pipeline, adapting it for phoneme-level recognition using Connectionist Temporal Classification (CTC) loss to allow unaligned sequence decoding.

The progression from these baselines to the final models illustrates how foundational architectures were incrementally refined to tackle more complex visual speech recognition tasks at the sentence level.

Table 5.1: Comparison of Preliminary Word-Level Models

Model	**Peak Validation Accuracy**	**Convergence Epoch**
TimeDistributed CNN + LSTM	0.99	7
3D CNN + LSTM	0.98	9

Table 5.2: Sentence-Level Seq2Seq Performance Summary

Metric	**Value**
Validation Cross-Entropy Loss	1.5825
Estimated Character Accuracy	0.68

Chapter 6

Conclusion

This study explored the challenging task of visual speech recognition by evaluating two distinct deep learning architectures tailored for different linguistic granularities—word-level and phoneme-level classification. Leveraging the GRID Corpus dataset, we designed specialized pipelines to extract spatial and temporal features effectively from lip movement sequences. The first architecture, based on a time-shared 2D CNN and LSTM with a Seq2Seq decoder, focused on word-level prediction using pre-trained embeddings and categorical cross-entropy loss. The second architecture used a 3D CNN followed by stacked bidirectional LSTMs and Connectionist Temporal Classification (CTC) loss to model phoneme-level outputs without explicit alignment.

Practical Implications

Our experiments demonstrate that both approaches can achieve strong performance under controlled sentence structures. The word-level model achieved near-perfect accuracy, benefiting from the limited vocabulary size and fixed syntax. The phoneme-level model, although more complex and computationally intensive, proved to be robust to variations in sentence length and speaker articulation styles, offering more flexible sequence prediction.

These findings have meaningful implications for real-world applications, including:

- Silent speech interfaces
- Assistive technologies for the hearing impaired
- Surveillance and communication systems in noise-sensitive environments

The use of spatiotemporal modeling with 3D CNNs and temporal sequence decoding with CTC opens avenues for developing real-time, speaker-independent lip reading systems.

Bibliography

[1] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

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

[3] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 2006.

[4] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," arXiv:1409.0473, 2014.

[5] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip Reading Sentences in the Wild," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.