**1. Problem Statement:**

* **Problem:** Implement a Bilstm based deep learning-based model for automatic lipreading, capable of converting video sequences of lip movements into sentences.
* **Scope:** Focus on training the model on video datasets and improving its performance to accurately predict sentences from lip movements.

**2. Meta Data of Dataset:**

* **Dataset:** GRID corpus with aligned text labels./oxford lip reading/lrw(lip reading in the wild)

### Metadata Description: **GRID Corpus Metadata:**

1. **Dataset Size:**
   * **Speakers:** 34 speakers (18 male and 16 female).
   * **Total Utterances:** 33,000 utterances (approximately 1,000 utterances per speaker).
2. **Data Types:**
   * **Video:** Video recordings of speakers' faces, focused on the mouth region, with a resolution of 720x576 pixels at 25 frames per second.
   * **Audio:** Synchronized audio recordings of the speakers' utterances, provided in 16-bit WAV format at a sampling rate of 25 kHz.
   * **Text Transcripts:** Corresponding text transcripts of the spoken sentences.
3. **Sentence Structure:**
   * **Format:** Each sentence follows a fixed structure with six components: <command> <color> <preposition> <letter> <digit> <adverb>.
   * **Example:** "place blue at A nine now."
   * This structured format ensures consistency across the dataset.
4. **Annotations:**
   * **Transcriptions:** Each video is annotated with the corresponding transcription, ensuring alignment between the video, audio, and text data.
   * **Speaker Information:** Metadata includes details about each speaker, such as gender and ID.
5. **Applications:**
   * Used for tasks like automatic speech recognition, lipreading, and multimodal fusion research.

**3. Exploratory Analysis:**

* **Data Exploration:** Analyze the video sequences to understand the variability in lip movements. Visualize sample videos and their corresponding text.
* **Challenges Identification:** Identify potential challenges such as varying lighting conditions, occlusions, and different speaking speeds.

**4. Preprocessing Pipeline Specific to Data:**

* **Video Preprocessing:** Extract frames from the videos and resize them to a fixed resolution. Normalize pixel values to improve model performance.
* **Data Augmentation:** Apply data augmentation techniques such as random cropping, flipping, or brightness adjustment to make the model robust.
* **Label Encoding:** Convert the text labels into sequences of tokens for training. Use character-level encoding for flexibility in predicting unseen words

**5. Relevant Performance Metrics:**

* **Accuracy:** Measure the accuracy of the predicted sentences compared to the ground truth.
* **Word Error Rate (WER):** Calculate the word error rate to evaluate how well the model transcribes lip movements into text.
* **CTC Loss (Connectionist Temporal Classification):** This is the appropriate loss function for sequence-to-sequence tasks like lipreading, where the length of the predicted sequence can vary from the length of the input.

**6. Project Objectives:**

* **Objective:** Train a model that can achieve high accuracy in predicting sentences from lip movements.
* **Milestones:**
  + Achieve baseline accuracy with basic preprocessing.
  + Implement and improve upon the spatiotemporal convolutional and recurrent neural network architecture.
  + Optimize training with Connectionist Temporal Classification (CTC) loss.

**Literature review of Models to be implemented:**

1. **Convolutional Neural Networks (CNNs) for Feature Extraction:**
   * + CNNs, particularly 3D CNNs, are widely used to capture spatiotemporal features from video frames. In your model:
     + **Conv3D layers** capture both spatial and temporal information from input video sequences.
     + **MaxPooling** reduces spatial dimensions, making the model computationally efficient while retaining the most important features.
     + Studies like [*"End-to-End Learning for Lip Reading"* (Chung et al., 2016)](https://arxiv.org/abs/1611.01599) highlight the importance of CNNs in extracting spatial features in lipreading tasks. Using multiple CNN layers with increasing filter sizes, followed by pooling, helps reduce the data's dimensionality while focusing on essential patterns.
2. **Recurrent Neural Networks (RNNs) for sequence learning :**
   * + Recurrent neural networks (RNNs) are used for handling sequential data, especially for modeling temporal dependencies. In your model, BiLSTM is currently employed for sequence learning. Below are the alternatives and their advantages:
     + **Bidirectional Long Short-Term Memory (BiLSTM):** BiLSTMs are commonly used in sequence-to-sequence models, especially in tasks where both past and future contexts are crucial. In lipreading, the ability to retain long-term dependencies can help capture information that spans across multiple frames.
     + **Bidirectional Gated Recurrent Units (BiGRU):** GRUs are a simpler variant of LSTMs with fewer parameters, which can sometimes lead to faster training and improved efficiency. They also handle vanishing gradient issues well.

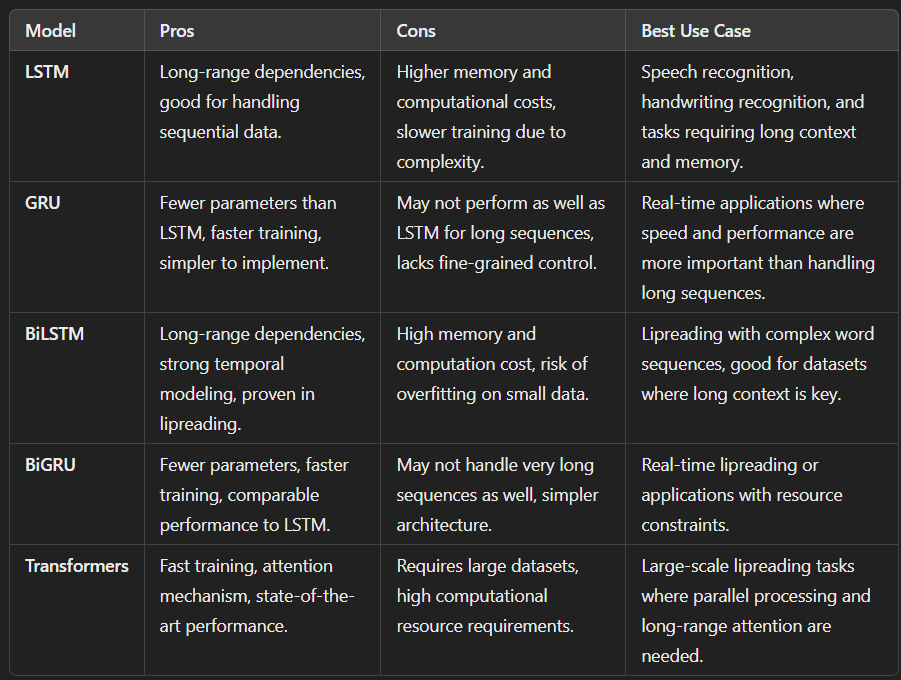
<https://www.sciencedirect.com/topics/computer-science/bidirectional-long-short-term-memory-network>

<https://www.researchgate.net/publication/13853244_Long_Short-term_Memory>

1. **Transformers for parallel execution:**
   * + Transformers, particularly self-attention mechanisms, have shown state-of-the-art performance in various sequence-to-sequence tasks. Instead of processing sequences in a strictly sequential manner like RNNs, transformers allow each position in the input sequence to attend to every other position directly, capturing long-range dependencies more effectively. For lipreading, transformers may help improve the accuracy by handling temporal sequences with varying importance, as demonstrated in models like *"*[*Attention Is All You Need" (Vaswani et al., 2017)*.](https://arxiv.org/abs/1706.03762)
     + In the case of our lipreading task:

[Transformer-Based Lip-Reading with Regularized Dropout and Relaxed Attention | IEEE Conference Publication | IEEE Xplore](https://ieeexplore.ieee.org/document/10023442)

**Pros and Cons Of Each Model:**

****

**Baseline:**

Starting with a simple CNN for spatial feature extraction followed by a single-direction LSTM for sequence modeling is a solid approach. This can serve as a good baseline for comparison with more advanced architectures. The softmax layer at the end will work well for classification tasks like lipreading.

* 3D Convolution Networks (x3): Using 3D CNNs will effectively capture both spatial and temporal features, as lip movements are inherently spatiotemporal. Three layers of 3D convolutions will help the network learn progressively deeper features.
* BiLSTMs (x2): Replacing single LSTMs with BiLSTMs is a good choice for lipreading since they process the input sequences in both directions, which is crucial for understanding the temporal dependencies across frames. Two layers of BiLSTMs will enhance the model's ability to capture long-term dependencies in lip movements.

**Error Metric For Evaluation:**

**CER(character error rate)-**

Character Error Rate (CER) is a metric used to measure the performance of a text recognition model or **automatic speech recognition (ASR)** system.

CER is the percentage of characters that are transcribed incorrectly.

CER is calculated by **dividing the total number of incorrect characters by the total number of characters in the reference text**.

A l**ower** CER indicates better performance, with a CER of 0 being a perfect score.

**Target:**

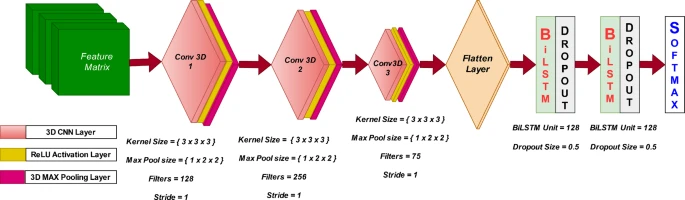
Spoken sentence in the input video

**Model Configuration**

**BiLstm:** Lr=0.0001 , Optimizer= Adam , Loss= CTCLoss

**BiRNN:** Lr=0.0001 , Optimizer=Adam , Loss=CTCLoss

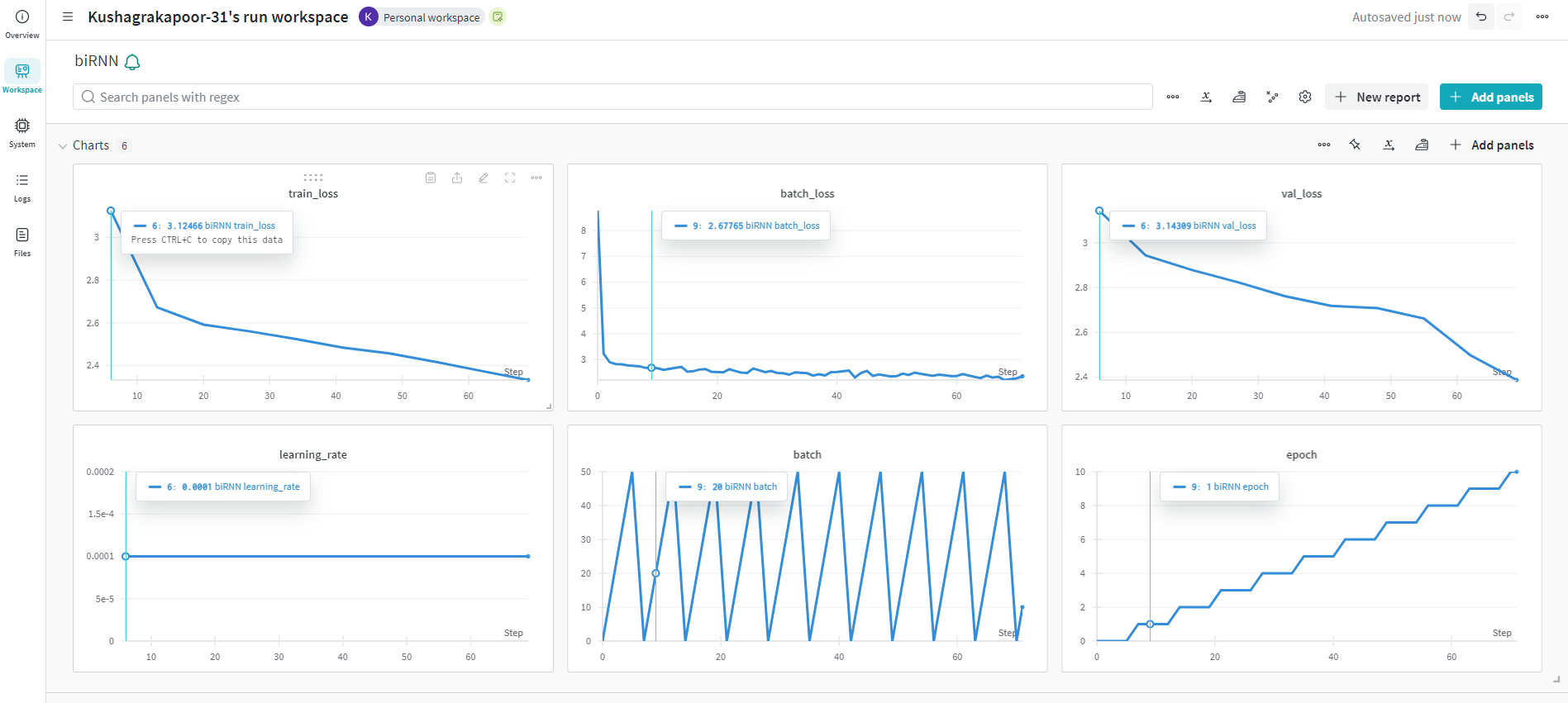
**Selected Model for the task:**

****

**Training Visualization for BiLSTM and BiRNN:**

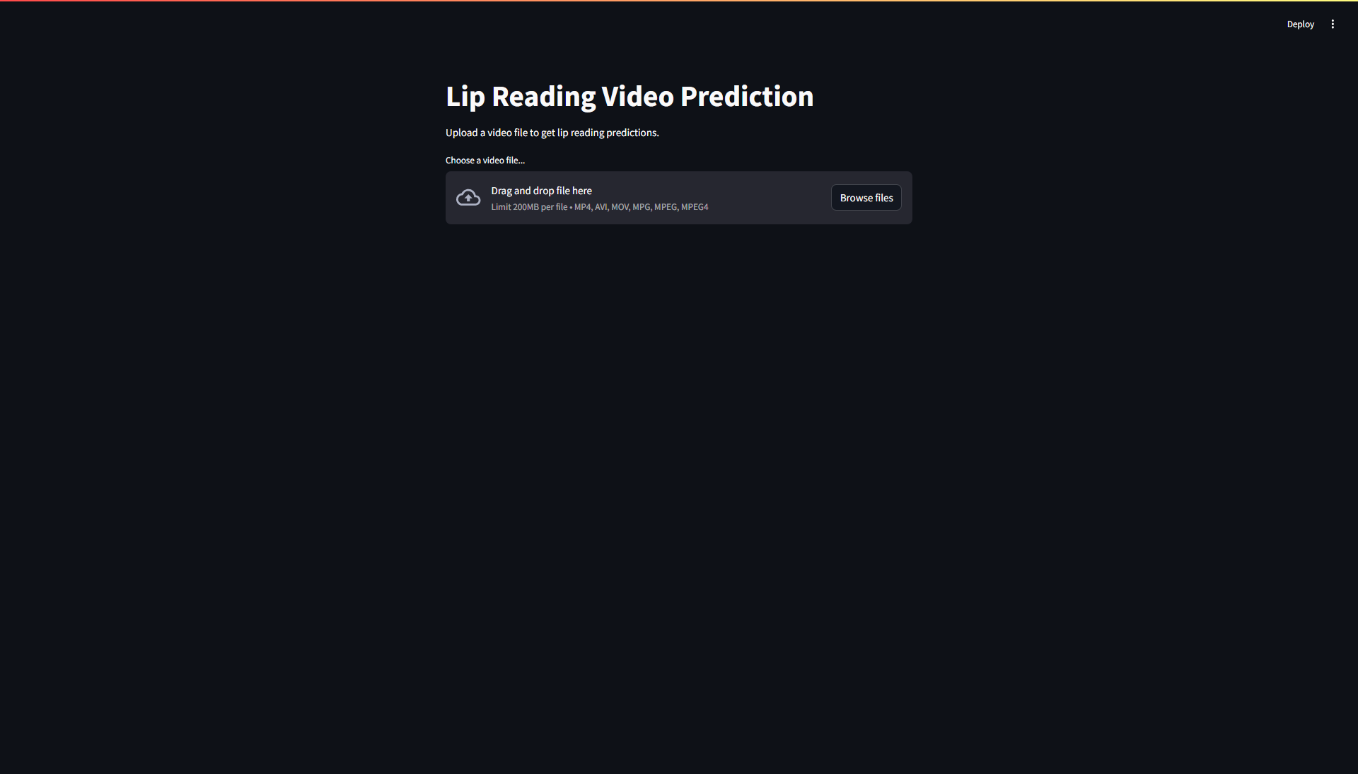
****

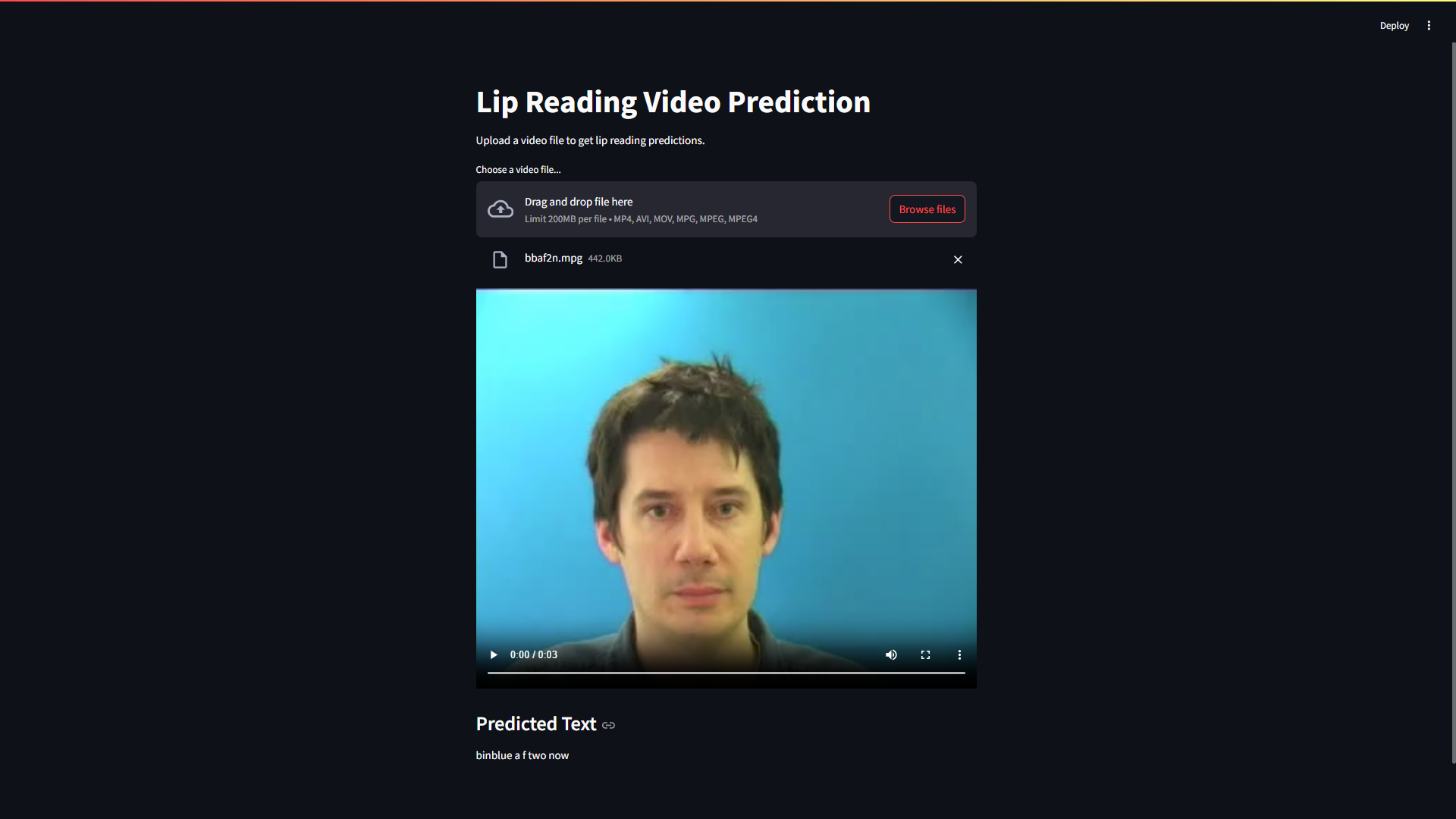
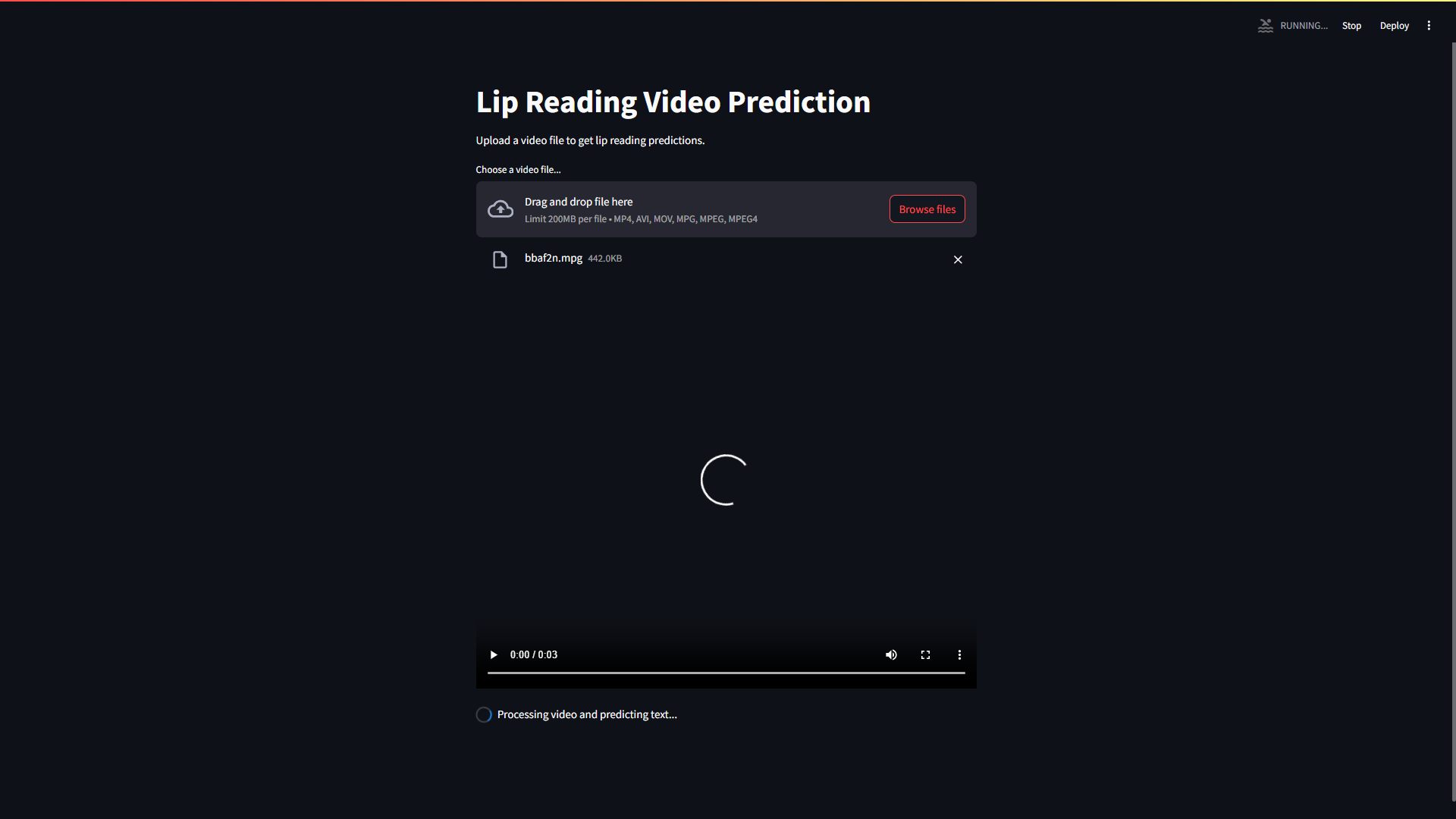
**Fig1:BiLSTM**

****

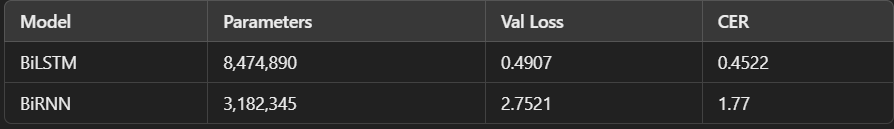
**Fig2:BiRNN**

**Deployment: (used streamlit)**

****

****

**Tabulation of Results obtained:**



The CER (Character Error Rate) for the BiLSTM model at 0.4522 indicates significantly better accuracy compared to the BiRNN model's 1.77. This suggests that BiLSTM handles sequential dependencies more effectively, leading to fewer errors in predictions.

**Hyperparameters:**

A batch size of 8 was used for 450 examples, resulting in 57 batches per epoch. The model was trained for 100 epochs using the Adam optimizer with a learning rate of 0.0001, ensuring identical training environments across experiments.

**Conclusion:**

In the case of lipreading, the selected BiLSTM model demonstrated effective performance, achieving a notably lower Character Error Rate (CER) compared to simpler recurrent models. The model's bidirectional LSTM layers were adept at capturing temporal dependencies in the visual sequences, which proved essential for interpreting subtle lip movements accurately. With a CER of 0.4522 and a validation loss of 0.4907, the BiLSTM model effectively learned context and improved recognition accuracy, making it a strong choice for lipreading tasks where precision in sequential visual cues is crucial

**Research Papers for reference:**

[[1611.01599] LipNet: End-to-End Sentence-level Lipreading (arxiv.org)](https://arxiv.org/abs/1611.01599)

[Transformer-Based Lip-Reading with Regularized Dropout and Relaxed Attention | IEEE Conference Publication | IEEE Xplore](https://ieeexplore.ieee.org/document/10023442)

<https://arxiv.org/abs/1706.03762> (Attention is all you need)

[Deep Learning for Visual Speech Analysis: A Survey | IEEE Journals & Magazine | IEEE Xplore](https://ieeexplore.ieee.org/abstract/document/10472054)