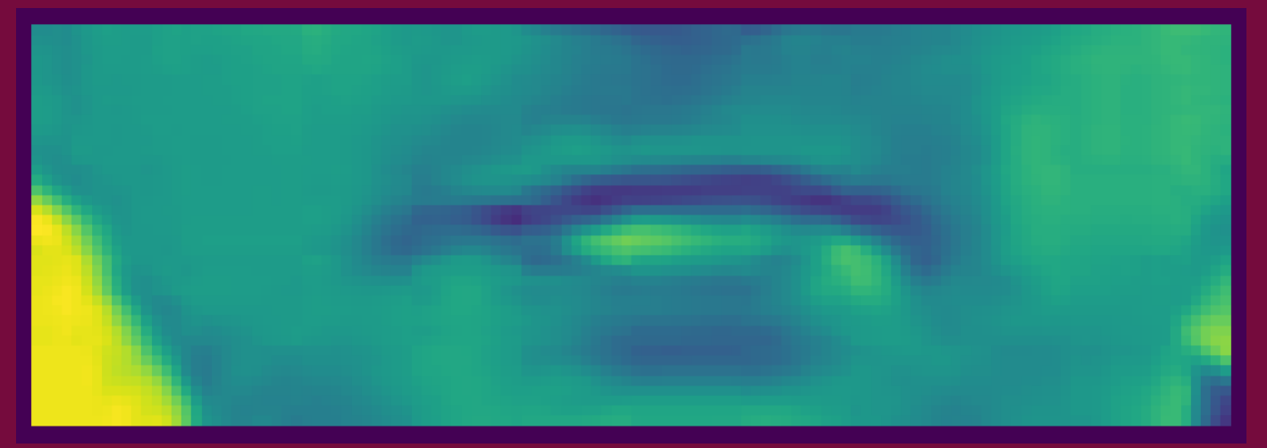


Real-Time Visual Speech Recognition with LIP-TRAC:

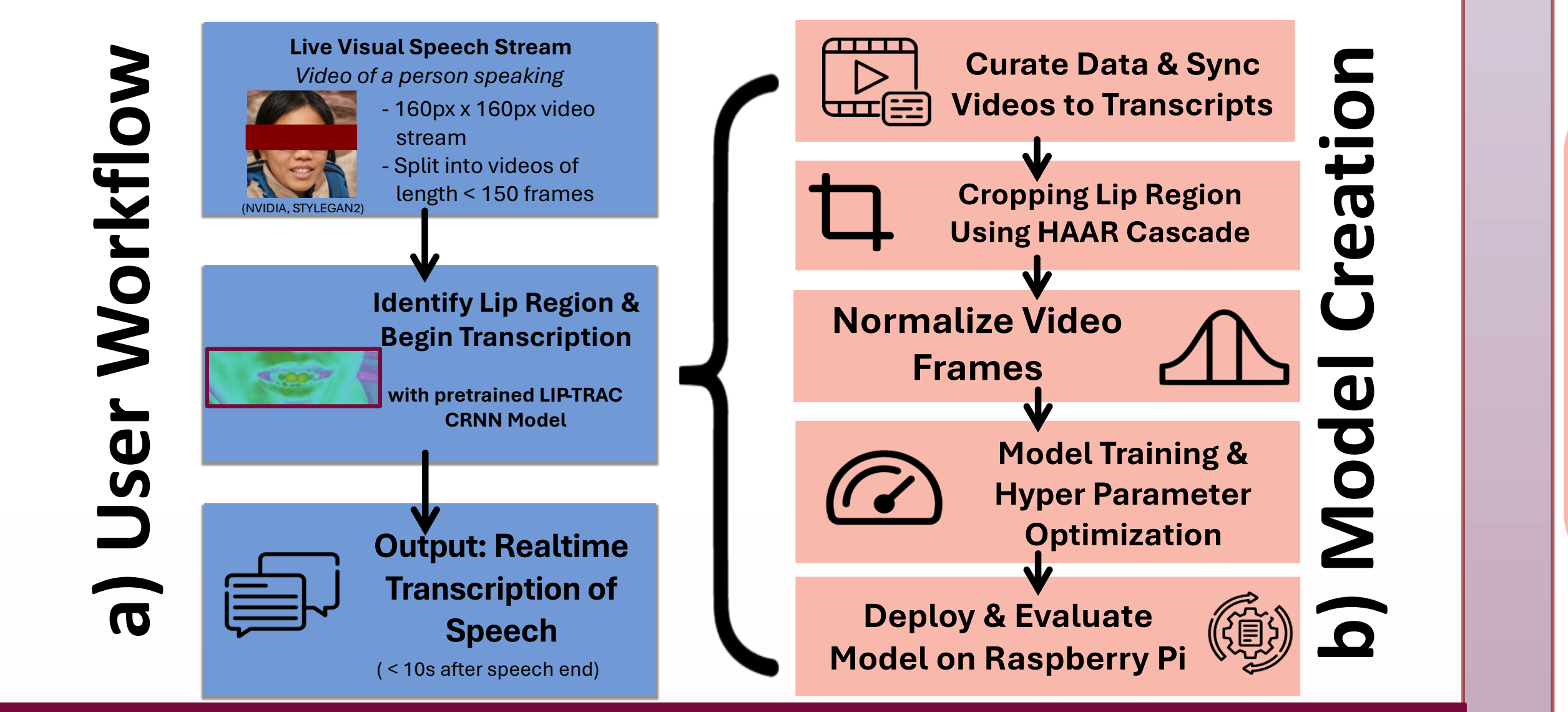
Lipreading through a Temporal Recurrent and Convolutional Neural Network Enhancing Practical Communication, Reducing Latency in Transcription, and Improving Accessibility for the Hearing Impaired

Monish Saravana Kumar Divya Sundari

Credit Line of Origin
References & Data: in references below.

All graphics, tables, and images have been created by Monish Saravana Kumar Divya Sundari, unless otherwise attributed.

VISUAL ABSTRACT


[redacted]

a) User Workflow

Live Visual Speech Stream
Video of a person speaking
- 160px x 160px video stream
- Split into videos of length < 150 frames

(NVIDIA, STYLEGAN2)

Identify Lip Region & Begin Transcription
with pretrained LIP-TRAC CRNN Model

Output: Realtime Transcription of Speech
(< 10s after speech end)

b) Model Creation

Curate Data & Sync Videos to Transcripts

Cropping Lip Region Using HAAR Cascade

Normalize Video Frames

Model Training & Hyper Parameter Optimization

Deploy & Evaluate Model on Raspberry Pi

INTRODUCTION

Hearing Loss

Hearing loss affects 1 in 5 people, 1.5 billion people worldwide, and in 25 years, this number is expected to rise to **2.5 billion. (World Health Organization).**

Visual Speech Reading AKA **Lipreading is crucial for the day-to-day life of these individuals.** However human lipreading accuracy remains very low, with 30% accuracy on the high end. (Assael et al.)

Current hearing aid technology is insufficient in noisy / outdoor environments. Additionally, the average price of a hearing aid is over $1000, severely restricting accessibility.

Lack of accessibility of existing technologies

Low Accuracy Of Human Lipreading

High inference time in existing models

Visual Speech Recognition (VSR)

Visual Speech recognition AKA Lipreading is the act of identifying speech using only visual cues. Existing models severely lack real-time & real-world capabilities.

Goal: Develop a low-inference time, high accuracy lipreading model that can be used in the real world to increase accessibility and integrate it into a physical device.

(1) Existing visual speech recognition models **are primarily "one speaker" models**, severely limiting their use cases.

(2) Models are **not catered to real-world usage**: minimal implementation of facial identification & lip extraction, and no physical device.

(3) **High inference time** in order to maximize accuracy: Not practical for day-to-day use.

The above problems prevent VSR models from being used by those with hearing loss, in day-to-day life.

RESEARCH OBJECTIVES

ENGINEERING PROBLEM – Current assistive technologies are not accessible to people with hearing loss, due to cost, situational limitations, or low inference speed. Current models are not suitable for real-world application.

RESEARCH QUESTION – *Can we develop a cost-accessible lip-reading device that balances speed and accuracy for better real-time usability?*

CLAIM – This research will develop a real-time lipreading device that balances speed and accuracy, achieving low inference time and high transcription accuracy. It will achieve close performance to state of the art models, while reducing inference time & being cheaper than traditional hearing aids.

Engineering Design Criteria for Lipreading Model

Versatility: Train model on variety of speakers to allow generalization to everyone

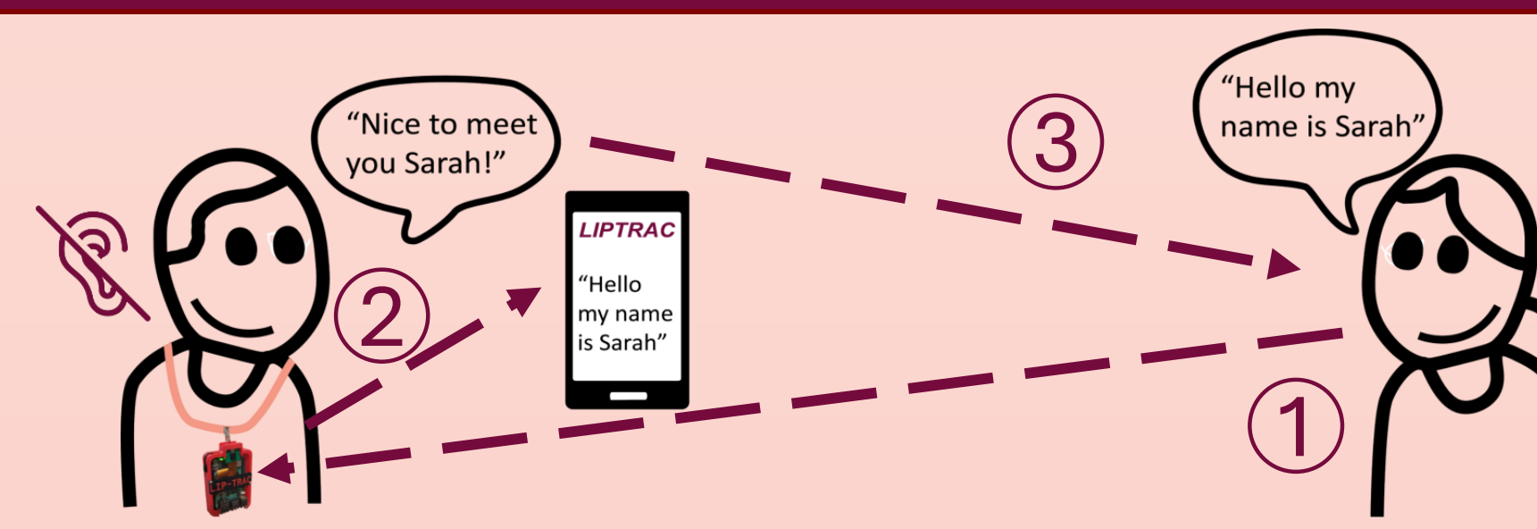
Functionality: Learn complex lip movements & transcribe them in real-time (<10s)

Accuracy: Achieves <35% WER and <20% CER.

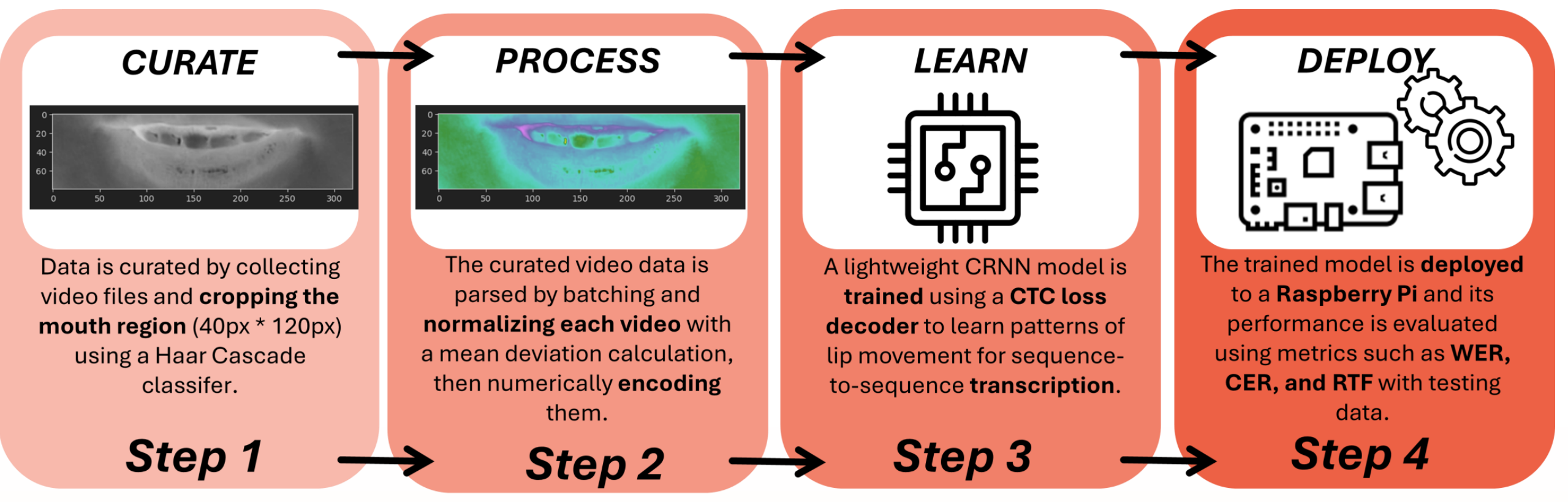
BRAINSTORMING A SOLUTION – How to achieve the criteria?

- ***Curation:*** Utilize BBC LRS2 Dataset with a 70/30 Training-Testing Split. Utilize computer vision (HAAR Cascade) to crop lip regions from each video frame.
- ***Normalization & Encoding:*** Normalize each video within itself and numerically encode each video.
- ***Training & Learning:*** Train a lightweight CRNN (convolutional recurrent neural network) architecture, using a CTC Loss function to address sequence-to-sequence learning.
- ***Analyzing & Evaluating:*** Evaluate the models and compare it to state-of-the-art models trained on similar datasets, using WER (Word Error Rate), CER (Character Error Rate), and inference time (time taken to output)
- ***Integrate:*** *Construct a physical prototype* as a proof of concept

PROBLEM FRAMING



ENGINEERING METHODOLOGY



CURATE – Data is curated by collecting video files and **cropping the mouth region** (40px * 120px) using a Haar Cascade classifer. **Step 1**

PROCESS – The curated video data is parsed by batching and **normalizing each video** with a mean deviation calculation, then numerically **encoding** them. **Step 2**

LEARN – A lightweight CRNN model is **trained** using a **CTC loss decoder** to learn patterns of lip movement for sequence-to-sequence **transcription**. **Step 3**

DEPLOY – The trained model is **deployed** to a **Raspberry Pi** and its performance is evaluated using metrics such as **WER, CER, and RTF** with testing data. **Step 4**

Overview & Summary of diagrams/figures below:

❑In the first step, each video (1064) is made **grayscale**, then the face is identified through a Haar Cascade, then using relative positioning, the mouth/lip region is masked & **cropped** and scaled to 40 x 120px.

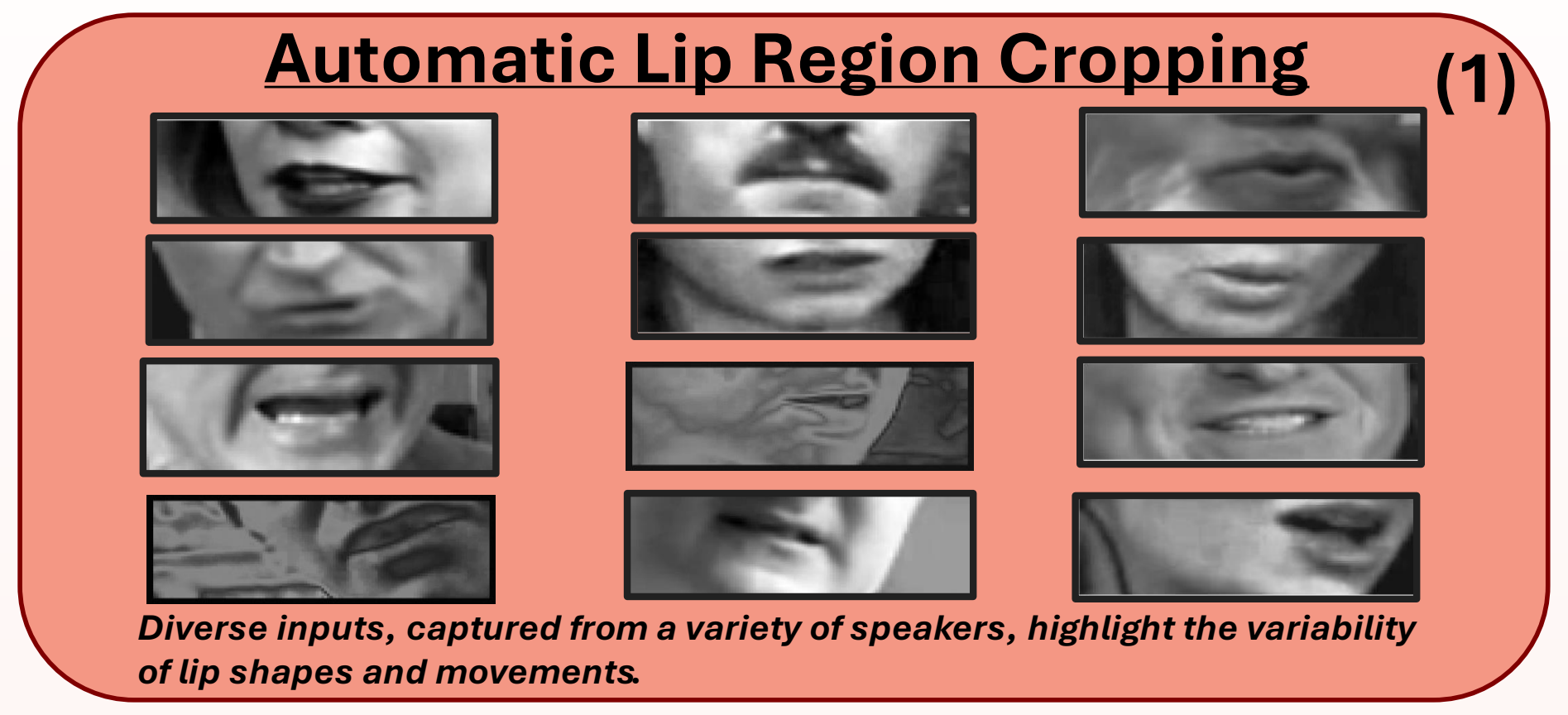
❑Then each frame is normalized, relative to the video based on their deviation from the average.

❑Next, a **lightweight temporal CRNN** model is trained using a novel architecture, along with a CTC Loss decoding function. Many variations with various hyperparameters were tested, all trained on an RTX 3060 TI.

❑The model is then deployed and evaluated on a **Raspberry Pi 5.**

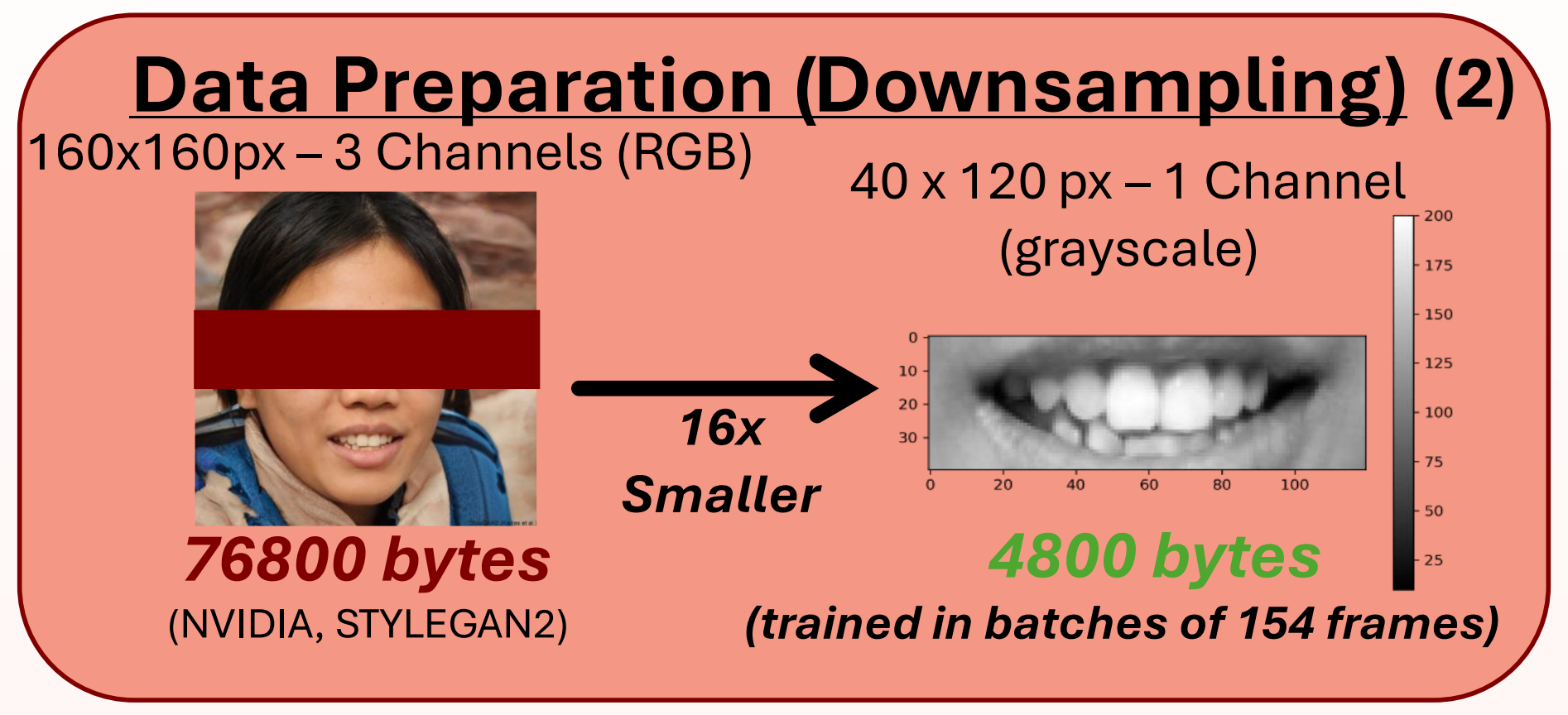
❑ The CAD model was assembled, as a proof of concept, usable device to allow for **real world testing.**

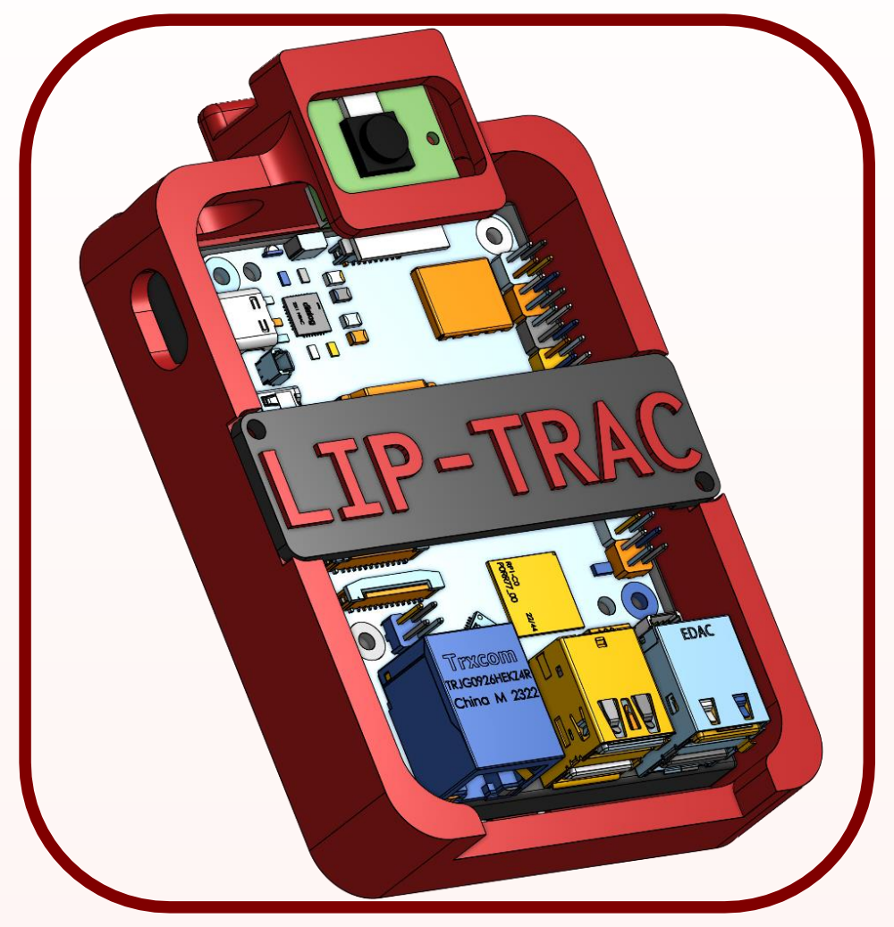
Automatic Lip Region Cropping (1)

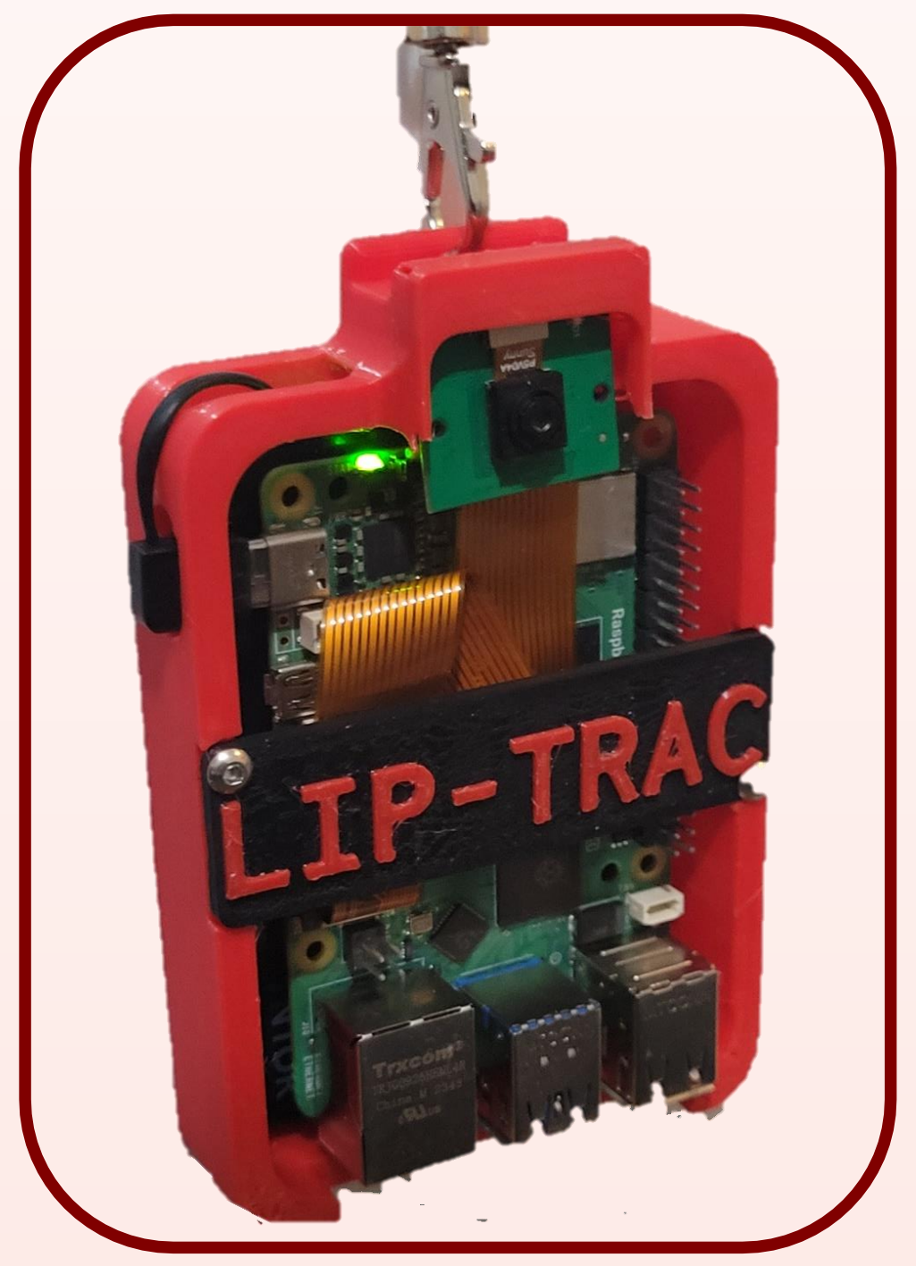


Diverse inputs, captured from a variety of speakers, highlight the variability of lip shapes and movements.

Data Preparation (Downsampling) (2)

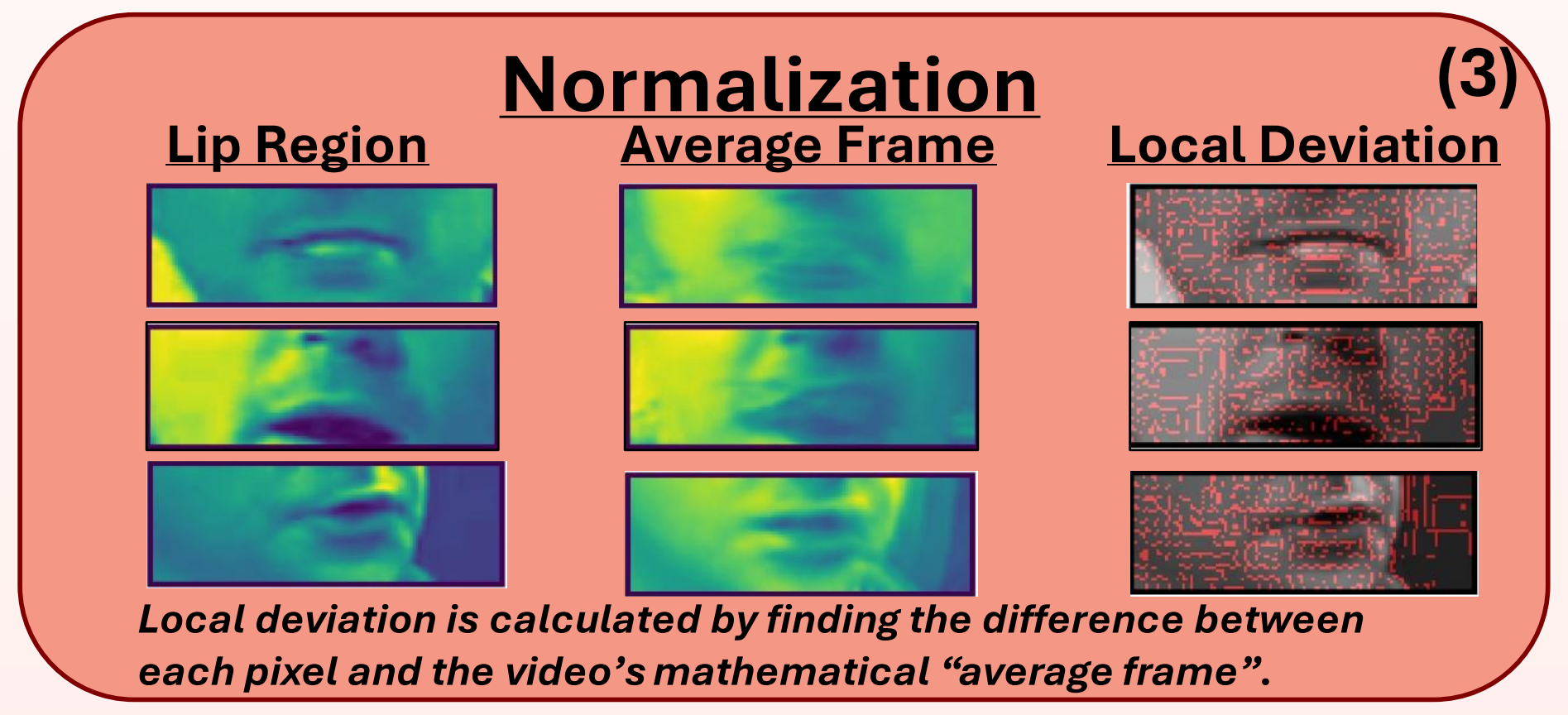

[redacted]
160x160px – 3 Channels (RGB)
40 x 120 px – 1 Channel (grayscale)
16x Smaller
76800 bytes
(NVIDIA, STYLEGAN2)
4800 bytes
(trained in batches of 154 frames)





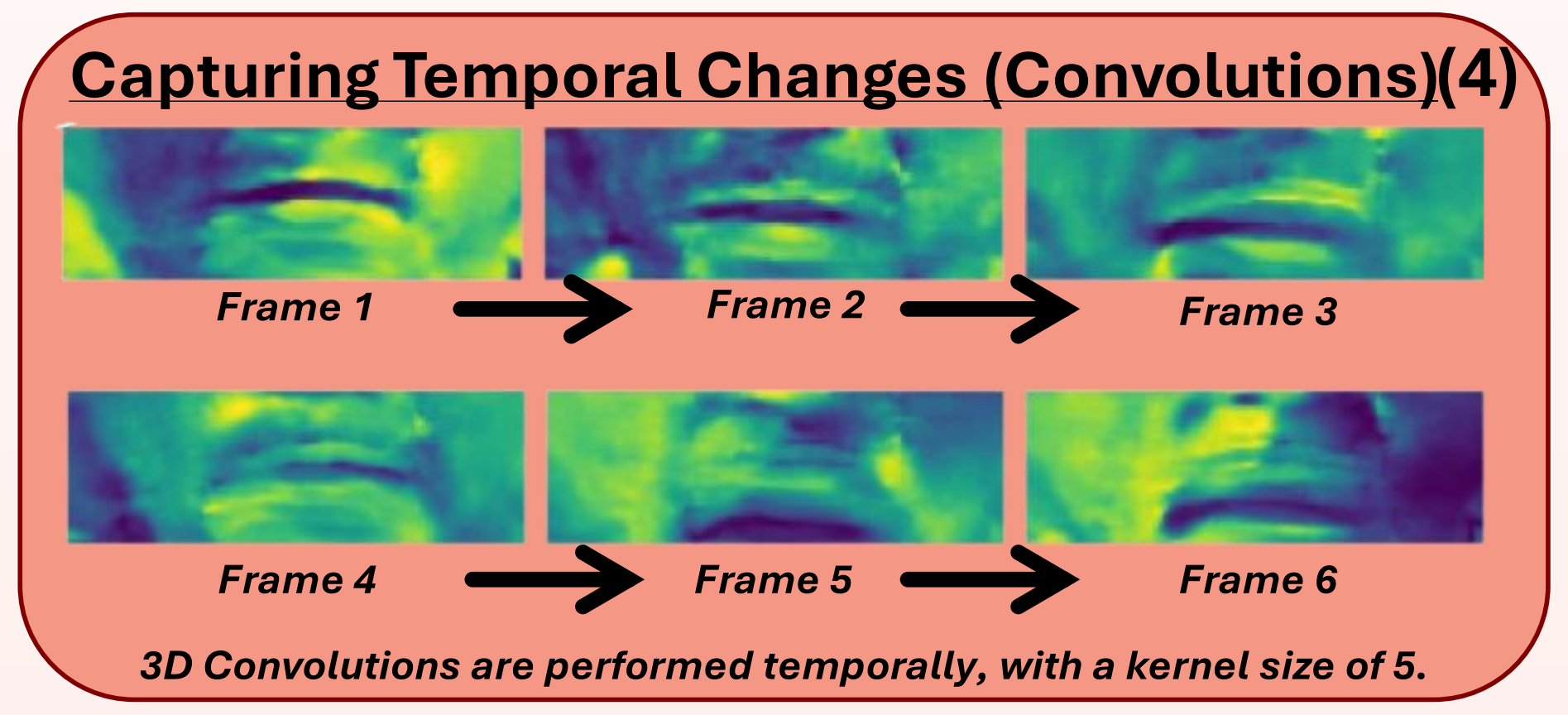
CAD & Photographs of LIP-TRAC Device (used for testing)

Normalization (3)



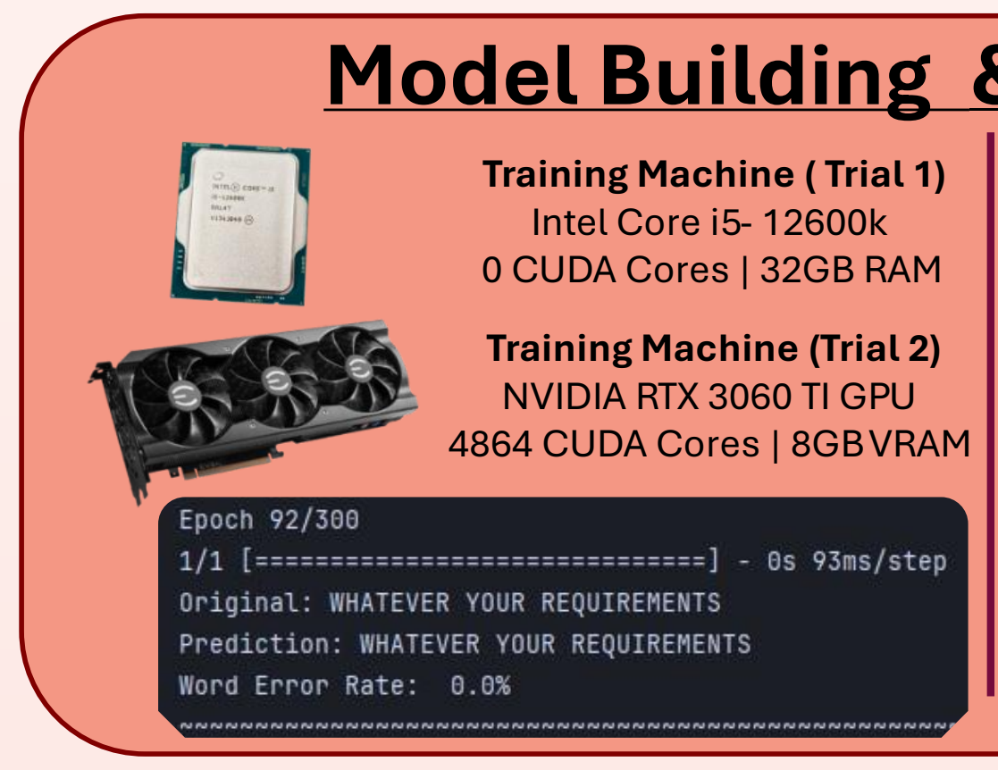
Local deviation is calculated by finding the difference between each pixel and the video's mathematical "average frame".

Capturing Temporal Changes (Convolutions)(4)



3D Convolutions are performed temporally, with a kernel size of 5.

Model Building & Training (5)



Training Machine (Trial 1)
Intel Core i5- 12600k
0 CUDA Cores | 32GB RAM

Training Machine (Trial 2)
NVIDIA RTX 3060 TI GPU
4864 CUDA Cores | 8GB VRAM

```
Epoch 92/300
1/1 [==============================] - 0s 93ms/step
Original: WHATEVER YOUR REQUIREMENTS
Prediction: WHATEVER YOUR REQUIREMENTS
Word Error Rate:  0.0%
```

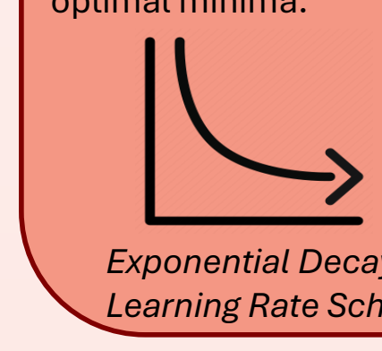
Final Model Architecture
(12 layers, 12.8million params)

1. 3D Convolution ~ 896 params
2. 3D Max Pooling
3. 3D Convolution ~ 55,360 params
4. 3D Max Pooling
5. 3D Convolution ~ 165,984 params
6. 3D Max Pooling
7. Time Distributed ~ 1,1455,488 params
8. Bidirectional GRU
9. Dropout
10. Bidirectional GRU ~ 1,145,588 params
11. Dropout
12. Dense ~ 14877 params

Model Tuning & Optimization (6)

Learning Rate

Decreases the learning rate over time, improving convergence, and reduces chances of overshooting the optimal minima.



Exponential Decay Learning Rate Scheduler

GRU vs. LSTM

The GRU layers have a higher inference speed than LSTM, making the model more practical for real-time lipreading applications

GRU (2 gates) | LSTM (3 gates)

Convolution Kernel

Due to selecting GRU's, a larger kernel was chosen in order to learn more features, and more patterns throughout the video, which increases accuracy.

3x3x3 Kernel | 5x5x5 Kernel

RESULTS & IN-SILICO VALIDATION

Fig 1. Word Level Accuracy of LIP-TRAC over 300 epochs

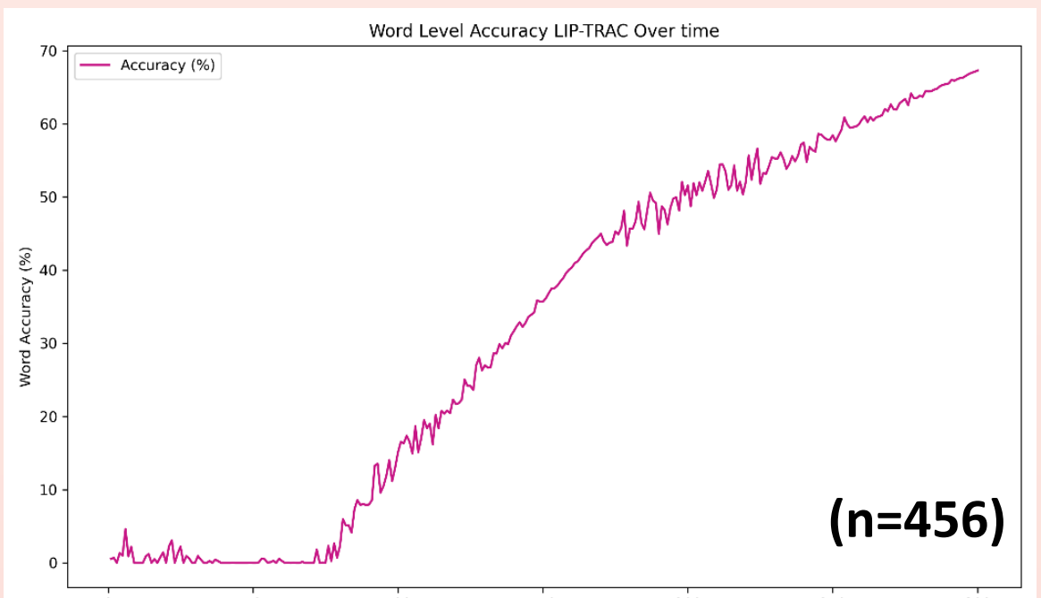


Fig 2. Character Level Accuracy Of LIP-TRAC over 300 epochs

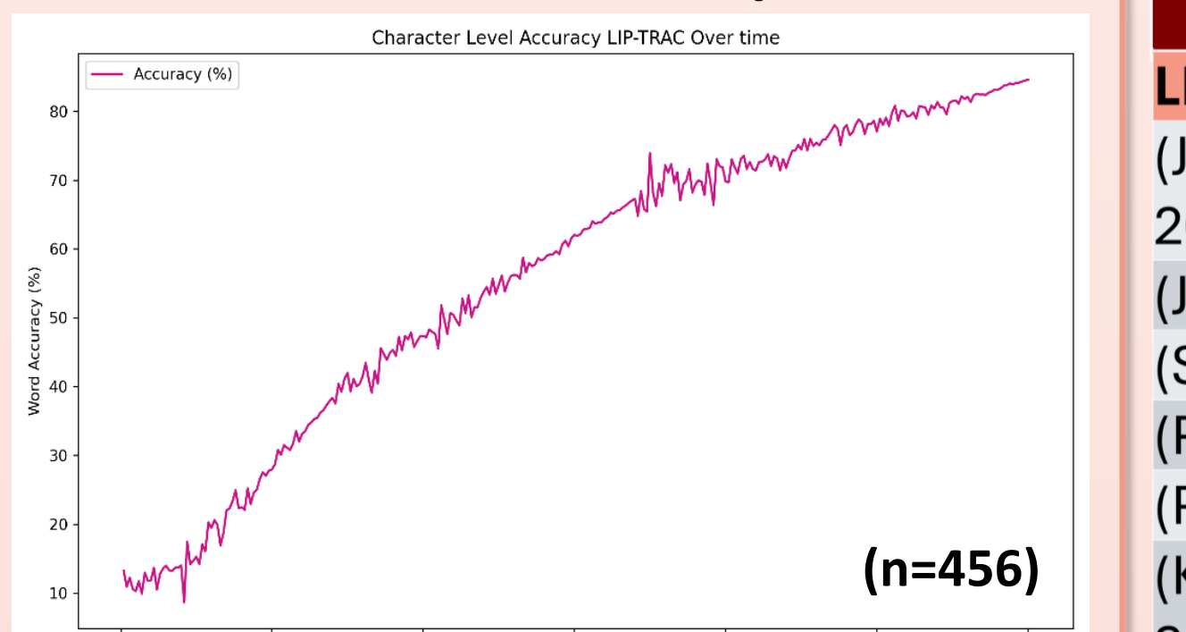


Fig 3. Word Level Accuracy of LIPTRAC vs Existing Models.

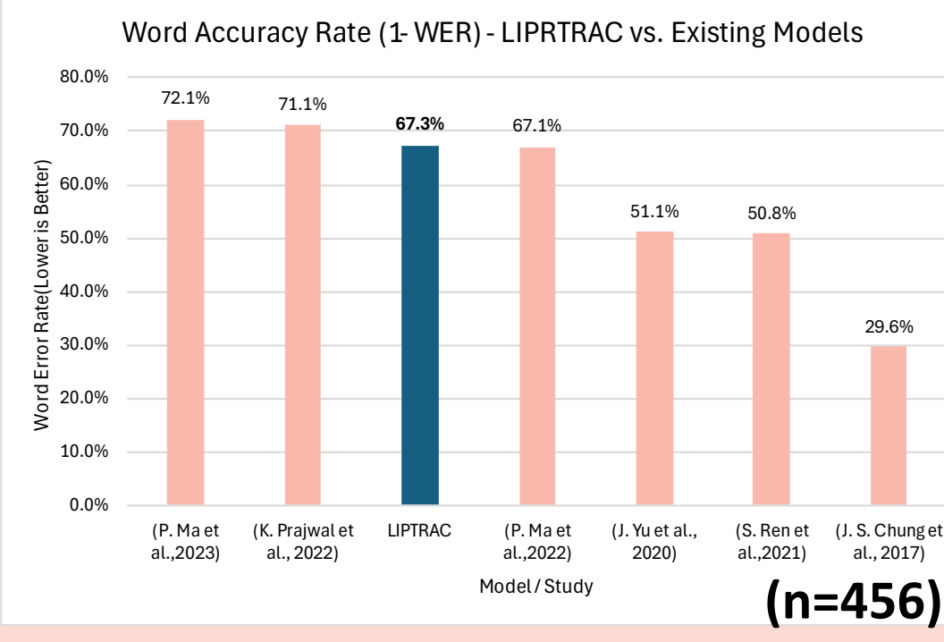
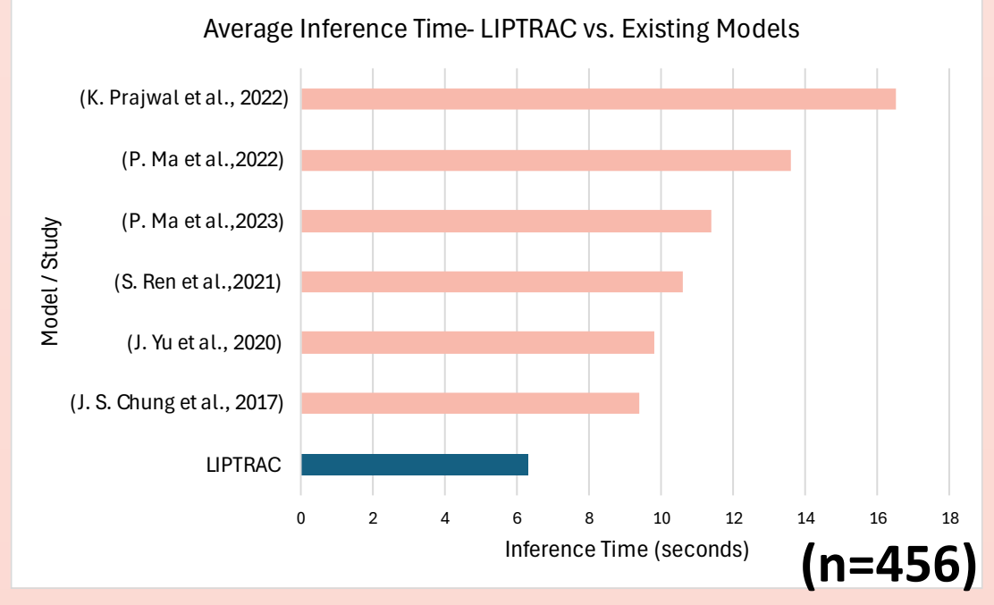


Fig 4. Avg. Inference Time of LIPTRAC vs. Existing Models



Model / Study	RTPS
LIPTRAC	**0.114444**
(J. S. Chung et al., 2017)	0.075638
(J. Yu et al., 2020)	0.069381
(S. Ren et al.,2021)	0.06156
(P. Ma et al.,2023)	0.044825
(P. Ma et al.,2022)	0.037353
(K. Prajwal et al., 2022)	0.017939

Fig. 7. RTPS of LIP-TRAC vs. Existing Models

Real Time Performance Score: Novel Metric

$$RTPS = \frac{Accuracy}{Inference\ Time}$$

❖**Real Time Performance Score (RTPS) is** a combination metric that utilizes both the accuracy (Word Level) and the inference time

❖These are the 2 most important metrics for real-world usage, and this metric **rewards low inference time** and **high accuracy.**

❖Most existing studies do not include inference time or any "real world" metric

Figure 5 & 6. Model Version 1 vs. Version 14 – Training Loss & Validation Loss Over Epochs

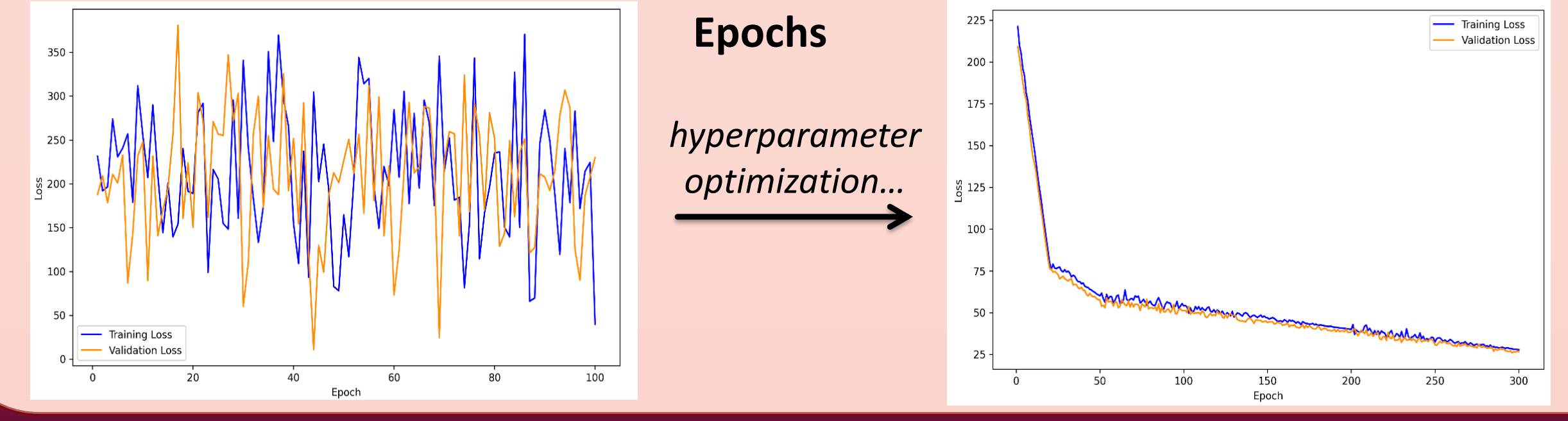
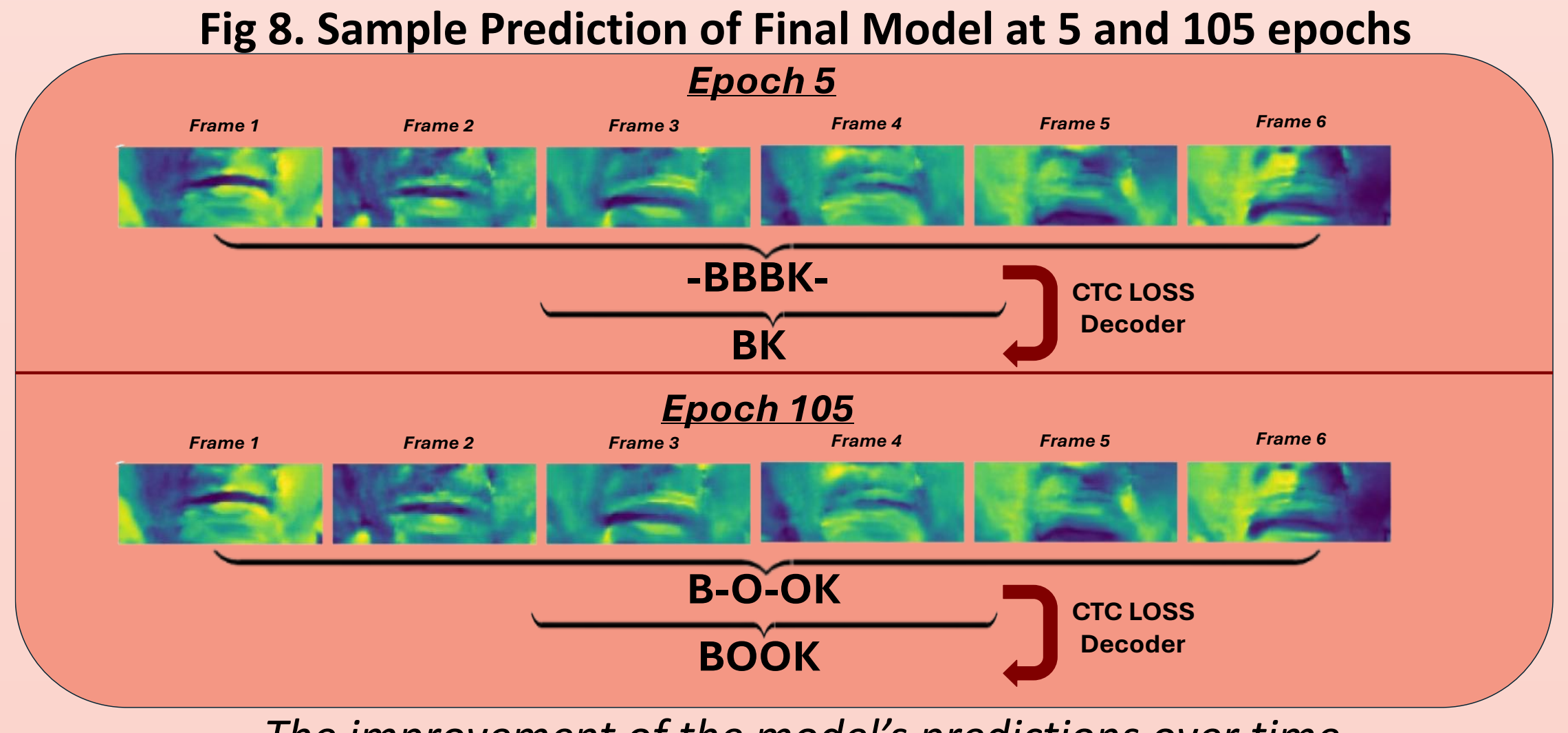


Fig 8. Sample Prediction of Final Model at 5 and 105 epochs



The improvement of the model's predictions over time.

DATA ANALYSIS

The testing data was 456 videos, and the training data was over 1000 videos from **The Oxford-BBC Lip Reading Sentences 2 (LRS2) Dataset**, which consists of videos up to a 1000 characters in length.

Although more videos were available, due to the limitations of the training machine, 1064/456 videos was chosen as the training and testing split.

CTC Loss Function

$$L = -\frac{1}{N}\sum_{i=1}^{N} \log P_{(y_i,X_i)}$$

(Graves et al.)

- ***L:*** *The average loss value for the batch*
- ***N:*** *The total number of samples in the batch*
- ***i:*** *Index representing a specific sample in the batch*
- y_i*: The true sequence of labels (text) for the "i-th" sample*
- X_i *: The input data (video frames) for the "i-th" sample*
- $P(y_{i,} X_i)$: The probability that the model assigns to the correct label sequence given the input data

By calculating the negative log probability of the correct label sequence for each input and averaging it across the batch, the model learns to map video frames to text sequences. This explicitly accounts for blank tokens allowing for multiple possible alignments between input and output sequences. For LIP-TRAC, the CTC Loss started high (~200) but decreased significantly during training, stabilizing around ~27, indicating improved alignment and predictions.

By plotting the average of loss and various accuracy metrics, against $\log(Learning\ Rate)$ with various optimizers & learning rate schedulers, "**ideal local minima**" was identified**,** and the ideal starting learning rate was determined to be $\mathbf{3.0 \times 10^{-6}}$.

Real World Trials using LIP-TRAC

- Custom tested on new sentences (to test lipreading capabilities)
- Tested with new speakers (to test multi speaker capabilities)
- **Tested on Raspberry Pi 5 with 3D Printed Prototype Frame** (to test facial recognition & LIP-TRAC design)

	Word Level Acc.	Character Level Acc.	Inference Time
Phrase 1	*61.2%*	*89.1%*	*7.1s*
Phrase 2	*83.5%*	*94.0%*	*6.6s*
Phrase 3	*59.7%*	*67.8%*	*6.4s*
Phrase 4	*67.8%*	*92.3%*	*6.3s*
Phrase 5	*56.6%*	*74.5%*	*5.8s*
Average	***65.8%***	***83.5%***	***6.4s***

The most frequent errors in the model were missing repeated characters (e.g. "helo" in place of "hello"), which is the reason the word error rate is significantly higher than the character error rate. However, in practical usage, this is not very important, as most words can be recognized in this form, with consecutive double letters merged into one.

CONCLUSIONS

This research demonstrates that a lightweight CRNN model can perform real-time lipreading, providing an accessible solution for those with hearing loss.

Revisiting Engineering Criteria:

- #1: LIP-TRAC was trained on a large variety of speakers, making it a multi-speaker model, capable of use in the real world
- #2: LIP-TRAC has an average inference time of ~ **6.3 seconds (per video)**
- #3: LIP-TRAC achieved **14% CER (<20%), 32.7% WER (<35% WER),** and RTPS OF **0.114**

Technical Impacts: This study has introduced several methods and results.

- Provides real time speech transcription, based on purely visuals, within **6.3 seconds**
- This significantly lowers costs, requiring only **training expenses and a $150 Raspberry Pi 5 for deployment.**
- Can serve as a basis for other multi speaker VSR models.

Non-Technical Impacts: LIP-TRAC can aid millions of people in the real-world.

- Serves as a **proof-of-concept** → for real time visual speech recognition.
- Highlights the speed-accuracy tradeoff within VSR modes.
- This is useful for individuals who have hearing impairments, or Aphonia. This enhances their accessibility and communication abilities, in their day-to-day life.

FUTURE WORK

- Integrate microphone to implement "validation" within real-world situations
- Word Level Prediction utilizing a dictionary rather than Character Level.
- N-Gram Implementation to predict sequence of words, utilizing context
- Utilize an attention mechanism and use entire face.

Not possible yet due to data availability:

- Test model architecture with multiple languages, or potentially one multilingual model, or with the International Phonetic Alphabet

KEY REFERENCES

Afouras, T., Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2019). Deep Audio-visual Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1. https://doi.org/10.1109/tpami.2018.2889052

Assael, Y. M., Shillingford, B., Shimon Whiteson, & Nando de Freitas. (2016). LipNet: End-to-End Sentence-level Lipreading. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1611.01599

Graves, A., Ch, A., Fernández, S., Gomez, F., Schmidhuber, J., & Ch, J. (2006). *Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks*. https://www.cs.toronto.edu/~graves/icml_2006.pdf

Ma, P., Haliassos, A., Fernandez-Lopez, A., Chen, H., Petridis, S., & Pantic, M. (2023). AUTO-AVSR: AUDIO-VISUAL SPEECH RECOGNITION WITH AUTOMATIC LABELS. https://arxiv.org/pdf/2303.14307v3

World Health Organization. (2024, February 2). *Deafness and hearing loss*. WHO; World Health Organization: WHO. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

Paper Bibliography Contains Full List