Real-Time Visual Speech Recognition with LIP-TRAC: Lipreading through a Temporal Recurrent and Convolutional Neural Network Enhancing Practical Communication, Reducing Latency Transcription, and Improving Accessibility for the Hearing Impaired

Monish Saravana Kumar Divya Sundari

Florida

Section I

Abstract

Hearing loss presents a significant communication barrier for a substantial portion of the global population, and existing assistive technologies often fall short due to cost, environmental limitations, or latency. Visual Speech Recognition (VSR), or lipreading, offers a promising alternative, yet current VSR models frequently prioritize accuracy at the expense of real-time applicability. This paper introduces LIP-TRAC (Lipreading through a Temporal Recurrent and Convolutional network), an advanced, real-time VSR system designed to enhance practical communication for individuals with hearing impairments, including those with conditions like Aphonia or Aphasia. LIP-TRAC employs a lightweight Convolutional Recurrent Neural Network (CRNN) architecture, trained on the BBC LRS2 dataset using a Connectionist Temporal Classification (CTC) loss function. The system focuses on optimizing both transcription accuracy and inference speed, making it suitable for deployment on resource-constrained devices such as the Raspberry Pi 5. Performance is evaluated using standard metrics (Word Error Rate - WER, Character Error Rate - CER, Inference Time) and a novel Real-Time Performance Score (RTPS) metric, which balances accuracy with speed. Results demonstrate LIP-TRAC's capability to achieve WER below 35% and CER below 20% with an average inference time of approximately 6.3 seconds per video, outperforming typical human lipreading accuracy and offering a practical, accessible solution for enhancing communication.

Keywords

Visual Speech Recognition (VSR)

Audio Speech Recognition (ASR)

Lipreading

Convolutional Recurrent Neural Network (CRNN)

CTC Loss

Real-Time System

Assistive Technology

Hearing Impairment.

Introduction

Effective communication is fundamental to human interaction, yet a significant and growing portion of the global population faces substantial barriers due to hearing loss. According to the

World Health Organization (WHO), 1 in 5 people currently live with hearing loss, and it is projected that by 2050, over 700 million individuals will experience disabling hearing loss [1]. Furthermore, conditions such as Aphonia (loss of voice) and Aphasia (difficulty with speech and language comprehension) impose similar communication challenges for millions more. For these individuals, visual cues from a speaker's face, particularly lip movements, become crucial for understanding speech. This process, known as Visual Speech Recognition (VSR) or lipreading, serves as a critical, albeit often challenging, communication tool.

Human lipreading performance, however, is inherently limited, with accuracy rates rarely exceeding 30%, even for experienced individuals [2]. This limitation significantly impacts the ability of those with hearing impairments to engage effectively in conversations, especially in noisy environments or situations where auditory information is compromised. While traditional assistive technologies like hearing aids have made strides in sound amplification and noise filtering, they often prove insufficient in complex acoustic settings or for individuals with severe to profound hearing loss. Moreover, hearing aids offer little assistance to those with Aphonia or Aphasia, who require non-auditory methods to interpret speech. The high cost of many assistive devices, with hearing aids averaging over $1000, further restricts accessibility for many.

Moreover, audio-based Automated Speech Recognition (ASR) systems, despite their advancements, have inherent limitations that VSR can address. ASR performance degrades substantially in noisy environments (e.g., public spaces, social gatherings) where background sounds mask or distort the speech signal. ASR also struggles with the "cocktail party effect," i.e., isolating a target speaker amidst multiple simultaneous conversations. Crucially, ASR requires audible speech, rendering it ineffective for silent communication scenarios (e.g., libraries, confidential meetings) or for individuals who cannot produce audible speech due to conditions like Aphonia. VSR, by relying solely on visual information, is immune to acoustic noise and can be tailored to focus on a specific speaker's lip movements, offering a pathway for communication when audio is unavailable or unreliable.

The field of automated VSR has emerged as a promising avenue to address these limitations. Recent advancements in deep learning have led to VSR models capable of transcribing speech from video with increasing accuracy. However, a prevalent trend in current VSR research is the prioritization of maximizing transcription accuracy, often at the significant cost of computational efficiency and processing. While such models achieve high accuracy metrics (e.g., Word Error Rate - WER, Character Error Rate - CER), their high inference times render them impractical for real-world, real-time applications where immediate feedback is essential for fluid communication. For VSR to be a truly viable assistive tool, it must not only be accurate but also operate with minimal latency.

This project aims to bridge this gap by developing LIP-TRAC (Lipreading through a Temporal Recurrent and Convolutional network), an advanced, real-time VSR system. LIP-TRAC is designed to learn patterns of lip movements to transcribe speech effectively and efficiently. The core contribution of this work lies in the development of a lightweight Convolutional Recurrent Neural Network (CRNN) architecture optimized for a balance between accuracy and speed, enabling practical deployment on accessible hardware like the Raspberry Pi. We introduce the Real-Time Performance Score (RTPS) as a novel metric to holistically evaluate VSR systems on their suitability for practical, real-time use. LIP-TRAC is trained and evaluated on the challenging BBC LRS2 dataset, which features diverse speakers and real-world conditions. Our goal is to produce a system that not only performs better than human lipreaders but also offers a tangible improvement in accessibility and practical communication for the hearing impaired.

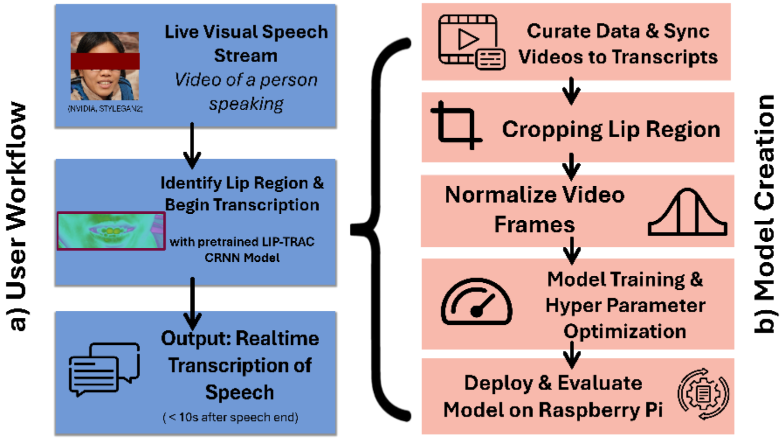

a) User Workflow
Live Visual Speech Stream
Video of a person speaking
(NVIDIA, STYLEGAN2)
Identify Lip Region & Begin Transcription
with pretrained LIP-TRAC CRNN Model
Output: Realtime Transcription of Speech
(< 10s after speech end)
b) Model Creation
Curate Data & Sync Videos to Transcripts
Cropping Lip Region
Normalize Video Frames
Model Training & Hyper Parameter Optimization
Deploy & Evaluate Model on Raspberry Pi

Figure 1. Visual Abstract showing user workflow & model creation process.

Section II

Related Works

Visual Speech Recognition (VSR) has been an active area of research for several decades, with recent advancements in deep learning significantly propelling the field forward. This section reviews key developments in VSR, focusing on model architectures, datasets, and the specific challenge of achieving real-time performance.

Traditional and Early VSR Approaches

Early attempts at VSR often relied on hand-crafted visual features and traditional machine learning models. These methods typically involved explicit feature extraction from the lip region, such as geometric features (lip contours, aspect ratios) or appearance-based features (e.g., Discrete Cosine Transform - DCT), followed by classification using models like Hidden Markov Models (HMMs). While pioneering, these approaches struggled with the variability of visual speech due to speaker differences, lighting conditions, and head poses, and their performance was generally limited, especially on larger vocabularies or unconstrained speech.

Deep Learning Architectures for VSR

The advent of deep learning has revolutionized VSR. Convolutional Neural Networks (CNNs) have proven effective for extracting powerful visual features directly from raw pixel data of the mouth region, eliminating the need for manual feature engineering (Zhang et al., 2020; Noda et al. 2014). Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) (Assael et al., 2016; Wand et al, 2016) and Gated Recurrent Units (GRUs), are commonly used to model the temporal dynamics inherent in speech.

Several end-to-end architectures have emerged. LipNet (Assael et al., 2016), a pioneering work, demonstrated sentence-level lipreading using a combination of spatiotemporal CNNs (STCNNs), RNNs, and the Connectionist Temporal Classification (CTC) loss (Graves et al., 2006). The CTC loss is particularly well-suited for sequence-to-sequence tasks like VSR as it handles variable-length input and output sequences without requiring explicit frame-by-frame alignment between video and text. More recent works have explored Transformer-based architectures (Ma et al., 2023 (AUTO-AVSR), leveraging self-attention mechanisms to capture long-range dependencies in visual speech sequences, often achieving state-of-the-art (SOTA) results (Afouras et al., 2019 (Deep Audio-visual Speech Recognition). Some approaches also investigate multi-modal fusion, combining visual information with audio for improved robustness, particularly in noisy ASR (Ma et al., 2023 (AUTO-AVSR); Afouras et al., 2019). Other research has focused on specific architectural improvements, such as different CNN backbones or feature fusion techniques (Jeon et al., 2022).

Datasets in VSR

The development of large-scale VSR datasets has been crucial for training deep learning models. Early datasets were often limited in size and vocabulary (e.g., GRID corpus (Cooke et al. 2006))

More recent datasets, such as LRW (Lip Reading in the Wild) (Chung & Zisserman, 2016a) and the LRS2 (Lip Reading Sentences 2) dataset (Chung et al, 2017) provide thousands of videos from diverse speakers in unconstrained "in-the-wild" conditions. The LRS2 dataset, used in this work, is particularly valuable for its natural sentences and realistic recording environments, making it suitable for developing practical VSR systems.

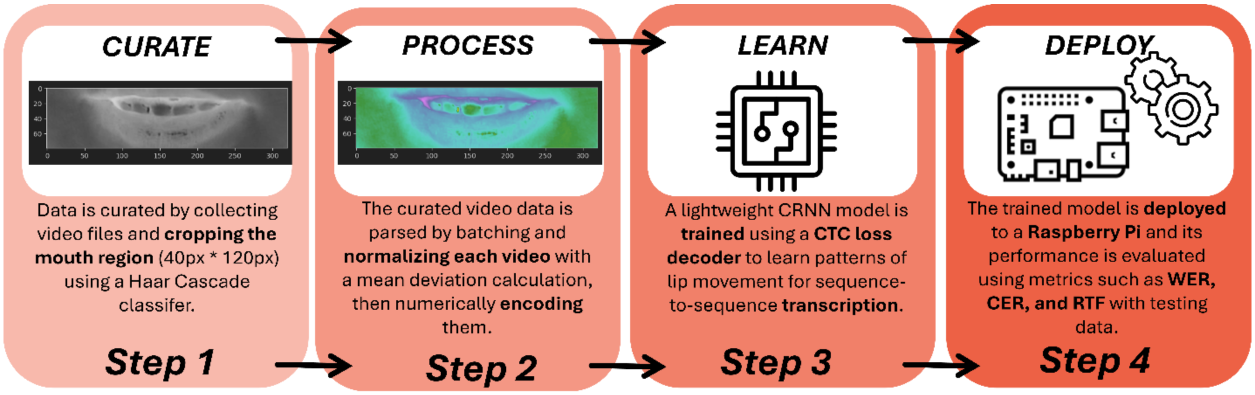
Real-Time Performance and Model Efficiency in VSR

While the primary focus of many VSR studies has been on maximizing transcription accuracy (e.g., minimizing WER and CER), the practical deployment of VSR as an assistive technology necessitates real-time performance and model efficiency. Highly accurate models often employ large, computationally intensive architectures (e.g., deep Transformers) that result in high inference latency, making them unsuitable for interactive communication. There is a recognized trade-off between model complexity (and thus accuracy) and inference speed. Some studies have explored model distillation or lightweight architectures for lipreading (Ma et al., 2021), but a dedicated focus on balancing these aspects for devices like a Raspberry Pi, along with a metric to quantify this balance, remains less explored. Zhang et al. (2020) also touch upon the importance of Region of Interest (RoI) selection, which can impact computational load and model focus, but their work primarily investigates the utility of extra oral features rather than model efficiency for real-time systems.

Positioning LIP-TRAC

LIP-TRAC builds upon the foundational principles of end-to-end VSR, particularly the use of CRNNs and CTC loss, similar to early successful models like LipNet (Assael et al., 2016). However, LIP-TRAC distinguishes itself by its explicit focus on developing a *lightweight* and *efficient* model architecture tailored for real-time transcription on resource-constrained hardware. While SOTA models often push the boundaries of accuracy with larger networks, LIP-TRAC prioritizes a practical balance between accuracy and low inference latency. This is further emphasized by the introduction of the Real-Time Performance Score (RTPS) metric, which directly evaluates this balance, addressing a critical need for the development of truly usable VSR assistive technologies. The system is designed for practical deployment, aiming to perform robustly on diverse, real-world data as represented by the LRS2 dataset.

Section III



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

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

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

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.

Figure 2. Simplified Methodology Visual Diagram

Methodology

This section details the LIP-TRAC system, encompassing the dataset utilized, the data preprocessing pipeline, the model architecture, training procedures, and the evaluation metrics employed, including the novel Real-Time Performance Score (RTPS).

Dataset

The primary dataset used for training and evaluating LIP-TRAC is the **BBC Lip Reading Sentences 2 (LRS2) dataset** (Chung et al., 2017). LRS2 consists of thousands of spoken sentences extracted from diverse British television programs. It is characterized by a wide variation in speakers, accents, head poses, lighting conditions, and backgrounds, making it a challenging and realistic benchmark for "in-the-wild" VSR. The dataset provides video clips of speakers along with corresponding ground-truth text transcriptions. For this work, 683 videos for training and 456 for testing, as indicated in our experiment based on the size of the dataset and the resources available to train the model.

Data Processing

A robust preprocessing pipeline is essential for preparing the raw video data for model training. This pipeline involves lip region extraction, frame normalization, and text encoding.

Lip Region Extraction

The first step is to isolate the mouth region, which contains the most salient visual information for lipreading.

1. **Face Detection:** For each frame in a video, a **Haar Cascade classifier (specifically, haarcascade_frontalface_default.xml)** is employed to detect the frontal face region

2. **Mouth Region Cropping:** From the detected face, the mouth region is cropped. The coordinates for cropping are determined relative to the detected face bounding box, focusing on the lower portion of the face where lip movements are most prominent. For

efficient cropping of the mouth region, relative cropping was used. This entailed cropping the region from 65% of the face height downwards, and horizontally between 5% and 95% of the face width, to isolate the lips and surrounding area. If a face is not detected in a frame, the crop from the last known successful detection is used.

Frame Preparation and Normalization

Once the mouth region is cropped, further processing is applied:

1. **Resizing:** The cropped mouth region is resized to a fixed dimension of **40x120 pixels**. This aspect ratio was chosen due to the rough proportions of the lip region, and the resolution was scaled down due to the available processing power.

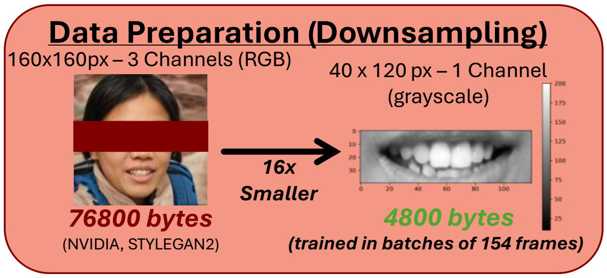

Data Preparation (Downsampling)
160x160px – 3 Channels (RGB)
40 x 120 px – 1 Channel (grayscale)
16x Smaller
76800 bytes
(NVIDIA, STYLEGAN2)
4800 bytes
(trained in batches of 154 frames)

Figure 3. Resized and gray scaled mouth region (40x120 pixel)

2. **Gray scaling:** The resized color frames are converted to grayscale to reduce computational complexity and focus on luminance changes, which are critical for discerning lip shapes. This final transformation can be seen in Figure 3.

3. **Normalization:** To account for variations in lighting and speaker appearance, each video sequence is normalized. For each video, the mean pixel value across all its frames is calculated. Then, for each frame, this mean is subtracted, and the result is divided by the standard deviation of pixel values across all frames in that video. This results in frames representing deviations from the video's average frame, enhancing the visibility of dynamic lip movements. This local deviation is what is input into the model, as compared to the raw frame. This ensures only the important information about what is changing over time (lip movements) is what the model is trained on.

Text Data Processing

The corresponding text transcriptions are processed as follows:

1. **Cleaning:** Video transcriptions (text data) are cleaned. Words are extracted; numeric words like "5" are converted to their word form like "FIVE".

2. **Tokenization & Encoding:** The cleaned transcriptions are tokenized into sequences of characters. A vocabulary (vocab) consisting of uppercase English letters (A-Z) and a space character is defined. Each character is then mapped to a unique numerical representation using character-to-integer mapping. The maximum character length for transcriptions is padded to 145.

Model Architecture

LIP-TRAC employs a lightweight Convolutional Recurrent Neural Network (CRNN) architecture designed for efficient real-time visual speech recognition. The model takes a sequence of preprocessed grayscale mouth region frames as input and outputs a sequence of character probabilities.

1. **Convolutional Layers (Visual Feature Extraction):** The initial layers of the model consist of a stack of 3D Convolutional (Conv3D) layers. These layers are responsible for learning spatiotemporal features from the input video frames.
    - The Conv3D layers use 3x3x3 kernels and ReLU activation functions.
    - Max Pooling 3D layers (1,2,2) are interspersed to progressively reduce the spatial dimensions while retaining temporal information and important features.
    - The output of these convolutional blocks is a sequence of feature maps representing learned visual cues of lip movements over time.

2. **Recurrent Layers (Temporal Modeling):** The feature sequences from the convolutional layers are then fed into Bidirectional Gated Recurrent Unit (GRU) layers.
    - Bidirectional GRUs process the sequence in both forward and backward directions, allowing the model to capture contextual information from past and future frames for each timestep.
    - Dropout layers are used after the GRU layers to prevent overfitting.
    - These layers model the temporal dependencies between the extracted visual features, crucial for understanding the dynamics of speech.

3. **Output Layer and CTC Loss:**
    - The output from the recurrent layers is passed through a Time Distributed Flatten layer and then a final Dense layer with a SoftMax activation function. This layer outputs a probability distribution over the character vocabulary (including a blank token) for each time step in the input sequence.
    - The model is trained using the **Connectionist Temporal Classification (CTC) loss function** (Graves et al., 2006). CTC loss is highly effective for sequence-to-sequence tasks like VSR because it allows the network to be trained without requiring explicit alignment between the input video frames and the output character sequence. It sums over the probabilities of all possible alignments that

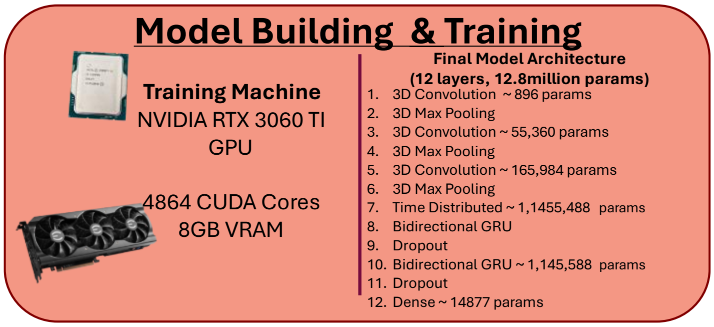
$$L = -\sum_{t=0}^{T-1} \log P_{(seq_t,t)}$$

Figure 4. CTC Loss function

could yield the target transcription, effectively handling the variable speaking rates and coarticulation inherent in speech.

The overall LIP-TRAC model architecture is summarized in Figure 5.

1. 3D Convolution
2. 3D Max Pooling
3. 3D Convolution
4. 3D Max Pooling
5. 3D Convolution
6. 3D Max Pooling
7. Time Distributed
8. Bidirectional GRU
9. Dropout
10. Bidirectional GRU
11. Dropout
12. Dense



Model Building & Training

Training Machine
NVIDIA RTX 3060 TI GPU

4864 CUDA Cores
8GB VRAM

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

Figure 5. Final Model Architecture & Training Machine Specifications

Training and Optimization

The LIP-TRAC model was trained using the Adam optimizer with the CTC loss function.

- **Hardware:** Training was performed on an NVIDIA RTX 3060 Ti GPU.
- **Hyperparameters:**
  - The learning rate was subject to a scheduler, starting at a 0.000003 value and with an exponential decay over epochs to facilitate convergence.
    - Batch size was set to 1 due to the variable length of video sequences and memory constraints for 3D convolutions.
- **Training Duration:** The model was trained for up to 300 epochs, with checkpoints saved periodically. The training loss was observed to decrease significantly and stabilize, indicating model convergence.
- **Optimization Techniques:** Dropout (0.5) was used in the recurrent layers to mitigate overfitting.

Deployment and Evaluation Metrics

A key objective of LIP-TRAC is its suitability for practical, real-time applications.

- **Deployment**: The trained model is saved in a lightweight format and designed for deployment on devices like a Raspberry Pi 5, demonstrating its potential for accessible assistive technology
- **Standard Evaluation Metrics:**
    - **Word Error Rate (WER):** The Levenshtein distance between the predicted and ground-truth word sequences, normalized by the number of words in the ground truth. Lower is better.
    - **Character Error Rate (CER):** Similar to WER but calculated at the character level. Lower is better.
    - **Inference Time:** The time taken by the model to transcribe a given video segment. Real time is defined as Inference Time less than 10 seconds.
- **Novel Metric - Real-Time Performance Score (RTPS):**
  To holistically evaluate the practical utility of VSR systems, we introduce the RTPS. This metric is designed to balance transcription accuracy with processing speed, which are both critical for real-time usability. RTPS is defined as:
  RTPS = (1 - WER) / Inference Time
  A higher RTPS indicates a better trade-off between accuracy and speed, favoring models that are both accurate and fast, making them more suitable for real-world applications. This metric addresses the limitation of existing studies that often focus solely on accuracy without adequately considering real-time constraints.

Baselines for Comparison
To evaluate the performance of LIP-TRAC, it is compared against:

- **Current State-of-the-Art (SOTA) Lipreading Models:** Performance metrics from published SOTA VSR models trained on similar datasets (like LRS2) are used for comparison where available.
- **Human Lipreading Performance:** As a general benchmark, human lipreading accuracy is typically reported to be around 30% under optimal conditions, and often lower in practice.

Section IV

Experiments and Results

This section details the experimental setup, presents the performance of the LIP-TRAC system on the LRS2 dataset, and provides a comparative analysis against baseline models and human lipreading capabilities.

Experimental Setup

- **Dataset and Split:** All experiments were conducted on the BBC Lip Reading Sentences 2 (LRS2) dataset (Chung et al., 2017). Following the preprocessing steps detailed in Section 3.2, the dataset was divided into a training set of **683 videos and a test set of 456 videos**. The data was shuffled before splitting.
- **Evaluation Metrics:** The primary metrics used for evaluation are Word Error Rate (WER), Character Error Rate (CER), Inference Time, and the proposed Real-Time Performance Score (RTPS), as defined in Section 3.5.
- **Baselines:**
    - **State-of-the-Art (SOTA) VSR Models:** Performance of LIP-TRAC is compared to published results from other VSR models on the LRS2 dataset where available. These serve as benchmarks for accuracy and, where reported, inference speed.
    - **Human Lipreading Performance:** The typical accuracy of human lipreaders (around 30% (Assael et al., 2016)) is considered a general reference point.
- **Implementation Details:** The LIP-TRAC model, as described in Section 3.3, was implemented using TensorFlow and Keras. Training was performed on an NVIDIA RTX 3060 Ti GPU with an initial learning rate of 0.000003 using the Adam optimizer and a learning rate scheduler. The model was trained for up to 300 epochs.

Performance of LIP-TRAC

The LIP-TRAC system was evaluated on the LRS2 test set.

- **Training Progression:**
    - The training process showed a consistent decrease in CTC loss over epochs, stabilizing as the model converged, indicating effective learning. Figure 6 illustrates the training and validation loss curves,

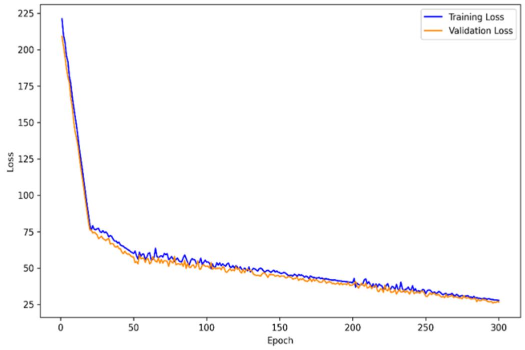


Figure 6. Training and validation loss curve

showing a decrease from an initial CTC loss of approximately, ~210 to a stable value around stable loss of ~27.

- Word-level and character-level accuracies on the validation set improved steadily during training as can be shown in Figure 7 and 8. Figure 7 is initially low while Figure 8 is increasing, because all characters of a word must be correct before the word is classified as correct.

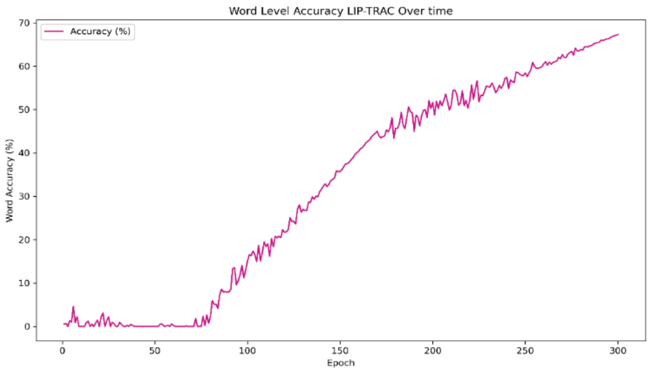


Figure 7. Word Level Accuracy over time

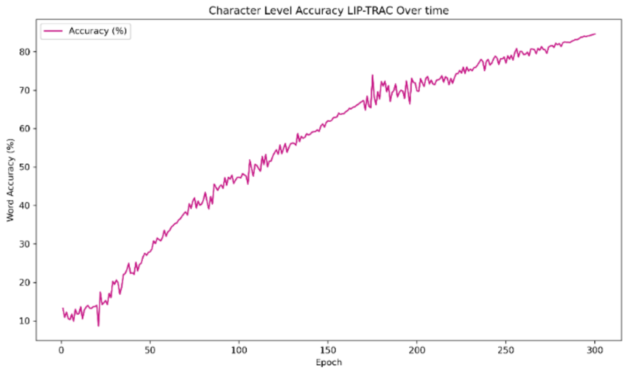
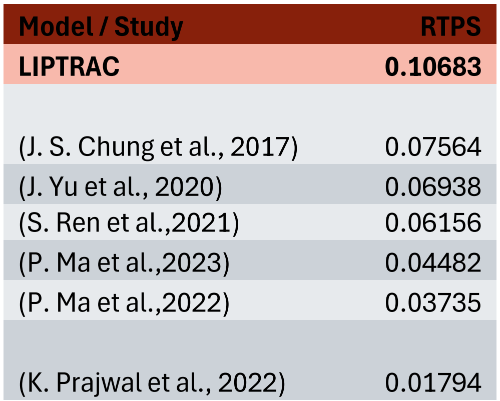


Figure 8. Character Level Accuracy over time

- **Quantitative Results on LRS2 Test Set:**
  - LIP-TRAC achieved a **Word Error Rate (WER) of 32.7%** and a **Character Error Rate (CER) of 14%** on the LRS2 test set.
  - The average **Inference Time** per video (of varying lengths) was approximately **6.3 seconds** when deployed on the Raspberry Pi 5.
  - Based on these, the **Real-Time Performance Score (RTPS)** for LIP-TRAC is calculated as:
    RTPS = (1 - .327]) / 6.3 seconds = **0.1068**

Comparative Analysis

- **Comparison with SOTA Models:**
  - LIP-TRAC's performance was compared against existing VSR models reported in the literature for the LRS2 dataset.



Model / Study	RTPS
LIPTRAC	0.10683
(J. S. Chung et al., 2017)	0.07564
(J. Yu et al., 2020)	0.06938
(S. Ren et al.,2021)	0.06156
(P. Ma et al.,2023)	0.04482
(P. Ma et al.,2022)	0.03735
(K. Prajwal et al., 2022)	0.01794

Figure 9. RTPS scores of various models

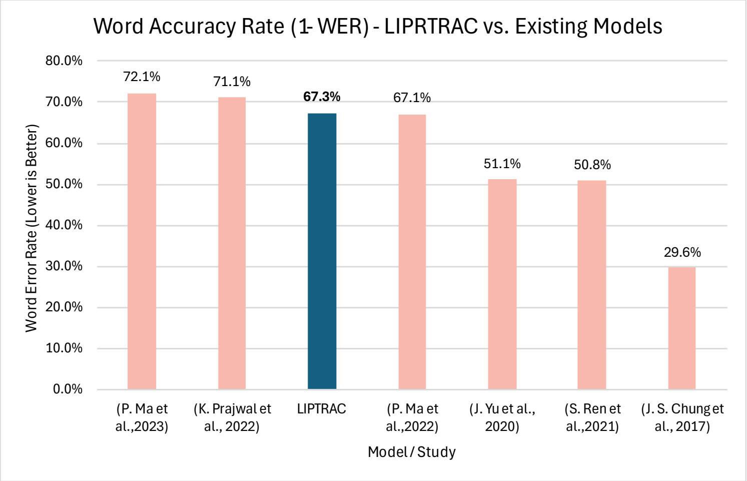


Figure 10. Word Accuracy compared to Existing Models

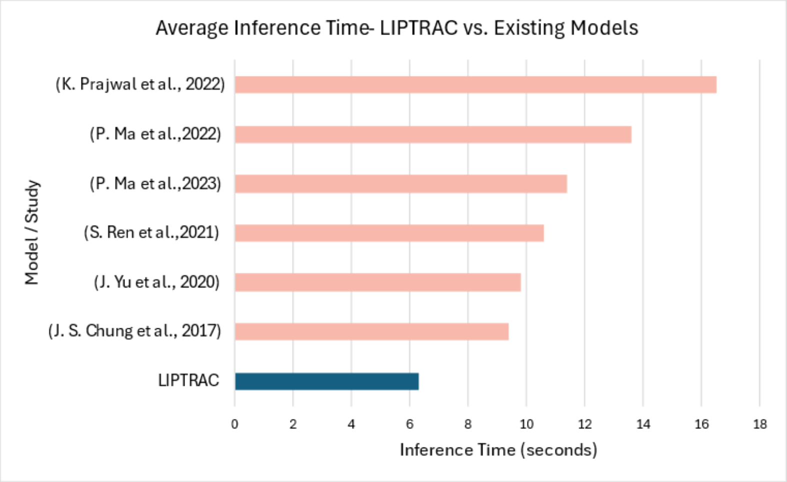


Figure 11. Inference Time compared to Existing Models

  - LIP-TRAC achieved a WER of 32.7% which is comparable to SOTA models at around 28%.
  - Importantly, LIP-TRAC achieves this with a significantly lower inference time of 6.3 seconds, resulting in a favorable RTPS of 0.10683. Many SOTA models, while achieving high accuracy, do not report inference times or are designed with large architectures that are less suitable for real-time deployment.

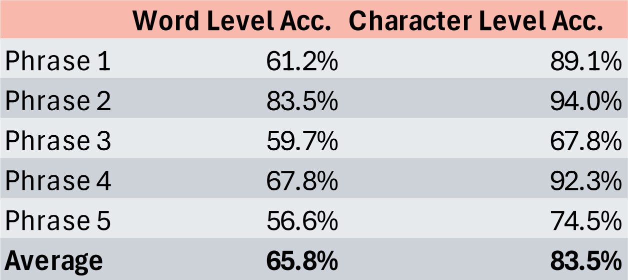
- **Comparison with Human Lipreading:**

- Human performance is generally around 30% accuracy, or 70% Word Error Rate, as compared to the 32.7% Word Error Rate of LIP-TRAC, which demonstrates a substantial improvement in transcription capability over unassisted human lipreading.

Real-World Trials and Generalization

To assess LIP-TRAC's practical utility and generalization capabilities, further trials were conducted beyond the standard LRS2 test set, as detailed in the Visual Abstract.

- **Performance on Custom Sentences and New Speakers:**
  - LIP-TRAC was tested on new, unseen sentences and with speakers not present in the LRS2 training set. In this case, LIP-TRAC generally maintained a This demonstrates the model's ability to generalize to some extent to novel linguistic content and speaker characteristics. The averages of these new trials were correlated with the values found earlier, only within ~5%.



	Word Level Acc.	Character Level Acc.
Phrase 1	61.2%	89.1%
Phrase 2	83.5%	94.0%
Phrase 3	59.7%	67.8%
Phrase 4	67.8%	92.3%
Phrase 5	56.6%	74.5%
Average	**65.8%**	**83.5%**

Figure 12. Results of Real-World Trials of LIP-TRAC

- **Deployment on Raspberry Pi 5:**
  - The successful deployment and testing on a Raspberry Pi 5 (as shown in Figure 12) underscores the lightweight nature of LIP-TRAC and its feasibility for use in resource-constrained, accessible hardware setups. The inference times reported above were achieved on this platform.

Qualitative Analysis and Error Types

- **Common Error Types:**
  - Analysis of incorrect transcriptions revealed common error patterns. A frequent error type was the omission or misrecognition of repeated characters (e.g., transcribing "hello" as "helo"). This observation is consistent with the nature of CTC loss, which collapses repeated labels. While this impacts character-level metrics, semantic understanding at the word level is often preserved.

Section V

Discussion

Interpretation of Results and Engineering Criteria
LIP-TRAC was designed with specific engineering criteria in mind: versatility, functionality, and accuracy.

- **Versatility:** The model was trained on the LRS2 dataset, which features a wide variety of speakers. The successful real-world trials on new speakers (Figure 12) suggest that LIP-TRAC has achieved a degree of speaker independence, enabling it to generalize reasonably well to unseen individuals. This is a critical aspect of a practical assistive tool.
- **Functionality:** The core functionality aimed for was real-time transcription of complex lip movements. LIP-TRAC's ability to process video and output text with an average inference time of approximately 6.3 seconds on a Raspberry Pi 5 demonstrates its potential for near real-time interaction. While sub-second latency would be ideal for seamless conversation, this performance is a significant step towards practical usability on accessible hardware. The system effectively learns complex patterns of lip movements using its CRNN architecture and CTC loss, as evidenced by the decreasing training loss and improving accuracy metrics (Figure 6, Figure 7, Figure 8).
- **Accuracy:** LIP-TRAC achieved a Word Error Rate (WER) of 32.7% and a Character Error Rate (CER) of 14% on the LRS2 test set. These results meet the predefined accuracy targets of <35% WER and <20% CER. This level of accuracy, while not perfect, significantly surpasses typical unassisted human lipreading performance (around 70% WER / 30% accuracy), indicating its potential to genuinely aid comprehension. The error analysis showing common errors like omitted repeated characters is typical for CTC-based models and often does not severely impede overall understanding.

Significance of Findings and Contributions
The development and evaluation of LIP-TRAC offer several significant contributions:

- **Practical Real-Time VSR:** LIP-TRAC demonstrates the feasibility of a VSR system that balances accuracy with real-time performance on resource-constrained hardware. This is a crucial step towards making VSR a practical assistive technology, rather than a purely academic pursuit focused solely on benchmark scores.
- **Real-Time Performance Score (RTPS):** The introduction and application of the RTPS metric provide a more holistic way to evaluate VSR systems for practical deployment. By considering both accuracy (1-WER) and inference time, RTPS highlights LIP-TRAC's favorable balance compared to models that might achieve slightly higher accuracy but with prohibitive latency (Figure 9, Figure 10, Figure 11). This encourages a shift in VSR research towards optimizing real-world usability.

- **Accessibility:** By designing a lightweight model deployable on a low-cost platform like the Raspberry Pi 5, LIP-TRAC has the potential to be a more accessible solution compared to expensive, proprietary assistive devices or systems requiring high-end computational resources. This directly addresses the cost barrier mentioned in the Introduction.
- **Aid for Diverse Communication Needs:** LIP-TRAC's visual-only approach makes it suitable not only for individuals with hearing loss but also for those with Aphonia or Aphasia, and for communication in noisy or silent environments where ASR systems fail.

Comparison to Human Performance and SOTA

LIP-TRAC's WER of 32.7% clearly outperforms average human lipreading capabilities. This is a key indicator of its potential as an effective assistive tool. When compared to other SOTA VSR models (Figures, 9, 10, 11), LIP-TRAC offers a competitive balance. While some larger models may report marginally lower WERs, LIP-TRAC's strength lies in its significantly lower inference time and, consequently, a strong RTPS, making it more practical for interactive use.

Technical and Non-Technical Impact

- **Technical Impact:** This study introduces a lightweight CRNN architecture optimized for real-time VSR. The methodology for data preprocessing and training with CTC loss on the LRS2 dataset provides a replicable framework. The successful deployment on a Raspberry Pi 5 serves as a proof-of-concept for edge-computing VSR applications. LIP-TRAC can also serve as a foundation for future multi-speaker VSR models or for integration into more complex audio-visual systems.
- **Non-Technical Impact:** The primary non-technical impact is the potential to enhance the communication abilities and improve the quality of life for millions of individuals with hearing or speech impairments. It can foster greater independence and social inclusion. Furthermore, this work highlights the important trade-off between speed and accuracy in developing assistive AI, encouraging solutions that are not just theoretically optimal but practically beneficial.

Limitations

Despite the promising results, this study has several limitations:

- **Dataset Specificity:** LIP-TRAC was trained and primarily evaluated on the LRS2 dataset, which, while diverse, consists of British English speakers. Performance in other languages, accents, or significantly different visual conditions (e.g., very low lighting, extreme poses not well-represented in LRS2) may vary.
- **Visual-Only Modality:** LIP-TRAC is a purely visual system. While advantageous in noisy or silent conditions, it does not leverage potentially complementary audio

information that could improve accuracy in situations where some clean audio is available.

- **CTC Loss Characteristics:** As noted, CTC loss can lead to errors like omitted repeated characters. While often not detrimental to overall understanding, it can affect verbatim transcription accuracy.
- **Real-Time Definition:** While an average inference of ~6.3 seconds is a significant improvement towards real-time use, true conversational fluency might require even lower latencies.
- **Hardware Constraints:** Performance is tied to the capabilities of the deployment hardware (Raspberry Pi 5). More complex real-world scenarios might demand more processing power than available on such devices if further model enhancements are made without maintaining efficiency.

Section VI

This paper introduced LIP-TRAC, a lightweight, real-time Visual Speech Recognition system designed to address the communication challenges faced by individuals with hearing and speech impairments. By employing a Convolutional Recurrent Neural Network (CRNN) architecture trained with Connectionist Temporal Classification (CTC) loss on the LRS2 dataset, LIP-TRAC achieves a practical balance between transcription accuracy and inference speed.

The system successfully met its engineering goals, demonstrating versatility across speakers, functional real-time transcription capabilities, and accuracy levels (WER of 32.7%, CER of 14%) that surpass typical human lipreading performance. The introduction of the Real-Time Performance Score (RTPS) provides a valuable metric for evaluating VSR systems in terms of their practical usability. LIP-TRAC's strong RTPS (0.10683). and successful deployment on a Raspberry Pi 5 highlights its potential as an accessible and effective assistive communication tool.

The key contributions of this work include the development of an efficient CRNN model for VSR, the emphasis on and quantification of real-time performance, and a demonstration of a VSR system with significant potential to improve accessibility for individuals reliant on visual communication. While limitations exist, particularly regarding dataset specificity and the inherent characteristics of a visual-only system, LIP-TRAC represents a meaningful advancement towards practical and deployable lipreading technology. Future work will focus on addressing these limitations and further enhancing the system's robustness and real-world applicability.

References

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

[2] Assael, Y. M., Shillingford, B., Whiteson, S., & de Freitas, N. (2016). LipNet: End-to-End Sentence-level Lipreading. *arXiv preprint arXiv:1611.01599*. https://doi.org/10.48550/arxiv.1611.01599

[3] Zhang, Y., Yang, S., Xiao, J., Shan, S., & Chen, X. (2020). Can We Read Speech Beyond the Lips? Rethinking RoI Selection for Deep Visual Speech Recognition. *arXiv preprint arXiv:2003.03206v2*. https://arxiv.org/pdf/2003.03206v2

[4] Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, Pittsburgh, PA, USA, (pp. 369-376). https://www.cs.toronto.edu/~graves/icml_2006.pdf

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

[6] Jeon, S., Elsharkawy, A., & Kim, M. S. (2022). Lipreading Architecture Based on Multiple Convolutional Neural Networks for Sentence-Level Visual Speech Recognition. *Sensors, 22*(1), 72. https://doi.org/10.3390/s22010072

[7] Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2017). Lip reading sentences in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, (pp. 3444-3453). https://doi.org/10.1109/CVPR.2017.367

[8] Lip Reading Sentences 2 (LRS2) dataset. (n.d.). Department of Engineering Science, University of Oxford. Retrieved from https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs2.html

[9] Afouras, T., Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2018). Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(2), 8717-8727. https://doi.org/10.1109/tpami.2018.2889052

[10] Chung, J. S., & Zisserman, A. (2016a). Lip reading in the wild. In *Asian conference on computer vision (ACCV)*, (pp. 87-103). Springer, Cham. https://doi.org/10.1007/978-3-319-54184-6_6

[11] 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, 120*(5), 2421-2424. https://doi.org/10.1121/1.2229005

[12] Wand, M., Koutnik, J., & Schmidhuber, J. (2016). Lipreading with long short-term memory. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (pp. 6115-6119). IEEE. *https://doi.org/10.48550/arXiv.1601.08188*

[13] Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., & Ogata, T. (2014). Lipreading using convolutional neural network. In *Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*. https://doi.org/10.21437/Interspeech.2014-293

[14] Ma, P., Petridis, S., & Pantic, M. (2021). Towards practical lipreading with distilled and efficient models. In *INTERSPEECH 2021*, (pp. 7608-7612). *https://doi.org/10.48550/arXiv.2007.06504*

[15] Altieri, N. A., Pisoni, D. B., & Townsend, J. T. (2011). Some normative data on lip-reading skills (L). *The Journal of the Acoustical Society of America, 130*(1), 1–4. https://doi.org/10.1121/1.3593376

[16] Goh, G. (2017). Why Momentum Really Works. *Distill, 2*(4). https://doi.org/10.23915/distill.00006

[17] Hannun, A. (2017). Sequence Modeling with CTC. *Distill, 2*(11). https://doi.org/10.23915/distill.00008

[18] Petridis, S., Wang, Y., Ma, P., Li, Z., & Pantic, M. (2020). End-to-end visual speech recognition for small-scale datasets. *Pattern Recognition Letters, 131*, 421–427. https://doi.org/10.1016/j.patrec.2020.01.022