Since 1962

Vivekanand Education Society's

Institute of Technology

An Autonomous Institute Affiliated to University of Mumbai
Hashu Advani Memorial Complex, Collector Colony, Chembur East, Mumbai - 400074.

Department of Artificial Intelligence and Data Science

Project Report on

Visual Speech Recognition using AI

In partial fulfillment of the Fourth Year, Bachelor of Engineering (B.E.) Degree in **Artificial Intelligence and Data Science** at **Vivekanand Education Society's Institute of Technology,** an Autonomous Institute Affiliated to University of Mumbai. Academic Year 2023-2024

Submitted by
Rupesh Dhirwani,
Arya Kurup,
Tejas Patne,
Akshat Tiwari

Project Mentors
Dr. (Mrs.) M. Vijayalakshmi
Mrs. Mamata Choudhari

(2023-24)

Vivekanand Education Society's

Institute of Technology

Since 1962

An Autonomous Institute Affiliated to University of Mumbai
Hashu Advani Memorial Complex, Collector Colony, Chembur East, Mumbai - 400074.

Department of Artificial Intelligence and Data Science

Project Report on

Visual Speech Recognition Using AI

Certificate Of Approval

This is to certify that ***Rupesh Dhirwani, Arya Kurup, Tejas Patne, Akshat Tiwari*** of Fourth Year, Department of Artificial Intelligence and Data Science have satisfactorily completed the project on "**Visual Speech Recognition using AI**" as a part of their coursework of PROJECT-I for Semester-VII under the guidance of their mentor **Dr. (Mrs.) M. Vijayalakshmi** and Co-Guide, **Mrs. Mamata Choudhari** in the year 2023-2024 .

Date-_______________

Head of the Department
(Dr. (Mrs.) M. Vijayalakshmi

Project Mentor
Dr. (Mrs.) M. Vijayalakshmi

Vivekanand Education Society's

Institute of Technology

Since 1962

An Autonomous Institute Affiliated to University of Mumbai
Hashu Advani Memorial Complex, Collector Colony, Chembur East, Mumbai - 400074.

Department of Artificial Intelligence and Data Science

Certificate of Approval

This is to certify that ***Rupesh Dhirwani, Arya Kurup, Tejas Patne, Akshat Tiwari*** of Fourth Year, Department of Artificial Intelligence and Data Science have satisfactorily completed the project on "**Visual Speech Recognition using AI**" as a part of their coursework of PROJECT-I for Semester-VII under the guidance of their mentor **Dr. (Mrs.) M. Vijayalakshmi** and Co-Guide, **Mrs. Mamata Choudhari** in the year 2023-2024 .

Date

Internal Examiner

External Examiner

Project Mentor

Head of the Department

Principal

Vivekanand Education Society's

Institute of Technology

An Autonomous Institute Affiliated to University of Mumbai
Hashu Advani Memorial Complex, Collector Colony, Chembur East, Mumbai - 400074.

Since 1962

Department of Artificial Intelligence and Data Science

DECLARATION

We, ***Rupesh Dhirwani, Arya Kurup, Tejas Patne, Akshat Tiwari*** from ***D16AD***, declare that this project represents our ideas in our own words without plagiarism and wherever others' ideas or words have been included, we have adequately cited and referenced the original sources.
We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our project work.

We declare that we have maintained a minimum 75% attendance, as per the University of Mumbai norms.

We understand that any violation of the above will be cause for disciplinary action by the Institute.

Yours Faithfully

1. Rupesh Dhirwani (10)

2. Arya Kurup (33)

3. Tejas Patne (41)

4. Akshat Tiwari (62)

ACKNOWLEDGEMENT

We are thankful to our college Vivekanand Education Society's Institute of Technology for considering our project and extending help at all stages needed during our work of collecting information regarding the project.

We are deeply indebted to the Head of the AI and Data Science Department and our project guide, **Dr. (Mrs.) M. Vijayalakshmi** and our co-guide, **Mrs. Mamata Choudhari** for her kind help and valuable advice during the development of project synopsis and for her guidance and suggestions.
It gives us immense pleasure to express our deep and sincere gratitude to our Principal **Dr. (Mrs.) J.M. Nair**, for giving us this valuable opportunity to do this project.

We express our hearty thanks to them for their assistance without which it would have been difficult in finishing this project synopsis and project review successfully.

We convey our deep sense of gratitude to all teaching and non-teaching staff for their constant encouragement, support and selfless help throughout the project work. It is a great pleasure to acknowledge the help and suggestion, which we received from the Department of AI and Data Science. We wish to express our profound thanks to all those who helped us in gathering information about the project. Our families too have provided moral support and encouragement several times.

Table of Contents

ABSTRACT

Visual Speech Recognition Using AI (VSR) is a cutting-edge technology that holds the potential to revolutionize various applications, from human-computer interaction to accessibility and security.

This interdisciplinary field explores the fusion of visual and auditory information to decipher speech, enabling machines to understand spoken language by analyzing the movements of a speaker's lips, face, and surrounding context. In this rapidly evolving domain, deep learning techniques, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have played a pivotal role in extracting intricate visual features and aligning them with corresponding audio signals.

This abstract highlights the challenges and advancements in VSR, emphasizing the importance of robust models that can adapt to diverse speaking styles, lighting conditions, and language variations.

Additionally, VSR has shown promise in various practical applications, such as enhancing automatic speech recognition, enabling silent speech interfaces, and improving accessibility for individuals with speech impairments.

However, several technical hurdles, like real-time processing and data privacy concerns, remain to be addressed. This abstract provides a glimpse into the evolving landscape of Visual Speech Recognition Using AI, where innovative solutions are poised to shape the future of human-machine communication.

Chapter 1

Introduction

1. Introduction

1.1 Motivation

The motivation for the project titled "Visual Speech Recognition Using AI" is multifaceted. Firstly, it serves as a means to enhance accessibility, offering people with hearing impairments a tool to understand spoken language through lip movements. Additionally, it facilitates improved communication for those with speech disorders or individuals who cannot speak. In the realm of security and surveillance, AI-driven lip reading can play a crucial role in monitoring and analyzing conversations in public areas, bolstering national security and public safety. Furthermore, it has the potential to revolutionize human-computer interaction by creating more intuitive interfaces with technology. This can find applications in voice assistants, robots, and smart devices. The project also contributes to the development of multimodal AI systems, which combine audio and visual information for more accurate speech recognition. In the realm of assistive technology, it can assist people with disabilities in their daily lives, translating lip movements into text or speech. Moreover, the project can break down language barriers, aiding cross-cultural communication, and has potential applications in healthcare, education, and research. Challenges like operating in noisy environments provide a compelling incentive, and the rapidly advancing field of AI and computer vision technologies makes it an exciting and pertinent area of exploration. Overall, "Visual Speech Recognition Using AI" is motivated by its capacity to improve accessibility, communication, security, and multiple other domains while addressing real-world challenges.

1.2 Problem Definition

The problem definition for the project revolves around the development of a robust and accurate visual speech recognising system, employing Artificial Intelligence (AI) techniques. This endeavor is driven by the need to transcribe spoken language by analyzing lip movements and visual cues, with the goal of addressing numerous real-world challenges and applications. Key challenges in this project include the necessity to ensure high accuracy and robustness, especially in varying environmental conditions and with different accents and languages. Acquiring and annotating a diverse dataset of lip movements and corresponding audio poses a significant data-related challenge. Moreover, integrating multiple modalities to enhance lip reading accuracy and adaptability, such as combining lip movements with audio and context, is a central aspect of this effort. The project also aims to tackle cross-language and cross-cultural variability, real-time processing requirements, ethical considerations regarding privacy, and user-friendliness. The ultimate objectives include developing an AI model that excels in transcribing spoken language from lip movements, while being adaptable and user-friendly, and effectively addressing the challenges inherent in this multifaceted task.

1.3 Relevance of the Project

The project holds significant relevance in several critical domains, making it a valuable and impactful endeavor.

Firstly, it addresses the pressing need for improved accessibility. By providing a means for people with hearing impairments to comprehend spoken language through lip movements, the project enhances their quality of life and participation in various aspects of society, such as education, employment, and social interactions.

Secondly, the project contributes to enhanced security and surveillance. AI-powered lip reading can be invaluable in monitoring and analyzing conversations in public spaces, airports, or crowded areas, thereby bolstering national security efforts and public safety measures.

Furthermore, the integration of lip reading with AI can revolutionize human-computer interaction. This technology can make voice assistants, robots, and smart devices more intuitive and efficient, bridging the gap between humans and technology.

In the realm of assistive technology, the project offers substantial relevance, assisting individuals with speech disorders or those who cannot speak to express themselves and communicate effectively, thereby fostering independence and improved well-being.

Moreover, this technology has the potential to facilitate cross-cultural communication by transcending language barriers. It can break down linguistic obstacles and create a more inclusive and connected global society.

In healthcare, it can be deployed for monitoring patients' well-being, particularly those with speech difficulties or individuals who are unconscious. This enhances the quality of healthcare delivery and patient care.

Additionally, the development of "Visual Speech Recognition Using AI" contributes to advancements in the fields of linguistics, psychology, and AI, fostering a deeper understanding of human communication and aiding in the research and development of multimodal communication systems.

Overall, the project's relevance lies in its potential to improve accessibility, communication, security, and various other domains, offering solutions to real-world challenges and positively impacting society.

1.4 Methodology used

The methodology used in the Project "Visual Speech recognition using AI" involves LipNet architecture is a structured approach designed to accurately transcribe spoken language through lip movements and audio cues. It begins with data collection, comprising video recordings of individuals speaking across different languages, accents, and environmental conditions. This data is meticulously synchronized with corresponding audio, then subjected to preprocessing to ensure precise alignment between visual cues and spoken language.

Feature extraction plays a pivotal role, focusing on capturing critical information from the video frames, including lip and facial movements. Techniques such as facial landmark detection and optical flow analysis are employed to extract these essential visual features. The core of the methodology involves the utilization of LipNet, a specialized deep learning architecture.

LipNet combines convolutional neural networks (CNNs) and Sequence Learning to comprehensively analyze and predict spoken language from lip movements. Training LipNet on the preprocessed data allows it to learn intricate patterns and relationships between visual and audio features, ensuring accurate transcription and understanding.

Rigorous validation processes are employed to ensure the LipNet model's accuracy and generalization. Separate datasets are used for validation, and evaluation metrics, including accuracy, precision, recall, and the F1 score, are applied to gauge its effectiveness in recognizing spoken language from visual cues.

Fine-tuning the model is essential to enhance accuracy, especially in challenging conditions and diverse environments. Attention is given to achieving real-time processing capabilities, ensuring the system can transcribe spoken language from visual cues as it unfolds.

Chapter 2

Literature Survey

2.1 Papers studied

2. Literature Survey

2.1 Papers studied

2.1.1 Nikita Deshmukh,Anamika Ahire,Smriti H Bhandari,Apurva Mali. Vision based Lip Reading System using Deep Learning. 2021 International Conference on Computing, Communication and Green Engineering (CCGE) JSPM's RSCOE, Pune, India. 2021.

i. This paper presents the method for the Vision based Lip Reading system that uses convolutional neural network (CNN) with attention-based Long Short-Term Memory (LSTM).
ii. The dataset includes video clips pronouncing single digits. The pretrained CNN is used for extracting features from preprocessed video frames which then are processed for learning temporal characteristics by LSTM.
iii. The SoftMax layer of architecture provides the result of lip reading. In the present work experiments are performed with two pre-trained models namely VGG19 and ResNet50 and the results are compared.
iv. To further improve the performance of the system ensembled learning is also used. The system provides 85% accuracy using ResNet50 and ensemble learning

2.1.2 Joon Son Chung, Andrew Senior, Oriol Vinyals, Andrew Zisserman. Lip Reading Sentences in the Wild. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6447-6456, 2017.

i. The goal of this work is to recognise phrases and sentences being spoken by a talking face, with or without the audio.
ii. Unlike previous works that have focussed on recognising a limited number of words or phrases, it tackles lip reading as an open-world problem – unconstrained natural language sentences, and in the wild videos.
iii. Model outputs at the character level, is able to learn a language model, and has a novel dual attention mechanism that can operate over visual input only, audio input only, or both.
iv. A 'Watch, Listen, Attend and Spell' (WLAS) network that learns to transcribe videos of mouth motion to characters.
v. A curriculum learning strategy to accelerate training and to reduce overfitting.
vi. a 'Lip Reading Sentences' (LRS) dataset for visual speech recognition, consisting of over 100,000 natural sentences from British television.

2.1.3 Pingchuan Ma, Stavros Petridis, Maja Pantic. Visual Speech Recognition Using AI for Multiple Languages in the Wild. arXiv:2202.13084, 2022

i. Visual speech recognition (VSR) aims to recognize the content of speech based on lip movements, without relying on the audio stream.
ii. Advances in deep learning and the availability of large audio-visual datasets have led to the development of much more accurate and robust VSR models

than ever before. However, these advances are usually due to the larger training sets rather than the model design.

iii. Approach revolves around addition of prediction-based auxiliary tasks to a VSR model, appropriate data augmentations and hyperparameter optimization of an existing architecture.

iv. It shows that combining multiple datasets further improves the performance.

v. It proposes two new metrics "Lip-Sync Error-Distance" (lower is better) and "Lip-Sync Error-Confidence" (higher is better), that can reliably measure the lip-sync accuracy in unconstrained videos. We see that the lip-sync accuracy of the videos generated using Wav2Lip is almost as good as real synced videos.

2.1.4 K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, C.V. Jawahar; Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13796-13805, 2020

i. This approach generates accurate lip-sync by learning from an "already well-trained lip-sync expert".

ii. Unlike previous works that employ only a reconstruction loss or train a discriminator in a GAN setup, we use a pre-trained discriminator that is already quite accurate at detecting lip-sync errors.

iii. It shows that fine-tuning it further on the noisy generated faces hampers the discriminator's ability to measure lip-sync, thus also affecting the generated lip shapes.

iv. It also employs a visual quality discriminator to improve the visual quality along with the sync accuracy.

2.1.5 Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, Nando de Freitas. LipNet: End-to-End Sentence-level Lipreading. arXiv:1611.01599v2.

i. Empirical results on the GRID corpus (Cooke et al., 2006), one of the few public sentence-level datasets, show that LipNet attains a 95.2% sentence-level word accuracy, in a overlapped speakers split that is popular for benchmarking lip reading methods.

ii. The previous best accuracy reported on an aligned word classification version of this task was 86.4% (Gergen et al., 2016). Furthermore, LipNet can generalize across unseen speakers in the GRID corpus with an accuracy of 88.6%.

iii. It also compares the performance of LipNet with that of hearing-impaired people who can lip-read on the GRID corpus task. On average, they achieve an accuracy of 52.3%, in contrast to LipNet's 1.69× higher accuracy in the same sentences.

iv. The end-to-end model eliminates the need to segment videos into words before predicting a sentence. LipNet requires neither hand-engineered spatiotemporal visual features nor a separately-trained sequence model.

2.1.6 Themos Stafylakis, Georgios Tzimiropoulos. Combining Residual Networks with LSTMs for Lipreading. arXiv:1703.04105.

I. They propose an end-to-end deep learning architecture for word-level visual speech recognition. Their system is a combination of spatiotemporal convolutional, residual and bidirectional Long Short-Term Memory networks.
II. We train and evaluate it on the Lipreading In-The-Wild benchmark, a challenging database of 500-size target-words consisting of 1.28sec video excerpts from BBC TV broadcasts.
III. The proposed network attains word accuracy equal to 83.0%, yielding 6.8% absolute improvement over the current state-of-the-art, without using information about word boundaries during training or testing.

2.1.7 Themos Stafylakisa, Muhammad Haris Khana, Georgios Tzimiropoulosa. Pushing the boundaries of audiovisual word recognition using Residual Networks and LSTMs. Elsevier, Science-Direct, Computer Vision and Image Understanding, Volumes 176–177, November–December 2018, Pages 22-32.

I. They present a deep learning architecture for lipreading and audiovisual word recognition, which combines Residual Networks equipped with spatiotemporal input layers and Bidirectional LSTMs.
II. The lipreading architecture attains 11.92% misclassification rate on the challenging Lipreading-In-The-Wild database, which is composed of excerpts from BBC-TV, each containing one of the 500 target words.
III. Audiovisual experiments are performed using both intermediate and late integration, as well as several types and levels of environmental noise, and notable improvements over the audio-only network are reported, even in the case of clean speech.
IV. A further analysis on the utility of target word boundaries is provided, as well as on the capacity of the network in modeling the linguistic context of the target word. Finally, they examine difficult word pairs and discuss how visual information helps towards attaining higher recognition accuracy.

2.1.8 Konstantinos Vougioukas, Pingchuan Ma , Stavros Petridis and Maja Pantic. Video-Driven Speech Reconstruction using Generative Adversarial Networks. Inproceedings, Interspeech September 2019.

I. They present an end-to-end temporal model capable of directly synthesising audio from silent video, without needing to transform to-and-from intermediate features.
II. Their approach is based on GANs and is capable of producing natural sounding, intelligible speech which is synchronised with the video.The performance of the model is evaluated on the GRID dataset for both speaker dependent and speaker independent scenarios.
III. It's the first method that maps video directly to raw audio and the first to produce intelligible speech when tested on previously unseen speakers.

IV. We evaluate the synthesised audio not only based on the sound quality but also on the accuracy of the spoken words.

2.1.9 Minsu Kim, Joanna Hong, Yong Man Ro. Lip to Speech Synthesis with Visual Context Attentional GAN. arXiv:2204.01726v1 [cs.CV] 4 Apr 2022.

I. They propose VCA-GAN which synthesizes the speech from local lip visual features by finding a mapping function of viseme-to-phoneme, while global visual context is embedded into the intermediate layers of the generator to clarify the ambiguity in the mapping induced by homophene.
II. A visual context attention module is proposed where it encodes global representations from the local visual features, and provides the desired global visual context corresponding to the given coarse speech representation to the generator through audio-visual attention.
III. In addition to the explicit modelling of local and global visual representations, synchronization learning is introduced as a form of contrastive learning that guides the generator to synthesize a speech in sync with the given input lip movements
IV. Extensive experiments demonstrate that the proposed VCA-GAN outperforms existing state-of-the-art and is able to effectively synthesize the speech from multi-speaker that has been barely handled in the previous works.

2.1.10 Gaoyan Zhang and Yuanyao Lu, Research on a Lip Reading Algorithm Based on Efficient-GhostNet, Journals, Electronics, Volume 12, Issue 5, 10.3390/electronics12051151.

I. This paper optimizes and improves GhostNet, a lightweight network, and improves on it by proposing a more efficient GhostNet, which achieves performance improvement while reducing the number of parameters through a local cross-channel interaction strategy, without dimensionality reduction.
II. The improved Efficient-GhostNet is used to perform lip spatial feature extraction, and then the extracted features are inputted to the GRU network to obtain the temporal features of the lip sequences, and finally for prediction.
III. They used Asian volunteers for the recording of the dataset in this paper, used the angle transformation of the dataset to deflect the recording process of the recorder by 15 degrees each to the left and right, improved the generalization ability of the model so that the model can be more consistent with recognition scenarios in real life. Efficient-GhostNet + GRU model can achieve the purpose of reducing the number of parameters with comparable accuracy.

Chapter 3

Requirements

3.1 Functional Requirements

3.2 Non-Functional Requirements

3.3 Constraints

3.4 Hardware & Software Requirements

3.5 System Block Diagram

3. Requirements

3.1 Functional Requirements

1. The system should preprocess data, aligning video frames with audio and standardizing visual cues for precise correspondence with spoken language.
2. The system must extract critical features from video frames, including lip and facial movements.
3. The system should support practical testing in real-world scenarios, including applications in assistive technology, security.
4. The system should generate a confidence score for each transcription, indicating the likelihood that the spoken language has been accurately transcribed.

Here are some additional requirements that may be important for some projects :

1. The system should be capable of recognizing spoken language from lip movements, irrespective of the language being spoken.
2. The system should be able to differentiate between spoken language for transcription and other audio or non-speech sounds.
3. The system should be able to transcribe spoken language in multiple languages, accents, and under diverse environmental conditions.
4. In scenarios like video conferencing, the system should be capable of identifying speakers to associate their speech with them.
5. The system should be able to perform real-time processing of incoming video and audio data, enabling live transcription of spoken language.

Here are some examples of how the system could be used :

1. The system could be employed to transcribe spoken language for individuals with hearing impairments, providing them with real-time subtitles during conversations, video calls, or public speeches.
2. Content creators could use the system to automatically generate subtitles or transcripts for videos and interviews, making their content accessible to a broader audience.
3. In surveillance and security applications, the system could be used for real-time speech recognition, aiding in monitoring conversations for potential security threats or incidents.
4. The technology could enhance the accuracy and efficiency of voice assistants, robots, and smart devices by understanding spoken language through lip movements, improving human-computer interaction.

3.2 Non-Functional Requirements

1. Accuracy and Precision:The system should maintain a high level of accuracy and precision in recognizing spoken language from lip movements, minimizing transcription errors.
2. Scalability:The system should be able to scale to accommodate increased processing demands as the user base grows, without a significant loss in performance.
3. Security and Privacy: Strong security measures must be in place to protect the confidentiality and privacy of audio and visual data, especially in applications involving sensitive or private conversations.
4. Usability and Accessibility: The user interface should be intuitive and accessible to a broad range of users, including those with disabilities, to ensure a positive user experience.
5. The system should provide quick responses, with minimal latency in processing and transcribing spoken language from lip movements.

Here are some additional non-functional requirements that may be important for some projects:

1. Robustness and Adaptability:The system should be robust enough to function in various environmental conditions, such as different lighting and background noise. It should also adapt to different accents and languages.
2. Adaptive User Feedback: The system should include mechanisms for collecting and utilizing user feedback to continually improve recognition accuracy and user satisfaction.
3. Availability and Reliability: The system should be available and reliable for use 24/7, ensuring that it can be depended upon for continuous speech recognition.

3.3 Constraints

1. Data Availability and Labeling: The system requires access to substantial datasets of labeled video and audio data for effective training, which can be challenging and costly to collect and annotate.
2. Bias in Data and Recognition: Similar to text-based models, Visual Speech Recognition Using AI systems can inherit biases from their training data, leading to variations in recognition accuracy based on factors like speaker gender, ethnicity, or accent.
3. Privacy Concerns: The collection and use of video and audio data raise privacy concerns, requiring compliance with regulations and ethical considerations, especially in applications involving user data.
4. Variable Environmental Conditions: The system's accuracy may be influenced by environmental factors, such as varying lighting and background noise, which can pose challenges in certain scenarios.

Here are some additional constraints that may be relevant for some projects:

1. +Computational Resources: The project demands significant computational resources for both model training and real-time processing, potentially limiting its implementation in resource-constrained environments.
2. Multilingual Support: Recognizing spoken language in multiple languages, dialects, and accents can be complex, necessitating extensive multilingual training data and specialized model architectures.
3. Real-Time Processing Challenges: Achieving real-time processing, especially in live conversations or interviews, can be computationally intensive, affecting the system's responsiveness and performance.

3.4 Hardware & Software Requirements

Hardware Requirements:

CPU/GPU Resources: For training deep learning models, GPUs (Graphics Processing Units) are highly recommended due to their parallel processing capabilities, which significantly accelerate model training. The choice of GPU depends on the model's size and complexity. Common options include:

Software Requirements: The software stack for a Visual Speech Recognition Using AI model includes various components for development, training, and deployment. Here are the key software requirements:

1. Python: Python is the primary programming language for developing machine learning and deep learning models. We recommend using Python 3.x for compatibility with modern libraries and frameworks.

2. Deep Learning Frameworks

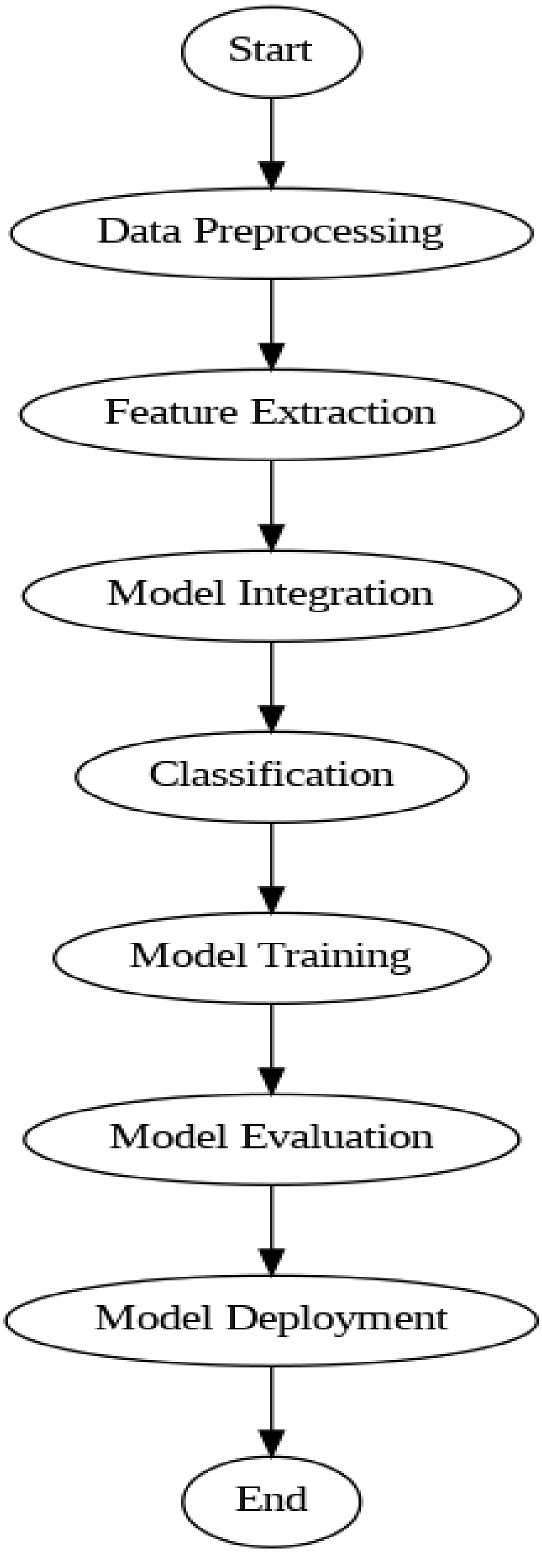
- TensorFlow: An open-source machine learning framework developed by Google.
- PyTorch: A popular open-source deep learning framework with a strong research community.
- Keras: High-level API that runs on top of TensorFlow and Theano.

3. Data Preprocessing Libraries

- NumPy: For numerical operations and array handling.
- NLTK: For natural language processing tasks.
- OpenCV: For image processing and performing computer vision tasks
- ImageIO: For reading and writing a wide range of image data, including animated images, volumetric data, etc.

4. GPU Drivers and CUDA (if using GPUs): Ensure that you have the necessary GPU drivers and CUDA (Compute Unified Device Architecture) libraries installed to leverage GPU acceleration.

3.5 System Block Diagram



Chapter 4

Proposed Design

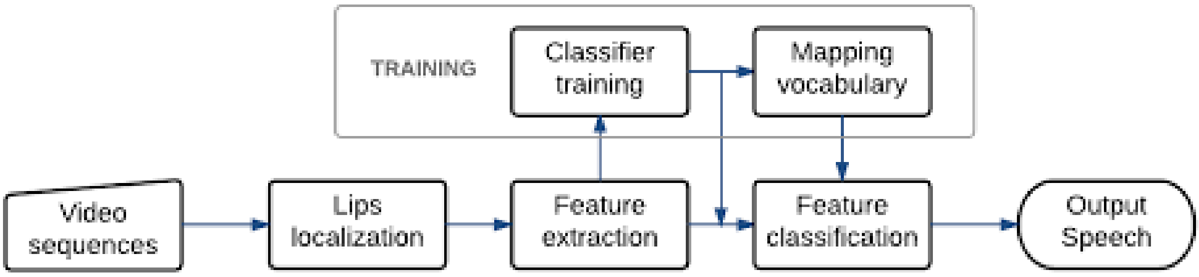
4.1 System Design

4.2 Detailed Design

4.3 Plan of Work

4. Proposed Design

4.1 System Design/Conceptual Design



In this system design report, we propose the architecture for a Visual Speech Recognition Using AI that leverages a GRID dataset containing 34 speakers(18 males, 16 females), each speaking 1000 english sentences. The proposed model combines Convolutional Neural Networks (CNN) and Bidirectional Long-Short-Term-Memory model which excels at capturing long-term dependencies, making it ideal for sequence prediction tasks. TimeDistributed class is a wrapper that allows you to apply a layer to every temporal slice of an input.

System Architecture:

The Visual Speech Recognition Using AI is designed to perform the following tasks:

1. Video Preprocessing: Extract frames from each video, detect and crop face from it, reduce the colored image to gray-scale image.

2. Alignment Extraction: Extract text from alignment files of grid dataset, characterize words and vectorize the sequence of characters into numbers.
   ST-CNN: Spatiotemporal Convolutional Neural Networks (STCNNs) are deep learning models used for analyzing and classifying spatiotemporal data, such as video data

Bi-LSTM: Bidirectional recurrent neural networks(RNN) are really just putting two independent RNNs together. This structure allows the networks to have both backward and forward information about the sequence at every time step.

3. Model Integration:
   Combine the feature vectors generated by the ST-CNN and Bi-LSTM models.
   Use a multi-modal approach to fuse the extracted features, taking into account the strengths of both models.

4. Classification: This problem is modeled as a classification problem. For each video we have 75 frames, which are to be classified into one of the 41 classes. These classes consist of English alphabets and symbols. This is done by taking into consideration the previous and next characters using Bidirectional LSTMs which are connected to a fully connected deep layer.

5. Model Training:
   Train the integrated model on the labeled dataset, employing loss functions and optimizers suitable for multi-class classification tasks. For the training purpose, we will be using a specific type of loss function called CTC loss. This takes care of sequential input data and allows the luxury to skip the step of aligning frames with respective characters spoken by the person in the sentence.
   Monitor the model's performance on a validation dataset to prevent overfitting.

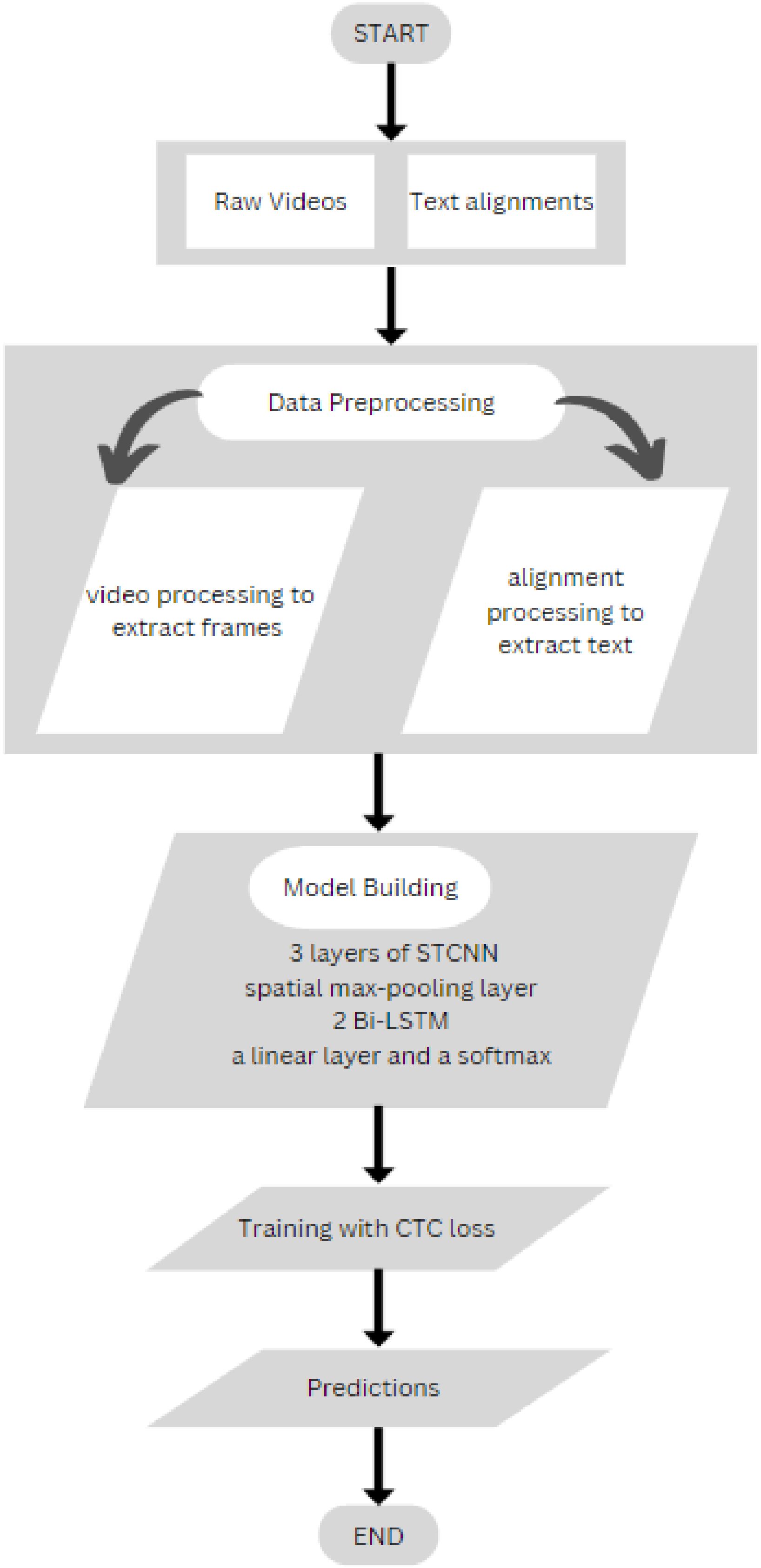
6. Model Evaluation:
   Evaluate the model's performance on a test dataset using metrics such as accuracy, precision, WER (Word Error Rate) scores.
   Word Error Rate = (Substitutions + Insertions + Deletions) / Number of Words Spoken
   Assess the model's ability to classify the characters accurately.

Data Source: The GRID corpus dataset contains 34 speakers(18 males, 16 females), each speaking 1000 english sentences.. A data pipeline should be developed to clean and preprocess to ensure consistent formatting and labeling.

Technologies and Frameworks: The Visual Speech Recognition Model will be implemented using the following technologies and frameworks:

- Python: As the primary programming language for model development.
- TensorFlow/Keras: For building and training the CNN model.
- scikit-learn: For model evaluation and metrics calculation.
- opencv-python: Frame capturing
- matplotlib: Visualization
- imageio: Image processing and analysis

4.2 Detailed Design



4.3 Plan of Work

A thorough review of relevant research papers and articles was conducted to gain insights into the state-of-the-art in visual speech recognition using AI. This literature review covered a wide range of topics, including:

- Existing ML and DL models used for cyberbullying detection.
- Research papers that combine different ML and DL models for improved detection accuracy.
- WER metric for evaluating the classified text(41 classes).

Selection of Research Papers: From the extensive literature review, 5-7 research papers were carefully selected based on their relevance to the research objectives. These papers presented innovative approaches and implementations of ML and DL models in the context of cyberbullying detection and classification.

Comparative Analysis: Each selected research paper was thoroughly analyzed with a focus on the following aspects:

- The ML and DL models employed, their architecture, and training techniques.
- The datasets used for training and evaluation.
- Performance metrics and evaluation results.
- Any unique features, advantages, or challenges posed by the models.

Next Steps: The next phase of this project will involve designing and implementing the chosen models for visual speech recognition, creating a severity rating model if applicable, and exploring practical applications in the field of visual speech recognition for AI. Regular progress reports and updates will be provided as the project advances.

Chapter 5

Implementation

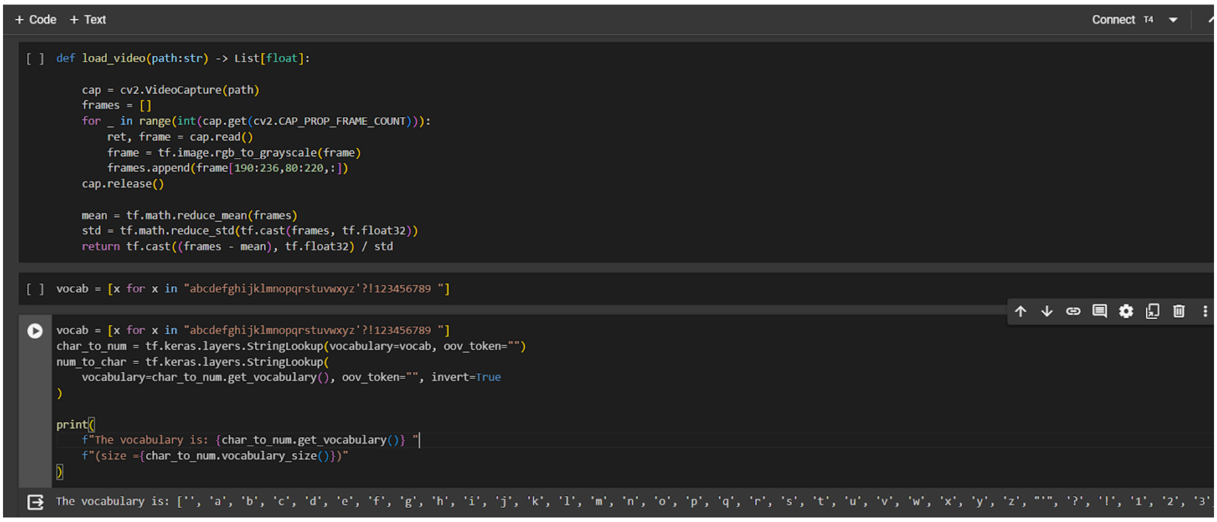
5. Implementation

1. Dataset Preprocessing

Dataset is preprocessed by opening the video file and initializes an empty list to store processed frames. It then iterates through each frame of the video, converting them to grayscale and cropping them to extract a mouth-centred crop of size 100 × 50 pixels per frame. These processed frames are then added to the list.

Once all frames are processed, the function calculates the mean and standard deviation of the frames. It then normalizes each frame by subtracting the mean and dividing by the standard deviation. Finally, it casts the frames to a specific data type (tf.float32) and returns the normalized frames.

List called vocab containing all the characters that may appear in the strings to be processed. This includes lowercase letters, punctuation marks like apostrophe, exclamation mark, and question mark, digits from 0 to 9, and a space.



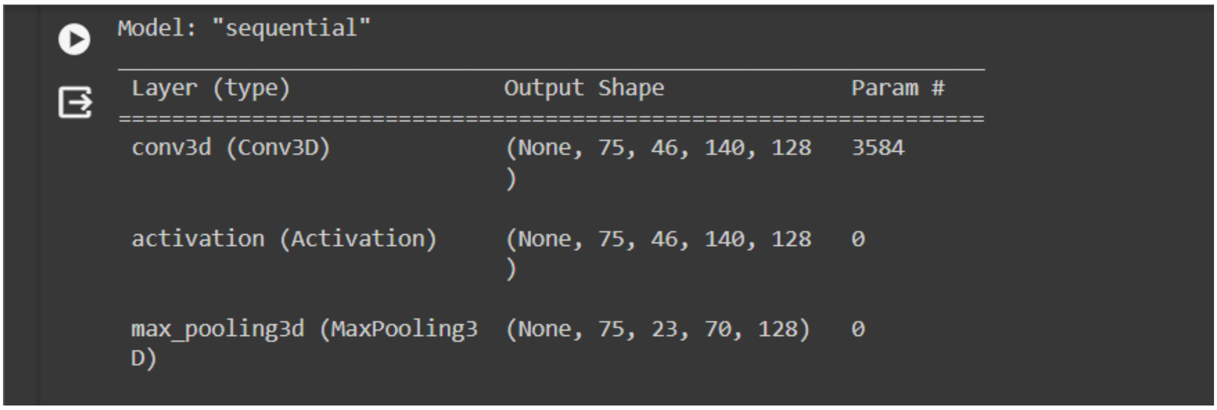
```
def load_video(path:str) -> List[float]:

    cap = cv2.VideoCapture(path)
    frames = []
    for _ in range(int(cap.get(cv2.CAP_PROP_FRAME_COUNT))):
        ret, frame = cap.read()
        frame = tf.image.rgb_to_grayscale(frame)
        frames.append(frame[190:236,80:220,:])
    cap.release()

    mean = tf.math.reduce_mean(frames)
    std = tf.math.reduce_std(tf.cast(frames, tf.float32))
    return tf.cast((frames - mean), tf.float32) / std
```

```
vocab = [x for x in "abcdefghijklmnopqrstuvwxyz'?!123456789 "]
```

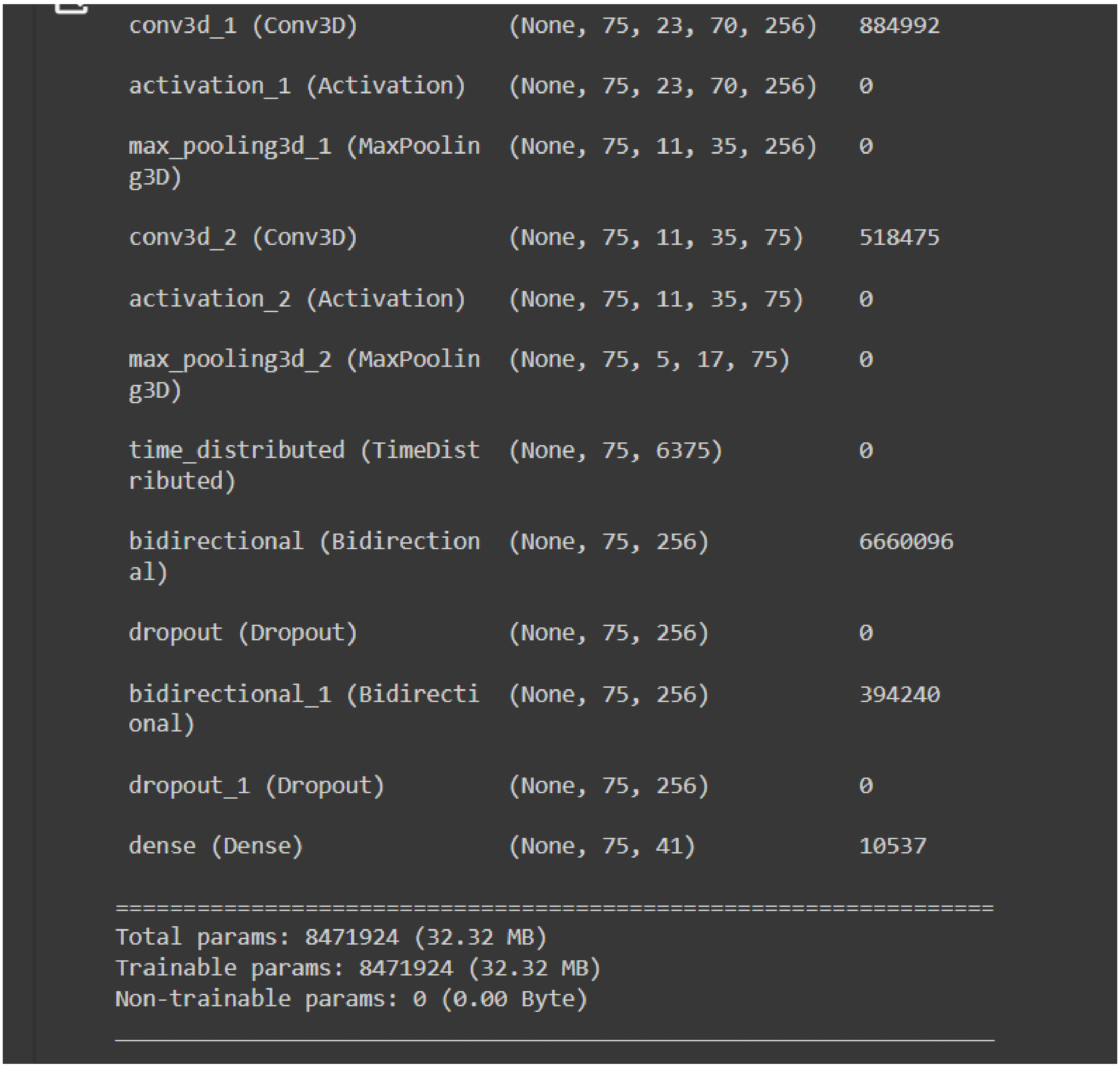
```
vocab = [x for x in "abcdefghijklmnopqrstuvwxyz'?!123456789 "]
char_to_num = tf.keras.layers.StringLookup(vocabulary=vocab, oov_token="")
num_to_char = tf.keras.layers.StringLookup(
    vocabulary=char_to_num.get_vocabulary(), oov_token="", invert=True
)

print(
    f"The vocabulary is: {char_to_num.get_vocabulary()} "
    f"(size ={char_to_num.vocabulary_size()})"
)
```

```
The vocabulary is: ['', 'a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j', 'k', 'l', 'm', 'n', 'o', 'p', 'q', 'r', 's', 't', 'u', 'v', 'w', 'x', 'y', 'z', "'", '?', '!', '1', '2', '3'
```

2. LipNet Architecture



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv3d (Conv3D)             (None, 75, 46, 140, 128   3584
                             )

 activation (Activation)     (None, 75, 46, 140, 128   0
                             )

 max_pooling3d (MaxPooling3  (None, 75, 23, 70, 128)   0
 D)
```



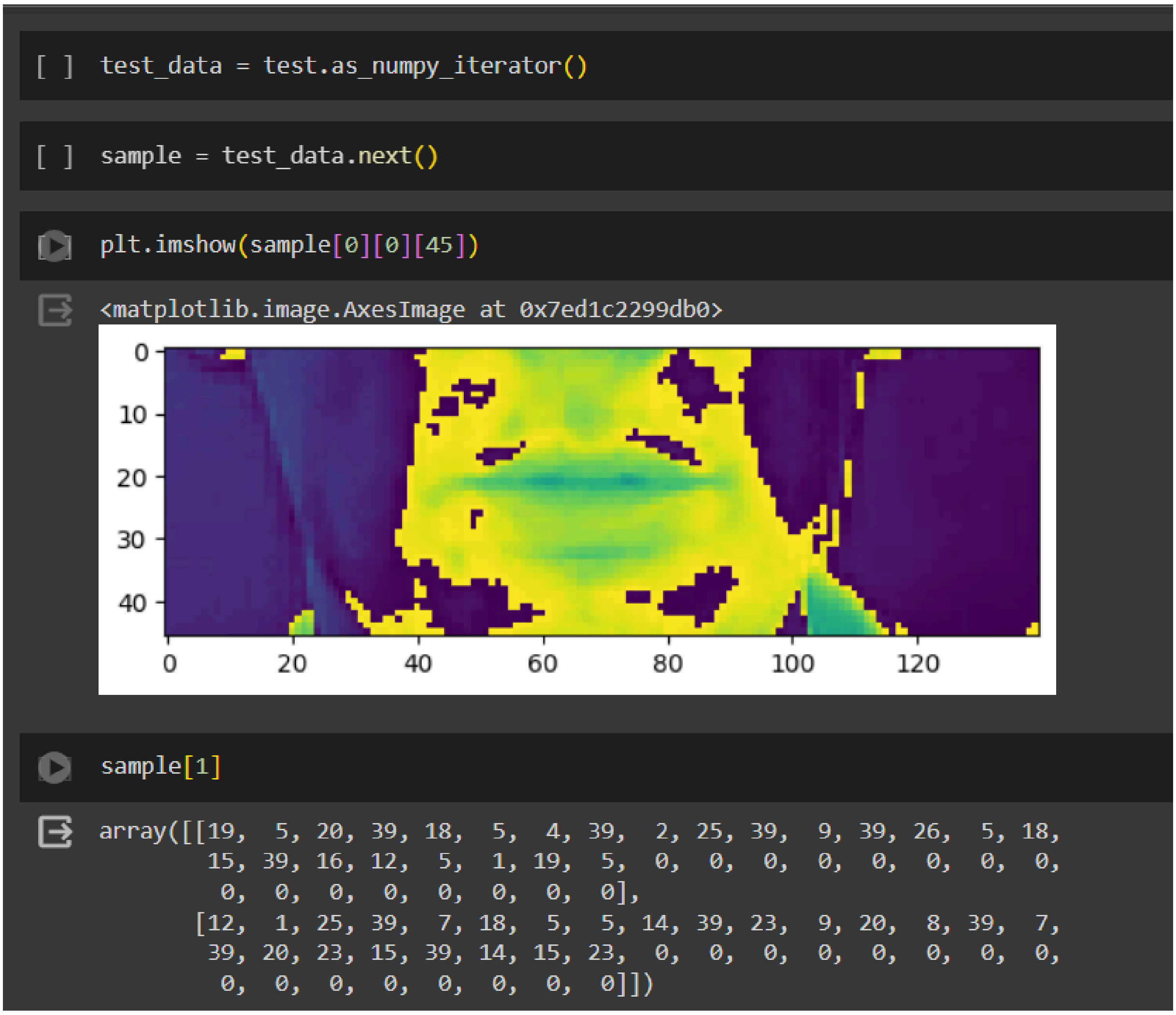
```
 conv3d_1 (Conv3D)           (None, 75, 23, 70, 256)   884992

 activation_1 (Activation)   (None, 75, 23, 70, 256)   0

 max_pooling3d_1 (MaxPoolin  (None, 75, 11, 35, 256)   0
 g3D)

 conv3d_2 (Conv3D)           (None, 75, 11, 35, 75)    518475

 activation_2 (Activation)   (None, 75, 11, 35, 75)    0

 max_pooling3d_2 (MaxPoolin  (None, 75, 5, 17, 75)     0
 g3D)

 time_distributed (TimeDist  (None, 75, 6375)          0
 ributed)

 bidirectional (Bidirection  (None, 75, 256)           6660096
 al)

 dropout (Dropout)           (None, 75, 256)           0

 bidirectional_1 (Bidirecti  (None, 75, 256)           394240
 onal)

 dropout_1 (Dropout)         (None, 75, 256)           0

 dense (Dense)               (None, 75, 41)            10537

=================================================================
Total params: 8471924 (32.32 MB)
Trainable params: 8471924 (32.32 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

The LipNet architecture, which starts with 3×(convolutions, max-pooling) and 2× . (channel-wise dropout). Subsequently, the features extracted are followed by a Bi-LSTM. The Bi-LSTM is crucial for efficient further aggregation of the CNN output. Finally, a linear transformation is applied at each time-step, followed by a softmax over the vocabulary augmented with the CTC blank, and then the CTC loss. All layers use rectified linear unit (ReLU) activation functions.

Chapter 6

Testing

6. Testing



```
[ ] test_data = test.as_numpy_iterator()
```

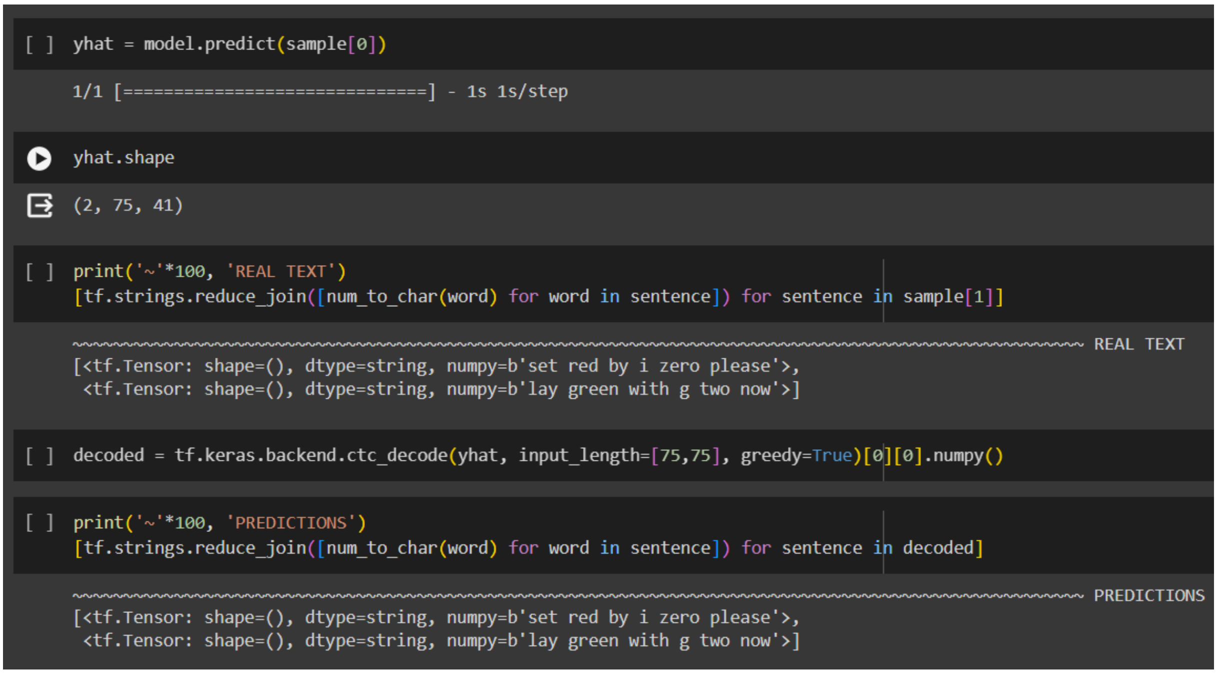
```
[ ] sample = test_data.next()
```

```
plt.imshow(sample[0][0][45])
```

```
<matplotlib.image.AxesImage at 0x7ed1c2299db0>
```

```
sample[1]
```

```
array([[19,  5, 20, 39, 18,  5,  4, 39,  2, 25, 39,  9, 39, 26,  5, 18,
        15, 39, 16, 12,  5,  1, 19,  5,  0,  0,  0,  0,  0,  0,  0,  0,
         0,  0,  0,  0,  0,  0,  0,  0],
       [12,  1, 25, 39,  7, 18,  5,  5, 14, 39, 23,  9, 20,  8, 39,  7,
        39, 20, 23, 15, 39, 14, 15, 23,  0,  0,  0,  0,  0,  0,  0,  0,
         0,  0,  0,  0,  0,  0,  0,  0]])
```



```
yhat = model.predict(sample[0])
```

```
1/1 [==============================] - 1s 1s/step
```

```
yhat.shape
```

```
(2, 75, 41)
```

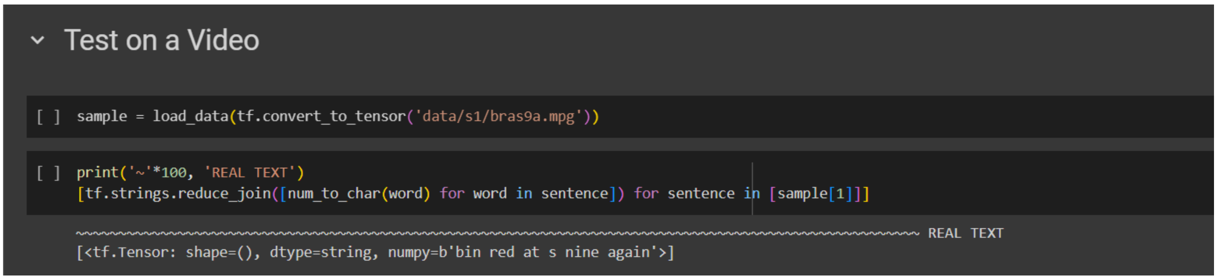
```
print('~'*100, 'REAL TEXT')
[tf.strings.reduce_join([num_to_char(word) for word in sentence]) for sentence in sample[1]]
```

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ REAL TEXT
[<tf.Tensor: shape=(), dtype=string, numpy=b'set red by i zero please'>,
 <tf.Tensor: shape=(), dtype=string, numpy=b'lay green with g two now'>]
```

```
decoded = tf.keras.backend.ctc_decode(yhat, input_length=[75,75], greedy=True)[0][0].numpy()
```

```
print('~'*100, 'PREDICTIONS')
[tf.strings.reduce_join([num_to_char(word) for word in sentence]) for sentence in decoded]
```

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ PREDICTIONS
[<tf.Tensor: shape=(), dtype=string, numpy=b'set red by i zero please'>,
 <tf.Tensor: shape=(), dtype=string, numpy=b'lay green with g two now'>]
```

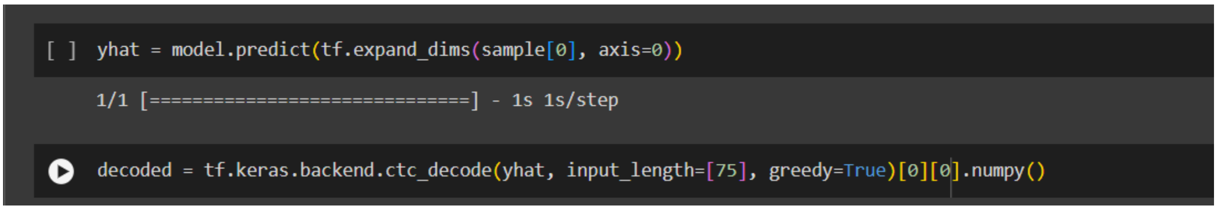


Test on a Video

```
sample = load_data(tf.convert_to_tensor('data/s1/bras9a.mpg'))
```

```
print('~'*100, 'REAL TEXT')
[tf.strings.reduce_join([num_to_char(word) for word in sentence]) for sentence in [sample[1]]]
```

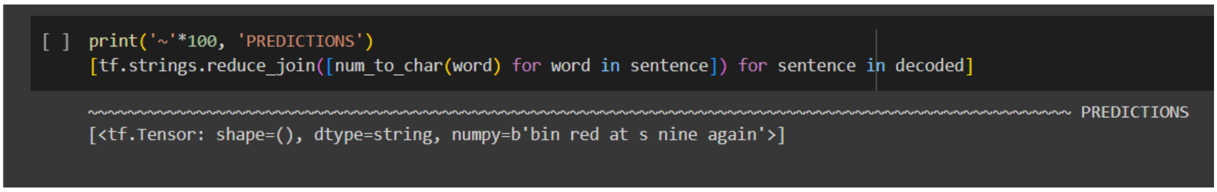
```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ REAL TEXT
[<tf.Tensor: shape=(), dtype=string, numpy=b'bin red at s nine again'>]
```



```
yhat = model.predict(tf.expand_dims(sample[0], axis=0))
```

```
1/1 [==============================] - 1s 1s/step
```

```
decoded = tf.keras.backend.ctc_decode(yhat, input_length=[75], greedy=True)[0][0].numpy()
```



```
[ ]  print('~'*100, 'PREDICTIONS')
     [tf.strings.reduce_join([num_to_char(word) for word in sentence]) for sentence in decoded]

     ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ PREDICTIONS
     [<tf.Tensor: shape=(), dtype=string, numpy=b'bin red at s nine again'>]
```

Chapter 7

Result Analysis

7.1 System Simulation

7.2 Parameters/Graphs

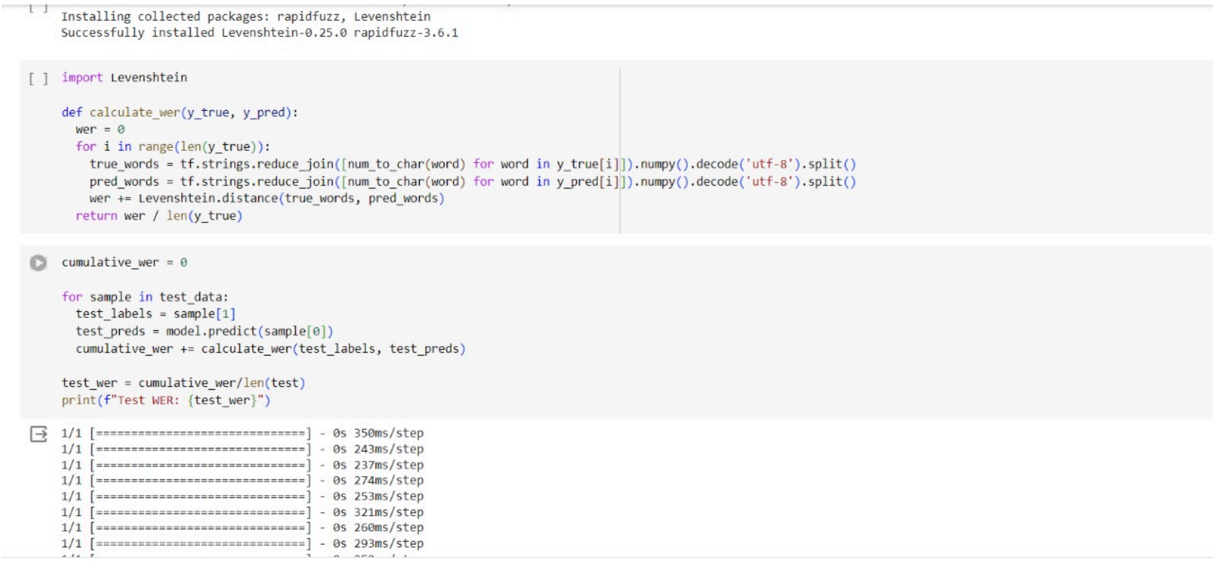
7.3 Output Printouts

7.4 Observation and Analysis

7. Result Analysis

7.3 Output Screenshots

To measure the performance of the LipNet Architecture we compute the metric WER(Word Error Rate), a standard metric for the performance of VSR models. WER (or CER) is defined as the minimum number of word (or character) insertions, substitutions, and deletions required to transform the prediction into the ground truth, divided by the number of words (or characters) in the ground truth. WER is usually equal to classification error when the predicted sentence has the same number of words as the ground truth, particularly in our case since almost all errors are substitution errors.

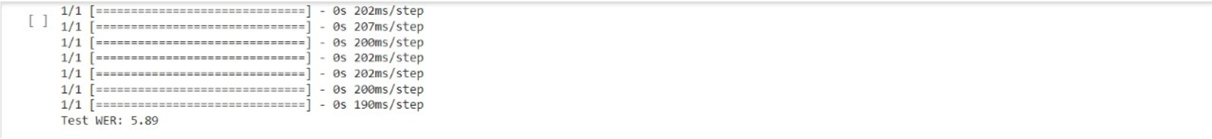


```
Installing collected packages: rapidfuzz, Levenshtein
Successfully installed Levenshtein-0.25.0 rapidfuzz-3.6.1
```

```
import Levenshtein

def calculate_wer(y_true, y_pred):
  wer = 0
  for i in range(len(y_true)):
    true_words = tf.strings.reduce_join([num_to_char(word) for word in y_true[i]]).numpy().decode('utf-8').split()
    pred_words = tf.strings.reduce_join([num_to_char(word) for word in y_pred[i]]).numpy().decode('utf-8').split()
    wer += Levenshtein.distance(true_words, pred_words)
  return wer / len(y_true)
```

```
cumulative_wer = 0

for sample in test_data:
  test_labels = sample[1]
  test_preds = model.predict(sample[0])
  cumulative_wer += calculate_wer(test_labels, test_preds)

test_wer = cumulative_wer/len(test)
print(f"Test WER: {test_wer}")
```

```
1/1 [==============================] - 0s 350ms/step
1/1 [==============================] - 0s 243ms/step
1/1 [==============================] - 0s 237ms/step
1/1 [==============================] - 0s 274ms/step
1/1 [==============================] - 0s 253ms/step
1/1 [==============================] - 0s 321ms/step
1/1 [==============================] - 0s 260ms/step
1/1 [==============================] - 0s 293ms/step
```



```
[ ] 1/1 [==============================] - 0s 202ms/step
    1/1 [==============================] - 0s 207ms/step
    1/1 [==============================] - 0s 200ms/step
    1/1 [==============================] - 0s 202ms/step
    1/1 [==============================] - 0s 202ms/step
    1/1 [==============================] - 0s 200ms/step
    1/1 [==============================] - 0s 190ms/step
    Test WER: 5.89
```

For computing the WER, we calculate the Levenshtein distance between the true words and the predicted words for a single sample and then calculate the cumulative WER for all the samples in the test dataset, which is the cumulative WER upon the length of the WER.

The lesser the WER value, greater the accuracy of the model fitted. The WER for the model is 5.89.

7.4 Observation and Analysis

According to the literature, the accuracy of human lip readers is around 20% (Easton & Basala, 1982; Hilder et al., 2009). As expected, the fixed sentence structure and the limited subset of words for each position in the GRID corpus facilitate the use of context, increasing performance. With Bi-LSTMs with CTC loss function, our model performs really good with 5.89 wer. Which indicates that Bi-GRU are better for this task as it gives a score of 4.6 wer.

Chapter 8

Conclusion

8. Conclusion

8.1 Limitation

Contextual Understanding: Visual speech recognition systems face challenges in comprehending language beyond lip movements due to ambiguity, idiomatic expressions, and the absence of non-verbal cues. Incorporating contextual language models and multimodal fusion techniques can aid in interpreting spoken language within its broader context by leveraging surrounding words, sentences, and complementary sources of information such as gestures and facial expressions.

Limited Vocabulary: Visual speech recognition systems excel in constrained domains with a well-defined vocabulary but struggle with broader vocabularies, leading to errors with out-of-vocabulary words and lexical ambiguity. Strategies such as data augmentation and transfer learning can mitigate these limitations by increasing dataset diversity and leveraging pre-trained models to improve generalization to a wider range of vocabulary.

Data Availability: Obtaining labeled data for visual speech recognition poses challenges in terms of annotation costs, domain specificity, and speaker diversity. Crowdsourcing platforms, transfer learning, and synthetic data generation techniques offer potential solutions by reducing annotation efforts, leveraging existing datasets, and augmenting data diversity through simulation or computer graphics. These strategies aim to address data scarcity issues and enhance the robustness of visual speech recognition models across various domains and speaker demographics.

8.3 Future Scope

Multilingual and Cross-Cultural Adaptation: Future developments will likely prioritize the creation of diverse datasets encompassing multiple languages, accents, and cultural contexts. This initiative aims to enhance the robustness and generalization capabilities of visual speech recognition models across linguistic and cultural diversity. Additionally, advancements in cross-lingual transfer learning techniques will enable models trained on one language to adapt and perform effectively on other languages, thereby fostering inclusivity and accessibility on a global scale.

Dataset Expansion and Diversity: The expansion and diversification of visual speech datasets will be crucial for improving the performance and scalability of AI models. Efforts may focus on collecting data from underrepresented populations, including individuals with diverse speech patterns, accents, and facial features. Moreover, the creation of specialized datasets tailored to specific domains, such as healthcare, education, and entertainment, will enable the development of domain-specific visual speech recognition systems optimized for diverse real-world applications.

Fine-Grained Annotation and Semantic Understanding: Future research endeavors will likely emphasize fine-grained annotation techniques to capture nuanced semantic information and contextual cues from visual speech data. By incorporating advanced annotation methodologies, such as semantic segmentation and facial action coding, AI models can achieve deeper semantic understanding of spoken language, enabling more accurate transcription and interpretation of speech content. This approach will pave the way for the development of next-generation visual speech recognition systems capable of understanding not only what is said but also how it is expressed, thereby facilitating richer and more contextually aware communication experiences across languages and cultures.

References

Journal Papers

8. References

Journal Paper

[1] Nikita Deshmukh,Anamika Ahire,Smriti H Bhandari,Apurva Mali. Vision based Lip Reading System using Deep Learning. 2021 International Conference on Computing, Communication and Green Engineering (CCGE) JSPM's RSCOE, Pune, India. 2021.

[2] Joon Son Chung, Andrew Senior, Oriol Vinyals, Andrew Zisserman. Lip Reading Sentences in the Wild. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6447-6456, 2017.

[3] Pingchuan Ma, Stavros Petridis, Maja Pantic. Visual Speech Recognition Using AI for Multiple Languages in the Wild. arXiv:2202.13084, 2022

[4] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, C.V. Jawahar; Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13796-13805, 2020

[5] Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, Nando de Freitas. LipNet: End-to-End Sentence-level Lipreading. arXiv:1611.01599v2.

[6] Themos Stafylakis, Georgios Tzimiropoulos. Combining Residual Networks with LSTMs for Lipreading. arXiv:1703.04105.

[7] Themos Stafylakisa, Muhammad Haris Khana, Georgios Tzimiropoulosa. Pushing the boundaries of audiovisual word recognition using Residual Networks and LSTMs. Elsevier, Science-Direct, Computer Vision and Image Understanding, Volumes 176–177, November–December 2018, Pages 22-32.

[8] Konstantinos Vougioukas, Pingchuan Ma , Stavros Petridis and Maja Pantic. Video-Driven Speech Reconstruction using Generative Adversarial Networks. Inproceedings, Interspeech September 2019.

[9] Minsu Kim, Joanna Hong, Yong Man Ro. Lip to Speech Synthesis with Visual Context Attentional GAN. arXiv:2204.01726v1 [cs.CV] 4 Apr 2022.

[10] Gaoyan Zhang and Yuanyao Lu, Research on a Lip Reading Algorithm Based on Efficient-GhostNet, Journals, Electronics, Volume 12, Issue 5, 10.3390/electronics12051151.