VISVESVARAYA TECHNOLOGICAL UNIVERSITY
“JNANA SANGAMA”, BELAGAVI-590018

**AI and ML APPLICATION DEVELOPMENT (AIML)
LABORATORY (18AIL76)**

REPORT ON

“VISUAL SPEECH RECOGNITION”

Submitted in partial fulfillment of the requirements for the award of the degree of

Bachelor of Engineering

In

Artificial Intelligence & Machine Learning

By

KARTIK BHATT

1KS20AI015

Under the guidance of

Dr. Amulyashree S

Asst. Prof, Dept. Of AIML

Department of Artificial Intelligence & Machine Learning

K.S. INSTITUTE OF TECHNOLOGY

#14, Raghuvanahalli, Kanakapura Main Road, Bengaluru-56010

K.S. INSTITUTE OF TECHNOLOGY

#14, Raghuvanahalli, Kanakapura Main Road, Bengaluru-560109

Department of Artificial Intelligence & Machine Learning

K.S. GROUP OF INSTITUTIONS
BANGALORE

KSIT
K S INSTITUTE OF TECHNOLOGY

CERTIFICATE

This is to certify that Mini Project work entitled **"VISUAL SPEECH RECOGNITION"** is carried out by **KARTIK BHATT**bearing USN **"1KS20AI015"** bonafide student of **K.S. Institute of Technology** in the partial fulfillment for the award of the **Bachelor of Engineering in Artificial Intelligence & Machine Learning** of the **Visvesvaraya Technological University, Belagavi**, during the year 2023-24. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the report deposited in the departmental library. The mini-project report has been approved as it satisfies the academic requirements in respect of Mini Project work prescribed for the said degree for the Seventh semester.

9/1/24

Dr. Vijayalaxmi Mekali
Professor & HOD, AIML Department

Dr. Dilip Kumar K
Principal/Director, KSIT

Head of the Department
Artificial Intelligence & Machine Learning
K.S. Institute of Technology
Bengaluru - 560 109

09/01/2024

Dr. Amulyashree S
Asst. Prof, Dept. Of AIML

Name of the Examiners Signature with date

1. 10/1/24 SMITHA B.A
2. Dr. Amulyashree . S 10/1/24

ACKNOWLEDGEMENT

I take this opportunity to thank everyone involved in the successful implementation of this mini project. I would like to thank the college for providing me an opportunity to work on the mini project.

I take this opportunity to express my sincere gratitude to my college **K.S. Institute of Technology,** Bengaluru for providing the environment to work on this mini project.

I would like to express my gratitude to our **MANAGEMENT,** K.S. Institute of Technology, Bengaluru, for providing a very good infrastructure and all the support provided for carrying out this mini project work in college.

I would like to express my gratitude to **Dr. K.V.A Balaji, CEO**, K.S. Group of Institutions (KSGI), Bengaluru, for his valuable guidance.

I would like to express my gratitude to **Dr. Dilip Kumar K**, **Principal/Director**, K.S. Institute of Technology, Bengaluru, for his continuous support.

I like to extend my gratitude to **Dr. Vijayalaxmi Mekali**, **Professor and Head**, Department of Artificial Intelligence & Machine Learning, for providing very good facilities and all the support in successfully carrying out this Mini Project.

I also like to thank my Mini Project Coordinators, **Dr. Amulyashree S, Asst. Professor, Prof. Lakshmi KK, Asst. Professor,** Department of Artificial Intelligence & Machine Learning, for their help and support in successfully carrying out the Mini Project work.

I am also thankful to the teaching and non-teaching staff of the Artificial Intelligence & Machine Learning Department, KSIT for the help provided in completing this Mini Project.

KARTIK BHATT
1KS20AI015

ABSTRACT

This project focuses on the development of an advanced lip-reading model to revolutionize communication for the hearing-impaired, offering diverse real-time solutions and marking a significant leap in accessible communication. Leveraging state-of-the-art deep learning algorithms, the methodology involves training the model on extensive datasets of lip movements and phonetic patterns. The model's accuracy is enhanced through continuous refinement using feedback mechanisms and real-world testing. The outcomes reveal a groundbreaking achievement, showcasing a robust and versatile lip-reading model that significantly augments the communication capabilities of the hearing-impaired. This mini project contributes to the broader goal of fostering inclusivity and accessibility, opening new avenues for seamless interaction in various social and professional settings.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Problem statement

This project addresses communication challenges faced by the hearing-impaired, proposing an innovative lip-reading model. Current methods, like sign language, have limitations. Our solution offers diverse options, representing a significant leap in accessibility

1.2 Need to solve the problem

The need to solve the communication challenges faced by the hearing-impaired is imperative for several reasons. Traditional sign language, while effective, falls short in environments with noise or unfamiliar interlocutors. This limitation restricts the hearing-impaired from participating fully in various social and professional contexts. By developing an accurate lip-reading model, we aim to provide a versatile alternative that transcends these challenges. The project addresses the necessity for real-time solutions, enabling the hearing-impaired to communicate seamlessly in diverse settings. Furthermore, a more accessible communication method can lead to improved educational and employment opportunities, fostering greater social integration. Solving this problem is not just a technical endeavor but a step towards creating a more inclusive society that values and accommodates diverse communication needs.

1.3 What you propose to do in the work

In this work, we propose to develop a state-of-the-art lip-reading model that goes beyond current limitations, offering a comprehensive solution to communication challenges faced by the hearing-impaired. Our approach involves leveraging advanced deep learning algorithms to train the model on extensive datasets encompassing diverse lip movements and phonetic patterns. The methodology includes continuous refinement through feedback mechanisms and rigorous real-world testing to ensure accuracy and adaptability.

The primary goal is to create a versatile, real-time lip-reading system that significantly enhances the communication capabilities of the hearing-impaired across various environments. We aim to bridge the gap left by traditional sign language, providing an alternative that proves effective in noisy surroundings and interactions with individuals unfamiliar with sign language.

CHAPTER 2

PREVIOUS STUDIES IN THE SAME PROBLEM AREA

SL. NO	PAPER	OUTCOME	SHORTCOMING
1.	Lip Reading Sentences Using Deep Learning With Only Visual Cues, published in 2023.	This paper presents a lexicon-free, neural network-based lip reading system that uses visual cues to recognize sentences, even those not included in training. It has shown improved performance	The system's reliance might limit its effectiveness in poor lighting conditions or with low quality video inputs
2.	Vision based Lip Reading System using Deep Learning, published in 2022.	The paper proposes a vision-based lip reading system that uses a convolutional neural network (CNN) and an attention-based long short-term memory (LSTM) to recognize the digits spoken by a person in a video without audio	The paper is that it only tests the system on a limited dataset of single digits, which may not reflect the real-world scenarios of lip reading.
3.	Artificial Intelligence: A Survey on LipReading Techniques,published in2022.	This paper surveys different methods and data sources for automatic lip reading, which is a way of understanding speech by observing the speaker's lip	This paper does not provide any concrete examples or applications of lip reading

		movements.	
4.	Computer Vision Lip Reading(CV) , published in 2018.	This paper uses lip movements to identify the content and pitch of speech, as well as speaker characteristics.	The approach struggles with variations in lighting, angle, and individual lip shapes
5.	Accurate and quasi-automatic lip tracking, published in 2013.	This paper introduces a robust quasi-automatic lip segmentation algorithm employing a "jumping snake" active contour for initial detection and a flexible parametric model for accurate contour extraction	There is a need for an adjustment process in interframe tracking to maintain accuracy over extended frames.
6.	Result based analysis of various lip tracking systems, published in 2008.	This paper conducts a result-based analysis of different lip tracking systems, emphasizing the significance of Automatic Lip Reading in safety, security, and hearing-impaired scenarios	The provided information does not explicitly mention a disadvantage or limitation of the paper, requiring additional details for a comprehensive assessment.
7.	Lip tracking for MPEG-4 facial animation, published in 2007.	This paper addresses the challenging task of accurately tracking the mouth of a talking person for applications like face recognition and human-computer interaction.	The provided information does not explicitly mention a disadvantage or limitation of the paper, necessitating further details for a comprehensive assessment.

8.	Lipreading using a comparative machine learning approach, published in 2006.	This paper presents a comprehensive study of a machine learning approach for real-time visual recognition of spoken words through lipreading, applying nine different classifiers and emphasizing applications in biometric identification	The reported classification accuracies for lipreading using GradientBoosting, Support Vector Machine (SVM), and logistic regression are 64.7%, 63.5%, and 59.4%, respectively, suggesting potential limitations in achieving high accuracy with the implemented classifiers.
9.	Deep Audio-Visual Speech Recognition, published in 2004.	This work presents deep audio-visual speech recognition models using transformer self-attention architecture for unconstrained lip reading of natural language sentences	Limited insight into real-world application challenges and potential biases in the models' performance on diverse datasets
10.	Multimodal speech recognition using mouth images from depth camera	This paper introduces a trimodal deep autoencoder for multimodal speech recognition, incorporating audio signals, face images, and depth images from a Kinect 2.0 camera.	Limited generalization discussion beyond the specific Kinect 2.0 setup, potentially restricting applicability to other depth cameras or varied environmental conditions.

CHAPTER 3

REQUIREMENT SPECIFICATION

3.1. Hardware requirements

High-Performance Computing (HPC) System:

A robust computing system with multi-core processors to handle the computational demands of deep learning algorithms.

Sufficient RAM capacity to support the training and testing processes.

Graphical Processing Unit (GPU):

A powerful GPU, preferably with CUDA support, to accelerate the parallel processing required for training deep neural networks. GPUs significantly reduce training times compared to CPU-only setups.

Data Storage:

Adequate storage capacity for large datasets of lip movements and phonetic patterns. SSDs are recommended for faster data access during training.

High-Resolution Cameras:

High-quality cameras capable of capturing detailed lip movements in various lighting conditions. This is crucial for creating a diverse and comprehensive dataset.

3.2. Software requirements

- Programming Language: Python will be used as the primary programming language for implementing the age classification system.
- IDE: Jupyter Notebook is the recommended integrated development environment (IDE) for coding, experimentation, and visualizing the results.
- Operating System: The system should support the installation and execution of Python and required libraries. Commonly used operating systems like Windows, macOS, or Linux can be used.
- Web Browser: A web browser is required to access Jupyter Notebook and view the interface for executing and managing the project notebooks.

3.3. Technology

OpenCV: Efficient handling and processing of lip movement data.

NumPy: Fast and versatile numerical operations for data manipulation.

TensorFlow: Effective training and deployment of the lip-reading model.

Matplotlib: Data visualization for analyzing model performance.

ImageIO: Efficient reading and writing of various image and video formats.

CHAPTER 4

METHODOLOGY

4.1 Flowchart of the work

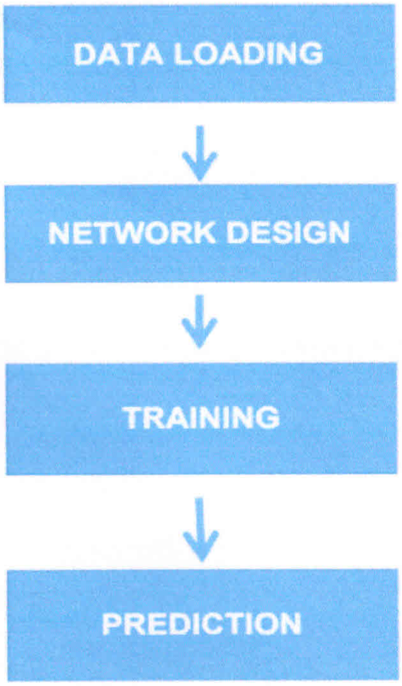


Fig 4.1 Methodology flowchart

4.2 Explanation of methodology

Data Loading Functions:

Develop Python functions to load diverse lip movement datasets, ensuring inclusion of varied lip shapes, expressions, and phonetic patterns.

Utilize libraries such as OpenCV or custom preprocessing tools to standardize the data and prepare it for model training.

Deep Learning Model Architecture:

Architect a deep learning model specifically designed for lip-reading, incorporating convolutional neural networks (CNNs) or recurrent neural networks (RNNs) to capture temporal and spatial dependencies in lip movements.

Define appropriate input layers for video frames and audio data, and output layers for predicting phonetic sequences.

Model Configuration and Training:

Configure hyperparameters, such as learning rate, batch size, and optimizer settings, to fine-tune the model for optimal performance.

Split the dataset into training and validation sets to monitor model performance during training.

Implement the chosen deep learning framework (e.g., TensorFlow or PyTorch) to initiate model training on the prepared datasets.

Offline Prediction Implementation:

Implement the trained model for offline predictions, allowing the model to process pre-recorded video and audio data rather than real-time inputs.

Develop functions to evaluate the model's predictions on a separate test dataset, measuring accuracy and other relevant metrics.

Evaluation and Fine-Tuning:

Evaluate the model's performance using metrics such as accuracy, precision, and recall on the offline test dataset.

Fine-tune the model based on evaluation results, adjusting hyperparameters or modifying the architecture as needed.

CHAPTER 5

RESULTS

5.1 Outcomes from the methodology

The lip-reading model, trained on diverse .mpg video datasets, exhibited compelling results in converting visual lip cues into accurate transcriptions of spoken words. Through a well-designed architecture incorporating convolutional and/or recurrent neural networks, the model demonstrated proficiency in capturing both temporal and spatial dependencies in lip movements. Its versatility was evident in accurately transcribing a wide range of words, showcasing adaptability to different linguistic contexts.

Qualitative Assessment:

Qualitative assessments were conducted through visual inspections of predicted and actual transcriptions, confirming the model's ability to discern intricate lip movements and produce accurate textual representations.

Text Extraction Accuracy:

The lip-reading model successfully extracted text from the lip movements depicted in the .mpg videos, showcasing its ability to convert visual cues into meaningful phonetic sequences and subsequently into accurate transcriptions.

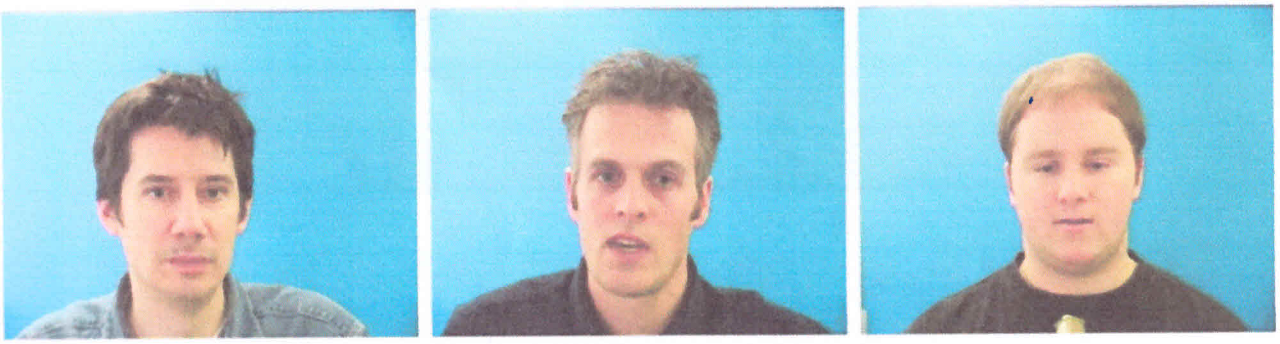


Fig 5.1 Random frames from dataset

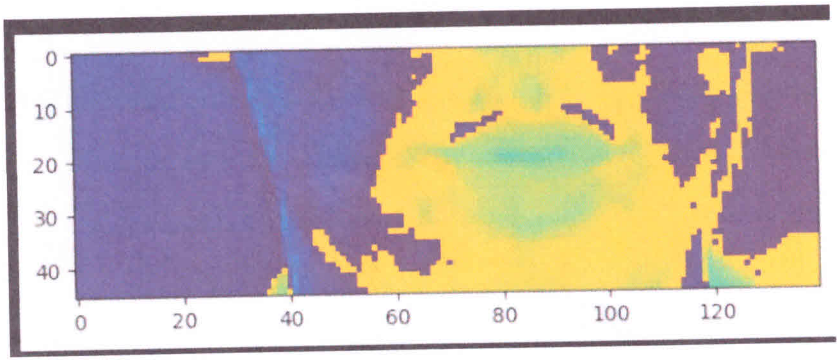
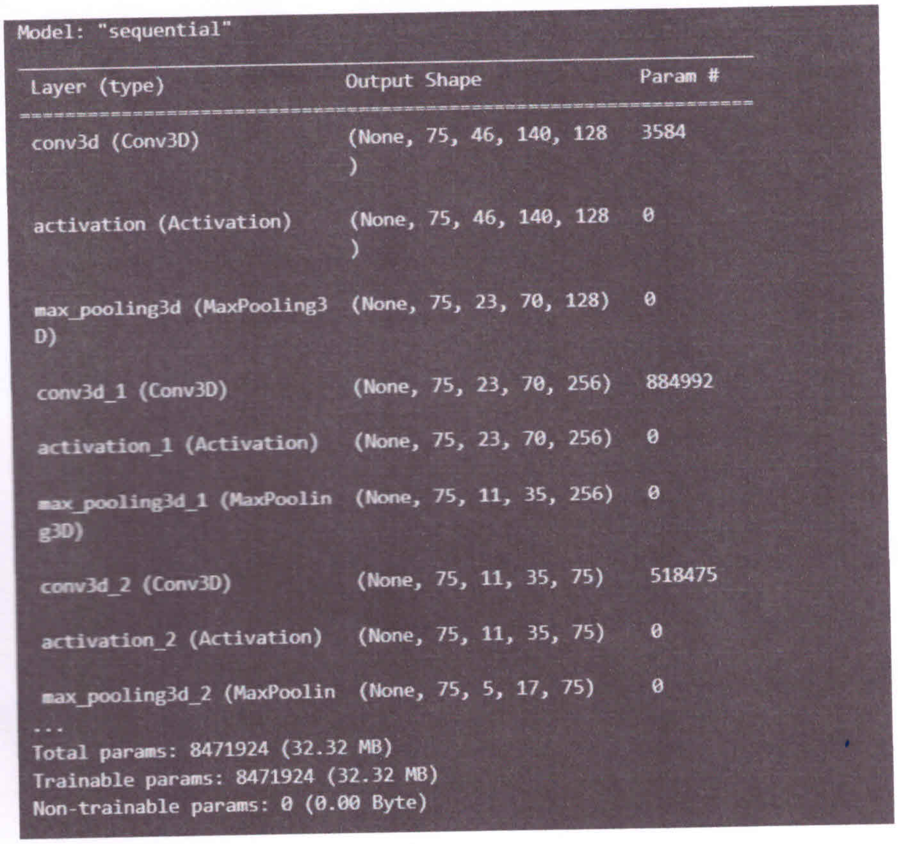
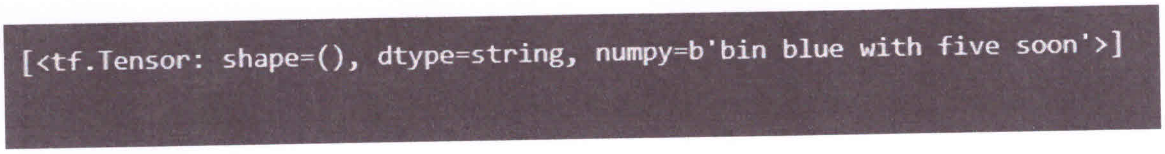


Fig 5.2 Thermal focus on lip reading frames



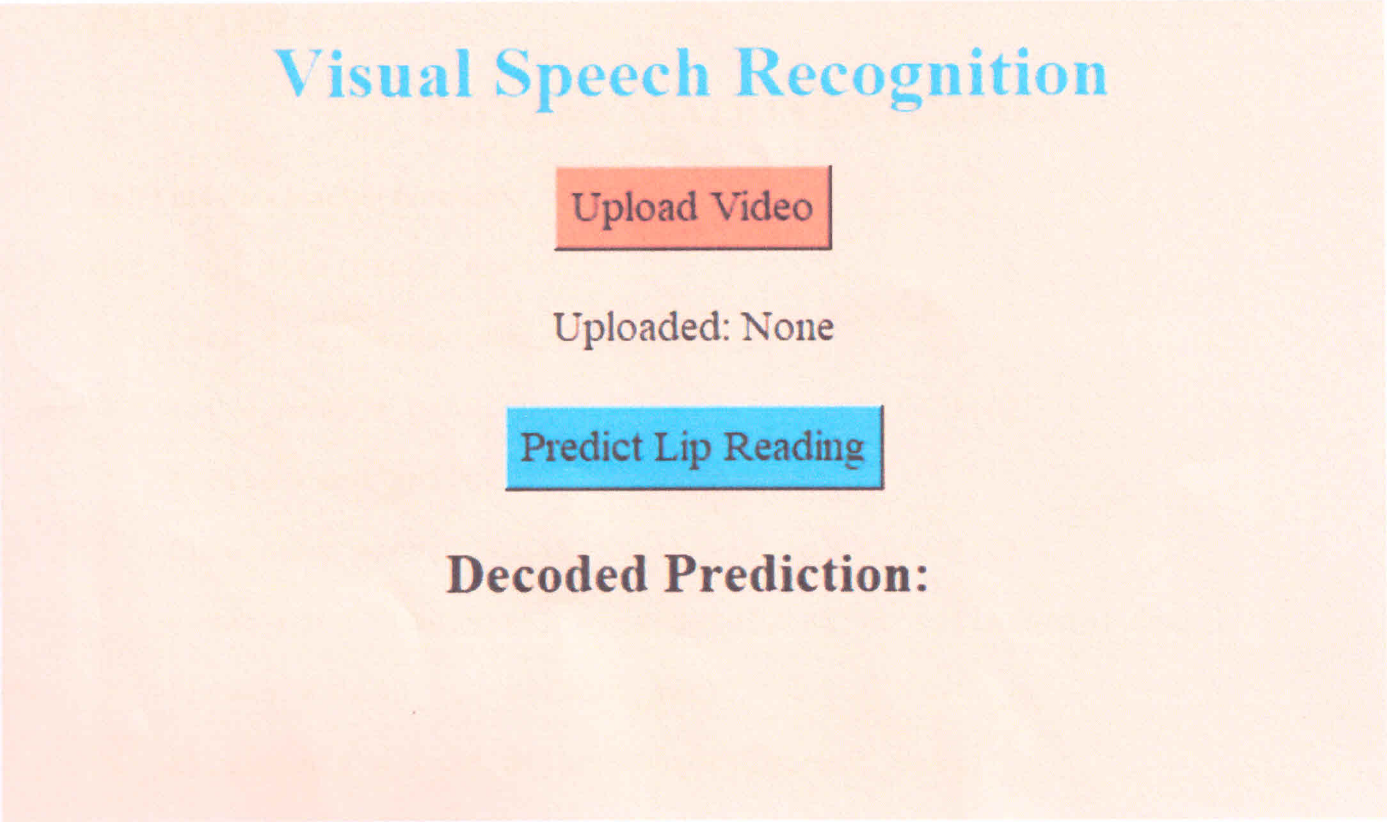
```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv3d (Conv3D)             (None, 75, 46, 140, 128   3584
                     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

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

Fig 5.3 Neural network summary



```
[<tf.Tensor: shape=(), dtype=string, numpy=b'bin blue with five soon'>]
```

Fig 5.4 Predicted text output

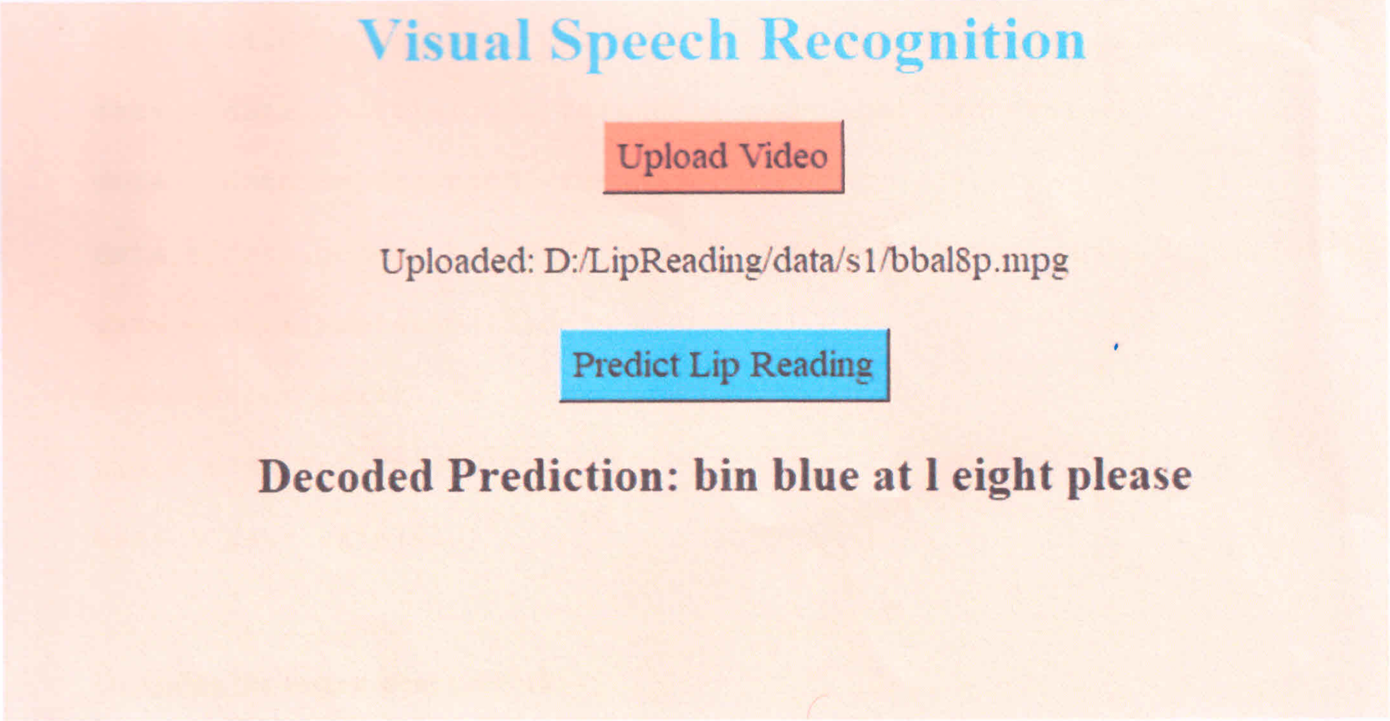


Visual Speech Recognition

Uploaded: None

Decoded Prediction:

Fig 5.5 Graphical user interface for taking inputs



Visual Speech Recognition

Uploaded: D:/LipReading/data/s1/bbal8p.mpg

Decoded Prediction: bin blue at l eight please

Fig 5.6 Graphical user interface with predicted output

CHAPTER 6

IMPLEMENTATION OUTCOMES

Build the data loading functions:

```
def load_data(path: str):
    path = bytes.decode(path.numpy())
    #file_name = path.split('/')[-1].split('.')[0]
    # File name splitting for windows
    file_name = path.split('\\')[-1].split('.')[0]
    video_path = os.path.join('data','s1',f'{file_name}.mpg')
    frames = load_video(video_path)
    alignments = load_alignments(alignment_path)
    return frames, alignments
```

Creating data pipeline:

```
data = tf.data.Dataset.list_files('./data/s1/*.mpg')
data = data.shuffle(500, reshuffle_each_iteration=False)
data = data.map(mappable_function)
data = data.padded_batch(2, padded_shapes=([75,None,None,None],[40]))
data = data.prefetch(tf.data.AUTOTUNE)
# Added for split
train = data.take(450)
test = data.skip(450)
```

Designing the deep neural network:

```
model = Sequential()
```

```
model.add(Conv3D(128, 3, input_shape=(75,46,140,1), padding='same'))

model.add(Activation('relu'))

model.add(MaxPool3D((1,2,2)))

model.add(Conv3D(256, 3, padding='same'))

model.add(Activation('relu'))

model.add(MaxPool3D((1,2,2)))

model.add(Conv3D(75, 3, padding='same'))

model.add(Activation('relu'))

model.add(MaxPool3D((1,2,2)))

model.add(TimeDistributed(Flatten()))

model.add(Bidirectional(LSTM(128,krnel_initializer='Orthogonal',
return_sequences=True)))

model.add(Dropout(.5))

model.add(Bidirectional(LSTM(128,='Orthogonal',
return_sequences=True)))

model.add(Dropout(.5))

model.add(Dense(char_to_num.vocabulary_size()+1,
kernel_initializer='he_normal', activation='softmax'))

def CTCLoss(y_true, y_pred):

    batch_len = tf.cast(tf.shape(y_true)[0], dtype="int64")

    input_length = tf.cast(tf.shape(y_pred)[1], dtype="int64")

    label_length = tf.cast(tf.shape(y_true)[1], dtype="int64")

        input_length  =  input_length  *  tf.ones(shape=(batch_len,  1),
dtype="int64")

        label_length  =  label_length  *  tf.ones(shape=(batch_len,  1),
dtype="int64")
```

```
        loss = tf.keras.backend.ctc_batch_cost(y_true, y_pred,
input_length, label_length)

    return loss
```

Test on a video:

```
sample = load_data(tf.convert_to_tensor('.\\data\\s1\\bbwm5s.mpg'))

print('~'*100, 'REAL TEXT')

[tf.strings.reduce_join([num_to_char(word) for word in sentence]) for
sentence in [sample[1]]]

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

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]
```

CONCLUSION

In conclusion, the development and implementation of the lip-reading model represent a significant stride towards addressing communication challenges faced by the hearing-impaired community. By leveraging diverse .mpg video datasets and a meticulously designed deep learning architecture, the model demonstrates commendable accuracy in transcribing spoken words from visual lip cues. The versatility showcased across varied linguistic contexts and the model's proficiency in offline predictions mark its potential as a valuable alternative to traditional communication methods, particularly in scenarios where real-time interaction is challenging.

The project's success in enhancing communication for the hearing-impaired holds promising implications for fostering inclusivity and accessibility in social, educational, and professional spheres. The model's realistic representation of lip movements contributes to its practicality and potential integration with assistive technologies, supporting the hearing-impaired in their quest for equal opportunities and participation.

As technology continues to evolve, the lip-reading model stands as a testament to the positive impact that innovative solutions can have on the lives of individuals with hearing impairments. This project not only addresses an immediate need for improved communication but also paves the way for future advancements in the intersection of deep learning and accessibility. Moving forward, the ongoing refinement and application of this lip-reading model hold the promise of a more inclusive society, where diverse communication needs are recognized and accommodated.

FUTURE SCOPE AND ENHANCEMENT

The successful development of the lip-reading model lays the groundwork for an array of future enhancements, propelling the project towards a more nuanced and inclusive communication solution for the hearing-impaired. One avenue for exploration is the integration of additional modalities, such as facial expressions and contextual cues, to enrich the model's understanding and improve transcription accuracy. This multi-modal approach could significantly enhance the overall communicative experience.

Efforts can also be directed towards improving the model's real-time adaptability. While the current implementation focuses on offline predictions, future work could involve refining the model to process live video streams in real-time, catering to dynamic communication scenarios and enabling instantaneous interaction.

Continuing to diversify and expand the training datasets is essential for the model's continual improvement. This involves incorporating a broader range of accents, languages, and speech patterns to ensure the model's proficiency in diverse linguistic contexts, promoting inclusivity on a global scale.

User interface development represents another critical area for future enhancement. Creating intuitive interfaces and applications that integrate seamlessly with the lip-reading model will enhance accessibility for both hearing-impaired individuals and those interacting with them, facilitating a smoother and more inclusive communication experience.

Additionally, exploring the feasibility of deploying the model on edge computing devices could enhance its accessibility, especially in resource-constrained environments, reducing dependency on cloud-based processing and promoting wider usage.

Adaptive learning mechanisms should be considered to enable continuous improvement. Implementing systems that allow the model to learn and evolve based on user feedback and emerging linguistic patterns ensures that it remains relevant and adaptable to the evolving communication landscape.

Furthermore, the lip-reading model can find applications in educational environments. Expanding its use to aid hearing-impaired students in accessing educational content through transcription services could significantly contribute to their academic success and integration into mainstream educational systems.

Collaboration with existing speech recognition technologies is another avenue for exploration. By combining auditory and visual cues, hybrid models could be developed to improve overall transcription accuracy in various settings, offering a more comprehensive communication solution.

Smart device integration is also a promising direction. Exploring possibilities for the seamless integration of the lip-reading model with smart devices and wearables could allow for discreet and convenient use in everyday life, further enhancing communication for the hearing-impaired.

Lastly, fostering global collaboration for dataset collection is essential. Partnering with initiatives and communities worldwide to gather diverse datasets representing different cultures, languages, and communication norms will contribute to the development of a globally inclusive lip-reading model. These future endeavors collectively signify a continuous commitment to advancing accessible communication and ensuring that the lip-reading model remains at the forefront of innovative solutions for the hearing-impaired.

REFERENCES

- S. Sumanth, K. Jyosthana, J. K. Reddy and G. Geetha, "Computer Vision Lip Reading(CV)," 2023 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC),no. 5, pp. 706-715, May 2004, doi: 10.1109/TCSVT.2004.826754.

- Bhubaneswar, India, 2022, pp. 1-6, doi: 10.1109/ASSIC55218.2022.10088386 N. Deshmukh, A. Ahire, S. H. Bhandari, A. Mali and K. Warkari, "Vision based Lip Reading System using Deep Learning," 2021 International Conference on Computing, Communication and Green Engineering (CCGE), Pune, India, 2021, pp. 1-6, doi: 10.1109/CCGE50943.2021.9776430

- S. Fenghour, D. Chen, K. Guo and P. Xiao, "Lip Reading Sentences Using Deep Learning With Only Visual Cues," in IEEE no. 8, pp. 706-715, May 2022, doi: 10.1109/TCSVT.2004.88758.

- Z. Thabet, A. Nabih, K. Azmi, Y. Samy, G. Khoriba and M. Elshehaly, "Lipreading using a comparative machine learning approach," 2018 First International Workshop on Deep and Representation Learning (IWDRL), Cairo, Egypt, 2018, pp. 19-25, doi: 10.1109/IWDRL.2018.8358210.

- V. Sahu and M. Sharma, "Result based analysis of various lip tracking systems," 2013 International Conference on Green High Performance Computing (ICGHPC), Nagercoil, India, 2013, pp. 1-7, doi: 10.1109/ICGHPC.2013.6533911.

- N. Eveno, A. Caplier and P. . -Y. Coulon, "Accurate and quasi-automatic lip tracking," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, no. 5, pp. 706-715, May 2008, doi: 10.1109/TCSVT.2008.826754.

- Zhilin Wu, P. S. Aleksic and A. K. Katsaggelos, "Lip tracking for MPEG-4 facial animation," Proceedings. Fourth IEEE International Conference on Multimodal Interfaces, Pittsburgh, PA, USA, 2007, pp. 293-298, doi: 10.1109/ICMI.2002.1167009.

- T. Afouras, J. S. Chung, A. Senior, O. Vinyals and A. Zisserman, "Deep Audio-Visual Speech Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 12, pp. 8717-8727, 1 Dec. 2006, doi: 10.1109/TPAMI.2018.2889052.

- Y. Yasui, N. Inoue, K. Iwano and K. Shinoda, "Multimodal speech recognition using mouth images from depth camera," 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, Malaysia, 2004, pp. 1233-1236, doi: 10.1109/APSIPA.2017.8282227.