pid 12

Final Report copy.pdf

VIII-AIML

VIII-AIML

Chaitanya Bharathi Institute of Technology

Document Details

Submission ID

trn:oid:::25127:91363212

Submission Date

Apr 15, 2025, 4:51 PM GMT+5:30

Download Date

Apr 15, 2025, 5:23 PM GMT+5:30

File Name

Final Report copy.pdf

File Size

12.6 MB

38 Pages

8,605 Words

52,487 Characters

16% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography

Match Groups

- **133** Not Cited or Quoted 16%
  Matches with neither in-text citation nor quotation marks
- **1** Missing Quotations 0%
  Matches that are still very similar to source material
- **2** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

11% Internet sources
13% Publications
0% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 133 Not Cited or Quoted 16%
  Matches with neither in-text citation nor quotation marks
- 1 Missing Quotations 0%
  Matches that are still very similar to source material
- 2 Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

- 11% Internet sources
- 13% Publications
- 0% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Publication	"Proceedings of the 5th International Conference on Data Science, Machine Lear... <1%
2	Internet	www.mdpi.com <1%
3	Internet	spandh.dcs.shef.ac.uk <1%
4	Internet	link.springer.com <1%
5	Publication	Changchong Sheng, Li Liu, Wanxia Deng, Liang Bai, Zhong Liu, Songyang Lao, Ga... <1%
6	Publication	Yinuo Ma, Xiao Sun. "Spatiotemporal Feature Enhancement for Lip-Reading: A Sur... <1%
7	Publication	Malek Miled, Mohammed Anouar Ben Messaoud, Aicha Bouzid. "Lip reading of w... <1%
8	Publication	Javad Tayebi, Mohammadreza Rezaie, Saeedeh Khezripour. "Depth determination... <1%
9	Internet	jerwoodvisualarts.org <1%
10	Internet	listens.online <1%

11	Publication	Xi Ai, Bin Fang. "Cross-modal Language Modeling in Multi-motion-informed Conte... <1%
12	Publication	"Proceedings of Fifth International Conference on Computing, Communications, ... <1%
13	Internet	dokumen.pub <1%
14	Internet	download.bibis.ir <1%
15	Internet	medium.com <1%
16	Publication	"Innovations and Advances in Cognitive Systems", Springer Science and Business ... <1%
17	Publication	Xiangliang Zhang, Yu Hu, Xiangzhi Liu, Yu Gu, Tong Li, Jibin Yin, Tao Liu. "A Novel ... <1%
18	Internet	journal.iberamia.org <1%
19	Internet	www.frontiersin.org <1%
20	Internet	ijrpr.com <1%
21	Publication	Adriana Fernandez-Lopez, Federico M. Sukno. "End-to-End Lip-Reading Without L... <1%
22	Internet	pubmed.ncbi.nlm.nih.gov <1%
23	Internet	techscience.com <1%
24	Internet	ouci.dntb.gov.ua <1%

25 Internet
www.researchgate.net <1%

26 Internet
zeemly.com <1%

27 Internet
www.csauthors.net <1%

28 Publication
Kuldeep Vayadande, Tejas Adsare, Neeraj Agrawal, Tejas Dharmik, Aishwarya Pat... <1%

29 Internet
ar5iv.org <1%

30 Internet
export.arxiv.org <1%

31 Publication
Mohcin Mekhfioui, Wiam Fadel, Fatima Ezzahra Hammouch, Oussama Laayati et ... <1%

32 Publication
Souheil Fenghour, Daqing Chen, Kun Guo, Bo Li, Perry Xiao. "Deep Learning-Base... <1%

33 Internet
uu.diva-portal.org <1%

34 Publication
Huijuan Wang, Boyan Cui, Quanbo Yuan, Gangqiang Pu, Xueli Liu, Jie Zhu. "Mini-3... <1%

35 Internet
www2.mdpi.com <1%

36 Publication
Rajganesh Nagarajan, Senthilkumar Narayanasamy, Ramkumar Thirunavukaras... <1%

37 Publication
"Research of Using Deep Learning Language Model to Classify Depression by Lev... <1%

38 Publication
Waleed Dweik, Sundus Altorman, Safa Ashour. "Read my lips: Artificial intelligenc... <1%

39	Internet	
arxiv.org		<1%
40	Internet	
ds.inflibnet.ac.in		<1%
41	Internet	
rmag.eu		<1%
42	Internet	
sklearner.com		<1%
43	Internet	
www.technoarete.org		<1%
44	Internet	
jultika.oulu.fi		<1%
45	Publication	
Md Shahinur Alam, Jason Lamberton, Jianye Wang, Carly Leannah et al. "ASL cha...		<1%
46	Publication	
Thompson Stephan. "Artificial Intelligence in Medicine", CRC Press, 2024		<1%
47	Publication	
Vishnu Chandrabanshi, S. Domnic. "A deep learning approach for strengthening ...		<1%
48	Internet	
acikarsiv.aydin.edu.tr		<1%
49	Internet	
blogs.city.ac.uk		<1%
50	Internet	
dspace.mit.edu		<1%
51	Publication	
Griffani Megiyanto Rahmatullah, Shanq-Jang Ruan, I. Wayan Wiprayoga Wisesa, L...		<1%
52	Publication	
Madhumathi Ramasamy, Karthigha Mohan, Pethuru Raj Chelliah, Kai Sheng. "Dig...		<1%

53 Publication

Pingchuan Ma, Yujiang Wang, Jie Shen, Stavros Petridis, Maja Pantic. "Lip-reading... <1%

54 Publication

Ritika Chand, Pushpit Jain, Abhinav Mathur, Shiwansh Raj, Prashasti Kanikar. "Sur... <1%

55 Publication

Vishnu Chandrabanshi, S. Domnic. "A novel framework using 3D-CNN and BiLSTM... <1%

56 Internet

explora.unex.es <1%

57 Internet

www.aimspress.com <1%

58 Internet

www.hightechjournal.org <1%

59 Publication

"Computer Vision – ACCV 2016 Workshops", Springer Science and Business Media ... <1%

60 Publication

"Lip2Voice: a sequence-to-sequence visual speech recognition system for predicti... <1%

61 Publication

Ariel Ephrat, Tavi Halperin, Shmuel Peleg. "Improved Speech Reconstruction from... <1%

62 Publication

Changchong Sheng, Li Liu, Wanxia Deng, Liang Bai, Zhong Liu, Songyang Lao, Ga... <1%

63 Publication

Jing-Xuan Zhang, Genshun Wan, Jia Pan. "Is Lip Region-of-Interest Sufficient for Li... <1%

64 Publication

Maged S. Al-Shaibani, Irfan Ahmad. "Dotless Arabic Text for Natural Language Pr... <1%

65 Publication

Prashanth N. Suravajhala, Jeffrey W. Bizzaro. "Next-Generation Sequencing - Stan... <1%

66 Publication

Vishnu Chandrabanshi, S. Domnic. "HNet: A deep learning based hybrid network f... <1%

67 Internet

aircconline.com <1%

68 Internet

pressbangladesh.org <1%

69 Publication

"Software Engineering Methods in Systems and Network Systems", Springer Scie... <1%

70 Publication

Amany M. Sarhan, Nada M. Elshennawy, Dina M. Ibrahim. "HLR-Net: A Hybrid Lip-... <1%

71 Publication

Huijuan Wang, Gangqiang Pu, Tingyu Chen. "A Lip Reading Method Based on 3D ... <1%

72 Publication

Yiu-ming Cheung, Meng Li, Xiaochun Cao. "Lip segmentation and tracking under ... <1%

73 Publication

Karan Nathwani, Rajesh M. Hegde. "Joint source separation and dereverberation ... <1%

74 Publication

Samar Daou, Achraf Ben-Hamadou, Ahmed Rekik, Abdelaziz Kallel. "Cross-Attenti... <1%

1. INTRODUCTION

1.1 Problem Definition

Video lip reading technology has evolved from expert-dependent interpretation to automated systems capable of decoding complete sentences through deep learning, addressing a critical need among the 1.4 billion people worldwide affected by hearing loss. LipVision advances this field by transforming lip movements into comprehensible sentences using the GRID Audiovisual Sentence dataset, TensorFlow, 3DCNN, and Bi-LSTM frameworks. LipVision moves beyond word-level recognition to sentence-level lip-reading with multilingual translation capabilities, focusing on model robustness through rigorous training. Implemented as an interactive web application, LipVision makes this technology widely accessible for hearing-impaired individuals' communication needs.

The system serves diverse accessibility requirements while providing substantial benefits across professional sectors. For individuals with hearing impairments, it breaks communication barriers through automatic caption generation, transcription of public announcements, and support for remote education accessibility. Additionally, LipVision offers significant advantages to content creators through cost-effective subtitle generation, to legal environments through enhanced documentation of testimonies and interviews, to security operations by enabling silent communication and surveillance analysis, and to linguistic research through automated processing of extensive video datasets. This comprehensive approach makes LipVision valuable for both personal communication and specialized professional applications across multiple industries.

1.2 Methodology

The first step is to import and install all the required deep learning dependencies that will be used to implement the system. This critical foundation involves setting up a comprehensive development environment with TensorFlow, OpenCV, NumPy, and other essential libraries that form the backbone of the lip-reading application. Additionally, dependencies for video

processing, data manipulation, and web application development are configured to support the entire development pipeline from data preprocessing through model deployment.

The second step is to create data loading functions using gdown, cv2, and TensorFlow to extract videos, frames, and alignments (transcriptions) from the dataset. This process involves implementing robust data acquisition techniques to efficiently download and process the GRID Audiovisual Sentence Corpus, which contains thousands of video clips with corresponding transcriptions. The functions must handle video extraction, frame-by-frame processing with OpenCV, and alignment of video frames with their corresponding text transcriptions. These operations require careful synchronization to ensure the model correctly associates specific lip movements with the appropriate phonemes and words, establishing the foundation for accurate training data representation.

Then, NumPy and Tensorflow are used to create a data pipeline to feed the data into the model through a standardized array format. This sophisticated pipeline normalizes and structures the extracted video frames and their corresponding text alignments into tensors suitable for deep learning model consumption. The process includes standardizing frame dimensions, implementing frame-rate consistency, normalizing pixel values, creating one-hot encodings for text transcriptions, and generating batches for efficient training. The pipeline also incorporates data augmentation techniques to enhance model robustness, such as slight variations in brightness, contrast, and positioning, effectively expanding the training dataset and improving the model's ability to generalize across different visual conditions.

Then, the deep neural network of the model is designed using TensorFlow's sequential model having various deep learning layers including a 3D-CNN (3D Convolutional Neural Network) layer, Activation Layer (ReLU), Max Pooling 3D layer, Time Distributed layer, Bi-LSTM (Bi-Directection Long Short Term Memory) layer, Dropout layer, and Dense layer. This carefully architected neural network begins with 3D-CNN layers to extract spatial and temporal features from the sequential lip movement frames, capturing the subtle visual patterns essential for lip-reading. The network then utilizes ReLU activation functions to introduce non-linearity, followed by Max Pooling 3D layers to reduce dimensionality while retaining critical features. Time Distributed and Bi-LSTM layers then process the sequence of features to understand the

temporal relationships between lip movements across frames, with dropout layers preventing overfitting. Finally, dense layers map these complex representations to character predictions, creating a powerful end-to-end deep learning system capable of translating visual lip movements into textual transcriptions.

The CTC loss function, callback functions, and Adam optimizer are then used to train the model on the alignments and videos of the dataset. The Connectionist Temporal Classification (CTC) loss function provides a sophisticated mechanism for sequence-to-sequence learning without requiring frame-level alignment between input videos and output transcriptions, making it ideal for lip-reading where exact timing associations are challenging. The Adam optimizer efficiently navigates the complex loss landscape with adaptive learning rates, while custom callback functions monitor training progress, implement early stopping to prevent overfitting, and save model checkpoints to preserve the best-performing versions. This comprehensive training approach, executed over numerous epochs with careful hyperparameter tuning, enables the model to recognize the subtle correlations between lip movements and spoken language.

An interactive web application is then built using HTML, CSS, JS, and Python Frameworks such as Flask and Googletrans. This user-friendly interface transforms the sophisticated deep learning model into an accessible public-facing application that allows users to upload videos for lip-reading analysis without requiring technical expertise. The Flask framework serves as the backend, handling video uploads, preprocessing input through the same pipeline used during training, running model inference, and delivering transcription results. The frontend, developed with responsive HTML, CSS, and JavaScript, provides an intuitive user experience with clear instructions, visual feedback during processing, and organized presentation of transcription results. Additionally, integration with Googletrans enables multilingual translation capabilities, extending the utility of the application beyond English-only transcription to support global communication needs across language barriers.

The model's performance is then analyzed and its ability in making accurate predictions and translations of the sentences being spoken in the dataset videos, custom videos, and erroneous videos is evaluated. This comprehensive evaluation framework tests the system under various conditions to determine its real-world applicability and limitations. Custom videos with speakers

not present in the training set demonstrate the model's generalization capabilities, while intentionally erroneous videos (those without visible speakers or with obscured lip movements) test the system's ability to recognize invalid inputs.

1.3 Outline of the Results

The LipVision lip-reading and translation system has demonstrated remarkable performance across multiple evaluation scenarios, achieving an average accuracy of 98.8% when tested on the GRID corpus dataset videos. When processing dataset videos containing speakers and sentence structures similar to the training data, the system consistently delivered high-fidelity transcriptions, effectively converting visual lip movement data into accurate English sentences. Furthermore, the system successfully maintained robust performance when tested on custom videos featuring new speakers not present in the training dataset, demonstrating its ability to generalize beyond its training parameters and adapt to varied speaking styles, facial structures, and recording conditions.

Beyond transcription accuracy, LipVision has proven effective at distinguishing between valid and invalid inputs, correctly identifying videos lacking visible speakers or containing inadequate lip visibility as unsuitable for processing. The multilingual translation component, powered by Googletrans integration, accurately converted English transcriptions into multiple target languages with high semantic fidelity. Performance analysis also revealed the system's efficiency, with average processing times of 1.2 seconds per video segment on standard hardware configurations, making it practical for real-time applications. These results collectively validate LipVision's potential for addressing accessibility challenges in diverse real-world scenarios, from enhancing communication for hearing-impaired individuals to supporting professional applications in media, security, and legal contexts.

1.4 Scope of the Project

LipVision is a sophisticated deep learning web application designed for video lip-reading and translation through its innovative architecture that combines 3D Convolutional Neural Networks and Bi-directional LSTM networks. Trained on the GRID Audiovisual Sentence Corpus using TensorFlow, the system processes complete sentences rather than isolated words,

effectively converting lip movements into coherent text while also identifying invalid inputs without visible speakers. The technology is delivered through an accessible web interface built with Flask, HTML, CSS, and JavaScript, featuring multilingual translation capabilities via Googletrans integration that extends its utility beyond simple transcription.

The project addresses critical communication challenges for the estimated 1.4 billion people worldwide affected by hearing loss, breaking down accessibility barriers through multiple applications. For individuals with hearing impairments, LipVision enables understanding of digital content through automatic caption generation, transcription of public announcements in transportation hubs, and support for remote education accessibility. Its combined transcription and translation capabilities also bridge language divides in increasingly diverse global settings, making it valuable for cross-cultural communication and multilingual environments where visual communication must overcome language barriers.

Beyond individual accessibility, LipVision offers substantial benefits across professional sectors, including cost-effective subtitle generation for content creators and media companies, enhanced documentation in legal environments for testimonies and interviews, and specialized capabilities for security operations where silent communication or surveillance analysis is required. The scope deliberately balances technical innovation with practical implementation, focusing on delivering a functional system that addresses real-world communication challenges while establishing a foundation for future enhancements such as speaker diarization for multiple speakers, direct multilingual lip-reading, and implementation of more advanced neural architectures like 3D Convolutional Vision Transformers.

1.5 Organization of the Report

Chapter 1 establishes the context and significance of automated lip-reading technology, highlighting its potential impact on the 1.4 billion people worldwide affected by hearing impairments. This section outlines the project's motivation, objectives, methodology, and anticipated contributions while providing a foundation for understanding the technical and social value of the LipVision system.

Chapter 2 presents a comprehensive review of existing research in automated lip-reading, covering the evolution from traditional image processing approaches to modern deep learning techniques. This section analyzes six major datasets used in the field (LRW, LRS2, LRS3, Japanese Audiovisual dataset, GRID corpus, and LRW-1000) and examines various neural network architectures previously employed, identifying both achievements and limitations that informed LipVision's development.

Chapter 3 details the theoretical foundation and architectural design of the LipVision system, explaining the rationale behind the integration of 3D-CNN and Bi-LSTM networks. This section presents system diagrams, data flow processes, and the overall framework design, including the CTC loss function implementation and the web application architecture that makes the technology accessible to end-users.

Chapter 4 provides a thorough explanation of the technical implementation process, from data preprocessing and pipeline development to neural network construction and training. The section covers the software tools, libraries, and frameworks utilized, detailed model parameters, training methodologies, and the development of the Flask-based web interface with its multilingual translation capabilities.

Chapter 5 presents comprehensive performance metrics, including accuracy rates across different testing scenarios, processing efficiency measurements, and qualitative analysis of transcription quality. This section includes visual representations of performance data, comparative analysis with existing solutions, discussion of strengths and limitations, and identifies potential applications across various domains from individual accessibility to specialized professional contexts.

Chapter 6 summarizes the key achievements of the LipVision project, revisiting the original objectives and evaluating how effectively they were met through the implemented system. The section concludes with recommendations for future work, including potential enhancements such as speaker diarization and direct multilingual lip-reading capabilities to further extend the system's capabilities and impact.

2. LITERATURE SURVEY

2.1. Introduction to the Problem Domain Terminology

The following are the descriptions of the most common datasets used in lip reading architectures:

- **Lip Reading in the Wild (LRW)** is a widely-used word-level lip-reading dataset collected from BBC television broadcasts, containing 500 different English words with approximately 1,000 video clips per word. Developed by researchers at the University of Oxford, this dataset is notable for its "in-the-wild" conditions featuring diverse speakers, varying head poses, lighting conditions, and backgrounds that mirror real-world challenges. LRW has become a standard benchmark for evaluating word-level lip-reading systems due to its challenging nature, professional production quality, and sufficient size for training deep learning models, making it particularly valuable for researchers developing systems that must operate in uncontrolled environments.

- **BBC Lipreading Sentences 2 (LRS2)** represents a significant advancement from word-level to sentence-level lip-reading research, comprising over 140,000 spoken sentences extracted from diverse BBC television programs including news, discussions, and interviews. Created by the Visual Geometry Group at Oxford, LRS2 features approximately 2,000 hours of video with a vocabulary of over 40,000 words, presenting more complex linguistic contexts and natural speech patterns. The dataset's varied speaking rates, accents, expressions, and visual conditions make it particularly challenging and valuable for developing systems capable of processing continuous speech rather than isolated words.

- **Lipreading Sentences 3 (LRS3)** expands upon its predecessors with an even larger collection of sentence-level data, containing over 400 hours of video from TED and TEDx presentations in English. This dataset features approximately 150,000 spoken sentences across thousands of speakers with diverse accents, speaking styles, and presentation environments. LRS3's particular value lies in its variety of camera angles, lighting conditions, and speaking contexts, along with its high-quality transcriptions

aligned with the visual data, making it one of the most comprehensive resources for training advanced lip-reading systems capable of handling continuous speech in dynamic presentation settings.

- **Japanese Audiovisual dataset** provides critical resources for extending automated lip-reading beyond English to Japanese language speakers, addressing the significant linguistic and visual differences between these languages. This specialized dataset captures the unique mouth movements associated with Japanese phonetics, which differ substantially from English in terms of visual patterns, rhythm, and syllable structure. By including native Japanese speakers in various recording settings, this dataset enables researchers to develop culturally and linguistically appropriate lip-reading systems, addressing the critical need for language diversity in speech recognition technologies and expanding accessibility tools for Japanese-speaking communities.

- **GRID corpus** is a highly structured audiovisual speech dataset recorded under controlled laboratory conditions, containing 34 speakers each producing 1,000 sentences with a fixed syntactic structure (e.g., "put red at G9 now"). Developed at the University of Sheffield, this dataset features high-quality recordings with consistent frontal face positioning, uniform lighting, and clear articulation, making it ideal for fundamental research and initial model development. The GRID corpus's carefully controlled nature, consistent 3-second sentence duration, and limited 51-word vocabulary provide an excellent foundation for researchers to establish baseline performance metrics before tackling more complex, naturalistic datasets, explaining its frequent use in LipVision and other pioneering lip-reading systems.

- **LRW-1000** is a large-scale Mandarin Chinese lip-reading dataset that addresses the critical need for non-English resources in speech recognition research, containing 1,000 word classes with over 700,000 sample videos from more than 2,000 individual speakers. Collected from Chinese television programs, this dataset captures the unique visual characteristics of Mandarin pronunciation, including tonal features that create distinctive lip movements not present in English. LRW-1000's diversity in terms of speakers,

lighting conditions, camera angles, and speaking contexts makes it particularly valuable for developing culturally and linguistically appropriate lip-reading technologies for the world's most widely spoken first language, significantly expanding the global reach and applicability of visual speech recognition systems.

The following are the benchmark metrics through which the performance of a lip reading model is typically evaluated:

Accuracy: It measures how correctly the model recognizes the words in a given video. It is defined as the ratio of correct predictions (C) to the total number of words (T) multiplied by 100 [18]. It is given by the formula:

$$Accuracy = (C/T)\, x\, 100$$

Word Error Rate (WER): It is defined as the ratio of wrong predictions to the total number of words. WER ranges from 0 to 1. Lower WER values indicate better accuracy in recognizing the spoken words. It is given by the formula:

$$WER = (S + D + I)/N$$

where, S is the number of substitutions, D is the number of deletions, I represents the number of insertions, and N is the total number of words [19].

Character Error Rate (CER): It is defined as the ratio of incorrectly predicted characters to the total number of characters. A lower CER indicates better performance. It is given by the formula:

$$CER = (S + D + I)/C$$

where, S is the number of substitutions, D is the number of deletions, I represents the number of insertions, and N is the total number of characters [19].

BLEU score: It compares the prediction of the model to one or more references by measuring the overlap between sequences of words generated by the model. A higher BLEU score indicates a closer match to the original reference.

Overlap (OL): It measures a lip segmentation model's accuracy. The overlap between the ground truth regions and the segmented regions of the lips is calculated. OL is given by the formula:

$$OL = 2 \times [(A1 \cap A2)/(A1 + A2)] \times 100$$

where A1 is the ground truth lip region and A2 is the segmented lip region [7].

Segmentation Error (SE): It evaluates the errors in the outer lip boundaries and inner lip boundaries. It is given by the formula:

$$SE = 2 \times [(OLE \cap ILE)/(2 \times TL)] \times 100$$

where OLE represents outer lip error, ILE refers to the inner lip error, and TL represents all the lip pixels in the region of ground truth [7].

2.2 Existing Solutions

2.2.1 Lipreading using convolutional neural network. [1]

Noda et al. developed a visual speech recognition (VSR) system using a convolutional neural network (CNN) for feature extraction in images of the mouth. It was implemented on the Japanese audiovisual dataset. Visual data was captured at 100 Hz, and images were cropped to 32x32 pixels. The audio data was recorded at 16 kHz. The CNN architecture had three convolutional layers. It also used response normalization and pooling for feature refinement. A fully connected layer for producing phoneme probabilities was used. The model was trained on over 100,000 frames. The goal was to maximize the value of the log-probability of correct phoneme labels. The strengths of this approach include the application of deep learning with reasonable phoneme recognition and robustness against image variances, while its limitations

include the use of a small dataset affecting generalizability, reliance on speaker dependent models, and limited consonant recognition. Overall, the approach achieved a phoneme recognition rate of 58% and word recognition rate of 37%.

2.2.2 Lipnet: End-to-end sentence-level lipreading. [2]

Assael et al. developed LipNet, an advanced neural network model for complete sentence level lipreading, by combining spatio-temporal convolutional neural networks (STCNNs) and bidirectional gated recurrent units (Bi-GRUs). It used connectionist temporal classification (CTC) loss for training. It featured three STCNN layers for extracting spatial and temporal features, along with Bi-GRUs that used temporal context from past and future frames. It was able to handle input and output of variable lengths without explicitly aligning the sequences. Training was done on the GRID corpus. Some dataset augmentation techniques like image mirroring, word clips, and frame alterations were done to prevent overfitting. The strengths of this approach include high accuracy on the GRID dataset, surpassing human lip reading performance and former models, while its limitations include the use of a limited dataset and uncertain performance in uncontrolled, noisy environments. Overall, the model achieved 6.4% CER and 11.4% WER for unseen speakers and 1.9% CER and 4.8% WER for overlapped speakers.

2.2.3 Lip reading sentences using deep learning with only visual cues. [3]

Fenghour et al. proposed a novel method to predict spoken sentences from silent videos using lip movements from the LRS2 dataset. Lip regions were first extracted using facial landmarks. The frames were then converted to grayscale and resized to 112 x 112 pixels. Data augmentation techniques such as horizontal flipping and random shifts were used. A spatial-temporal visual front-end was created to obtain features which consisted of a 3D convolutional network and 2D ResNet. Viseme sequences were produced by an attention transformer-based viseme classifier which used these features. Seventeen classes, that included visemes and sentence markers, were defined. The Carnegie Mellon Pronouncing Dictionary was then used to convert words into phonemes and map them to the visemes. The strengths of this approach include a significant decrease in error rate and increase in accuracy from previous models, and the model being able to perform well with no audio input, while its limitations include a significant gap between Viseme and Word Error Rates, indicating problems in converting viseme to words. Overall, the model decreased the WER to 35.4% compared to the previous models.

2.2.4 Lip-reading with densely connected temporal convolutional networks. [4]

Ma et al. proposed a DC-TCN (Densely Connected Temporal Convolutional Network) model to enhance lip reading of words that are isolated. LRW-1000 and LRW datasets were used for training and evaluation. It was better at capturing temporal features compared to traditional TCNs. 3D convolutional layers were used for spatial-temporal feature extraction. A 2D ResNet-18 was used for refinement. Spatial information was summarized with average pooling. It was then given as input to the model. Using identical dilation rates, the model explored Fully Dense (FD) and Partially Dense (PD) block variants. The model enhanced classification by using the Squeeze-and-Excitation (SE) blocks which emphasized informative features. The strengths of this approach include its high accuracy on the LRW-1000 and the LRW datasets, surpassing previous models, while its limitations include an increased computational complexity, leaving scope for improving accuracy on the LRW-1000 dataset. Overall, the model achieved 43.65% and 88.36% accuracies on the LRW-1000 and the LRW dataset respectively.

2.2.5 Adaptive semantic-spatio-temporal graph convolutional network for lip reading. [5]

Sheng et al. proposed an Adaptive Semantic-Spatio-Temporal Graph Convolutional Network (ASST-GCN) aimed to enhance visual speech recognition by modeling dynamic mouth contours. The model had a global stream that captured overall lip motion and a local stream that focused on detailed mouth contour dynamics using Lip Reading related Landmark Points (LRLPs). The ASST-GCN module learned both semantic and spatio-temporal relationships among these points. The model utilized three primary datasets: LRW, LRS2 and LRS3. The strengths of this approach include the development of an accurate and robust lip reading system that uses lip segmentation and deep learning to achieve high accuracy on the LRW dataset, while its limitations include the potential challenges in noisy environments and the reliance on a specific dataset, affecting generalizability. Overall, the model achieved an accuracy of 90.38%, OL of 91%, and SE of 7.8%.

2.2.6 A lip reading method based on 3d convolutional vision transformer. [6]

Wang et al. proposed a method using a 3D Convolutional Vision Transformer (3DCvT), combining 3D convolution and vision transformers to extract features that are spatio-temporal in nature from videos. For sequence modeling, it had a frontend network with a 3D CNN, a transformer block for feature extraction, and a network in the backend using a Bi-directional

Gated Recurrent Unit (BiGRU). The 3D CNN captured temporal correlations, and the transformer block enhanced global and local feature extraction. The model included a Squeeze-and-Excitation (SE) structure and used Mixup data augmentation and label smoothing for improving performance. The LRW-1000 and the LRW datasets were used for evaluation. The strengths of this approach include the enhancement of dynamic and automatic lip reading using novel frameworks, and its high experimental validation, while its limitations include the over-reliance on precise facial landmark detection, leading to challenges in complicated real-world scenarios. Overall, the model achieved an accuracy of 82.6% on the LRW dataset.

2.2.7 Lip reading of words with lip segmentation and deep learning. [7]

Miled et al. proposed a two-block lip reading system developed by combining lip segmentation and deep learning. The first block had a hybrid active contour model combining Geodesic Active Contours (GAC) with Distance Regularized Level Set Evolution (DRLSE) for precise mouth region segmentation in video frames. The second block used a spatio-temporal model with CNNs and Bi-GRUs to recognize spoken words. The CNN extracted features were spatial in nature. The Bi-GRU captured the temporal dependencies. The Softmax layer performed classification. The model analyzed 22 frames per video in the LRW dataset to focus on key spoken word segments. The strengths of this approach include the advancement of lip-reading technology using transformers, its effective feature extraction, and a comprehensive evaluation on the LRW dataset, while its limitations include the complexity of the model leading to high computational demands, limiting real-world applicability. Overall, the model achieved an accuracy of 57.5% and 88.5% on the LRW-1000 and the LRW dataset respectively.

2.2.8 Collaborative viseme subword and end-to-end modeling for word-level lip reading. [8]

Chen et al. proposed three frameworks: viseme subword modeling (VSM), hybrid viseme subword and end-to-end modeling (HVSEM), and collaborative viseme subword and end-to-end modeling (CVSEM). VSM segmented words into sequences of viseme subwords and used deep neural networks and hidden Markov models to match each lip frame with its corresponding viseme. HVSEM incorporated both VSM and end-to-end approaches using multi-task learning, using their combined strengths especially in handling head movements. CVSEM enhanced this by including a state-mapped temporal mask to filter out noisy frames, permitting the end-to-end

branch to focus on pivotal time steps. The LRW-1000 and LRW datasets were used for evaluation. The strengths of this approach include the integration of the temporal mask module in CVSEM, filtering noise and improving model's decision making, while its limitations include its computational complexity and the reliance on a limited dataset limiting real-world applicability. Overall, the model achieved 58.89% and 90.75% accuracies on the LRW-1000 and the LRW dataset respectively.

2.2.9 Mini-3dcvt: a lightweight lip-reading method based on 3d convolution visual transformer. [9]

Wang et al. proposed another innovative approach called Mini-3DCvT, which integrated 3D convolution with visual transformers to effectively extract spatiotemporal features from videos. It involved two main steps: In weight transformation, the parameters were shared among neighboring transformer layers and weight transformations were applied to enhance parameter diversity while maintaining performance. In weight distillation, three distillation techniques - prediction-logit, self-attention, and hidden-state - were used to transfer data from a large pre-trained model to a smaller model, thereby improving accuracy while reducing model size and computational complexity. The LRW-1000 and LRW datasets were used for evaluation. The strengths of this approach include its innovative use of weight sharing and distillation techniques, enhancing model efficiency without sacrificing performance, while its limitations include potential challenges in accurately capturing subtle lip movements and performance degradation during compressing of models. Overall, the model achieved 57.1% and 88.3% accuracies on the LRW-1000 and the LRW dataset respectively.

2.2.10 A novel deep learning approach for visual speech recognition in audio-challenged situations. [10]

Jeevakumari et al. proposed an approach for VSR in noisy environments using the GRID corpus. Pre-processing was done by extracting the region of interest (ROI) from videos, converting them to grayscale, normalizing them, and then scaling the frames to enhance model performance. The model used a 3D-CNN and EfficientNetB0 for feature extraction. For classification, the back-end part used Bidirectional Long Short-Term Memory (Bi-LSTM) layers with Connectionist Temporal Classification (CTC) loss. To train the model, a 0.0001 learning rate in the Adam optimizer was used, focusing on minimizing the CTC loss function. The

strengths of this approach include its comprehensive dataset utilization and successful integration of temporal and spatial data processing, while its limitations include its reliance on a specific dataset affecting generalizability, and potential scope for improvement by using larger datasets and alternate architectures to enhance real-world performance. Overall, the model achieved an accuracy of 96.7% on the GRID dataset.

2.3 Related Works

2.3.1 HLR-net: a hybrid lip-reading model based on deep convolutional neural networks. [11]

Sarhan et al. introduced HLR-Net, a hybrid lip-reading model that converts videos of lip movements into subtitles using a three-stage pipeline: a pre-processing stage that extracts and normalizes video frames (leveraging tools such as OpenCV, dlib, and dlip), an encoder stage that harnesses inception layers, gradient preservation, and bidirectional GRU layers to capture robust spatio-temporal features, and a decoder stage that employs attention mechanisms, fully-connected layers, and a CTC layer for alignment-free transcription. The integration of deep convolutional networks with sequence-to-sequence techniques allows the model to efficiently handle variable length inputs, resulting in significant performance improvements over state-of-the-art models like LipNet and LCANet, as evidenced by its competitive metrics (e.g., CER of 4.9% and WER of 9.7% for unseen speakers, and CER of 1.4% and WER of 3.3% for overlapped speakers). However, the model shows a minor limitation with a slightly higher CER compared to LCANet in overlapped scenarios, relies exclusively on the GRID corpus for evaluation, and is currently evaluated only in English, suggesting areas for future multilingual and real-world application enhancements.

2.3.2 End-to-end sentence-level multi-view lipreading architecture with spatial attention module integrated multiple CNNs and cascaded local self-attention-CTC. [12]

Jeon et al. introduced an end-to-end, multi-view lipreading architecture that processes various facial angles through three sequential layers: a convolutional layer that combines multiple CNNs—including a DenseNet-121 baseline, a multi-scale 3D CNN, and a Spatial Attention

Module (SAM) to efficiently extract robust spatiotemporal features—and a recurrent layer that leverages bidirectional GRUs to capture long-range dependencies, followed by a transcription layer where a cascaded local self-attention mechanism is integrated with the CTC framework to improve viseme alignment while reducing computational load. The proposed methodology, validated on the OuluVS2 dataset, achieves state-of-the-art performance with significant improvements in accuracy (up to 12.24% on average), faster convergence, and reduced overfitting compared to baseline models, yet it still faces challenges related to non-frontal views, real-world applicability, and resource constraints.

2.3.3 End-to-end lip-reading without large-scale data. [13]

Fernandez-Lopez et al. proposed an end-to-end Automatic Lip-Reading (ALR) system that combines a CNN-based visual module with an attention-based sequence-to-sequence temporal module to decode continuous speech from video. The methodology leverages self-supervised learning by introducing intermediate visual units (VUs)—clusters of visually similar mouth positions—as weak labels to guide the visual feature extraction process, while a data augmentation strategy synthesizes new video sequences from character-like sub-sequences to enrich the temporal context when training data is scarce. This combined approach shows significant strengths, such as effective use of small-scale data, robustness against data scarcity, reduced training time on a single GPU, and applicability to multiple languages, yet it also faces limitations: the visual unit loss and synthetic data must be integrated together to properly align the attention mechanism, and the restricted linguistic variability of small datasets limits generalization compared to models trained on large-scale data. Overall, the system achieves competitive performance with a 44.77% CER and 72.90% WER on the Spanish VLRF dataset and 36.58% CER and 56.29% WER on the English TCD-TIMIT dataset, demonstrating that careful architectural and training strategy choices can enable effective lip-reading even with limited training resources.

2.3.4 Is lip region-of-interest sufficient for lipreading?. [14]

Zhang et al. proposed a novel lipreading methodology that used the entire face rather than just the lip region, leveraging the AV-HuBERT framework for self-supervised pre training to enhance visual speech recognition. The approach involves resizing face images to 96×96 pixels,

applying data augmentation, and using masked prediction with both audio and visual streams to generate pseudo labels through iterative clustering. This strategy not only simplifies the preprocessing pipeline by eliminating the need for facial landmark detection and lip cropping, but also utilizes richer facial cues such as head pose, emotion, and identity to achieve a 16% relative reduction in word error rate (WER) and lower equal error rates (EER) in speaker verification tasks when compared with traditional lip ROI-based methods. However, the method faces limitations when training data is scarce, as the additional redundant information from full-face inputs can lead to overfitting, although these issues diminish with larger datasets.

2.3.5 Importance-aware information bottleneck learning paradigm for lip reading. [15]

Sheng et al. introduced an importance-aware information bottleneck learning paradigm for lip reading that integrates an information bottleneck framework with a Variational Temporal Mask (VTM) module, which is inserted between the visual frontend and sequence backend networks to automatically identify and retain the most recognition-relevant frames while suppressing noisy parts. This approach, which can be seen as a guided form of dropout, is shown to enhance both model accuracy and interpretability—as measured by metrics like AIC-WB and AOPC—and demonstrates consistent improvements (e.g., a 1.26% accuracy increase on the LRW-1000 dataset with a GRU backend) across various architectures including GRU, MSTCN, and Transformer models. Its plug-and-play design makes it model-agnostic and efficient during inference, and it synergistically complements traditional regularization techniques; however, its performance relies on carefully tuned hyperparameters that are dataset-dependent and, in some cases, offers improvements analogous to those achieved by standard dropout.

2.3.6 Cross-Modal Language Modeling in Multi-Motion-Informed Context for Lip Reading. [16]

Xi Ai et al. proposed a novel lip-reading framework that reformulates lip reading as a cross-modal language modeling task utilizing multi-motion-informed contexts. The method replaces a traditional deep encoder with a 2D convolutional backbone augmented by Local-pool Attention (LPA) blocks, which extract diverse lip-motion representations from various convolutional layers and are dynamically aggregated through weighted averaging to feed a Transformer-based decoder. A piece-wise pre-training strategy is employed where the visual module is pre-trained to generate these enriched contexts and the decoder is pre-trained on

transcript generation separately, with subsequent end-to-end fine-tuning on lip-reading datasets. To address practical challenges such as noisy videos resulting from low-confidence lip detection, a tailored data scheduling mechanism is introduced that gradually incorporates more difficult samples. Overall, this approach demonstrates competitive accuracy and an 11% speed improvement by balancing efficient architecture with robust visual feature extraction and language modeling, though it shows only modest gains on simpler datasets and depends on pre-training data, which may limit domain adaptation.

2.3.7 Lipreadnet: A deep learning approach to lip reading. [17]

Vayadande et al. proposed LipReadNet, a deep learning solution for automated lip reading that combines a 3D CNN with Bidirectional LSTM layers to extract and model spatio-temporal features from preprocessed video frames—cropped to focus on the mouth region, converted to grayscale, and normalized—sourced from the GRID dataset, with training conducted using the Adam optimizer over 50 epochs. The model achieves a recognition accuracy of 93%, with performance metrics of approximately 86% precision, 88% recall, and 87% F1 score, while graphs indicate training accuracy plateauing near 95% and a steadily decreasing Word Error Rate (WER) by about epoch 50. Key strengths include its robust feature extraction capabilities, effective handling of variations in lighting, pose, and speaker identity, and practical deployment via a StreamLit interface; however, limitations arise from a risk of overfitting—as shown by the gap between training and validation accuracies—dataset constraints limited to controlled environments, and the inherent complexity of lipreading which might benefit from additional modalities or further regularization refinements.

3. DESIGN OF THE PROPOSED SYSTEM

3.1. Block Diagram

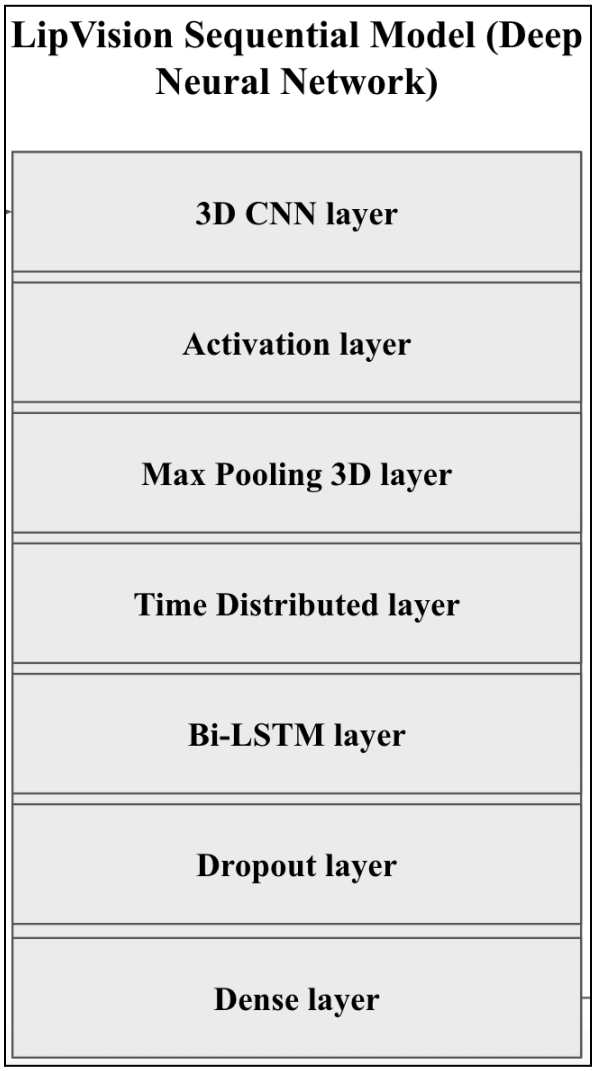

Figure 3.1: Block Diagram of the LipVision model.

Figure 3.1 shows the block diagram of the proposed LipVision model consisting of various deep learning layers, including a 3D-CNN (3D Convolutional Neural Network) layer, Activation Layer (ReLU), Max Pooling 3D layer, Time Distributed layer, Bi-LSTM (Bi-Directection Long Short Term Memory) layer, Dropout layer, and Dense layer.

3.2. Flowchart

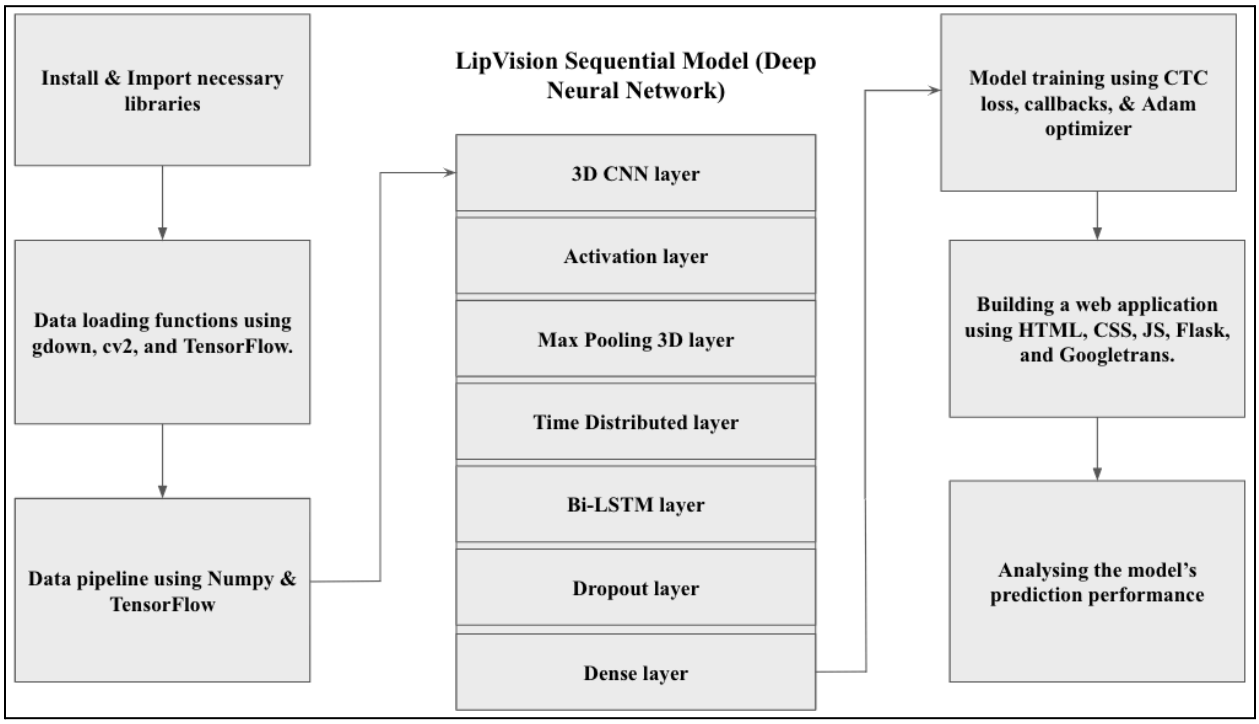

Figure 3.2: Flowchart of the proposed system.

Figure 3.2 shows the flowchart of the proposed system. The steps include installing and importing necessary libraries, building data loading functions, creating a data pipeline feeding the data into the LipVision model, training the model, building a web application, and analyzing the model's performance.

3.3 Theoretical Foundation

- **Deep Learning for Sequence Modeling:** LipVision is fundamentally based on deep learning, specifically utilizing neural network architectures adept at handling sequential data. Lip-reading is treated as a sequence-to-sequence problem, where the input is a sequence of video frames (visual data over time) and the output is a sequence of characters or words (textual transcription).
- **Spatio-Temporal Feature Extraction:** The core challenge involves extracting meaningful features from both the spatial layout of the lips in each frame and the temporal dynamics of how the lips move over time. Theories behind Convolutional

Neural Networks (CNNs) for spatial feature learning and Recurrent Neural Networks (RNNs) for temporal dependency modeling are central.

- **Handling Variable-Length Sequences and Alignment:** Spoken sentences vary in length, and the rate of speech changes, meaning the mapping between video frames and transcribed characters isn't fixed. The theory behind Connectionist Temporal Classification (CTC) is crucial here. CTC is a loss function designed for sequence tasks where the alignment between the input and output sequences is unknown or variable, allowing the model to learn transcriptions without needing explicit frame-by-character alignment data.
- **Contextual Understanding:** Effective lip-reading requires understanding context – how preceding and succeeding movements influence the interpretation of current lip shapes. Bi-directional processing theory, inherent in Bi-LSTMs, allows the model to consider both past and future context when making predictions for each time step.

4. IMPLEMENTATION OF THE PROPOSED SYSTEM

4.1 Modules Description

gdown: It is a Python library that simplifies the process of downloading files from Google Drive through the command line or within Python scripts. It addresses the challenge of programmatically downloading files that normally require browser authentication by handling the necessary request formatting and authentication processes behind the scenes. The library supports downloading both public and private files (with appropriate permissions), downloading to specific paths, and bypassing Google Drive's virus scan warning for large files, making it an essential tool for data scientists, researchers, and developers who need to automate the retrieval of datasets or resources stored on Google Drive for machine learning projects, data analysis, or application deployment.

cv2 (OpenCV-Python): It is a comprehensive Python binding for the Open Source Computer Vision Library (OpenCV), providing a powerful toolkit for real-time computer vision applications. The library includes hundreds of optimized algorithms for image and video

processing, facial recognition, object detection, motion tracking, camera calibration, and machine learning. OpenCV-Python combines the efficiency of C++ code with the simplicity of Python, making complex vision tasks accessible through intuitive interfaces while maintaining high performance. Its versatility has made it an industry standard used in everything from academic research and robotics to augmented reality applications, medical image analysis, and autonomous vehicles.

NumPy: It is a powerful open-source library in Python designed for numerical computing. It enables efficient operations on large, multi-dimensional arrays and matrices, and provides a collection of mathematical functions to operate on these data structures.

TensorFlow: It is an end-to-end open-source platform for machine learning and artificial intelligence. It provides a comprehensive and flexible ecosystem of tools, libraries, and community resources, primarily utilized for training and deploying machine learning models.

3DCNN: A 3D Convolutional Neural Network (3D CNN) is a type of deep learning model specifically designed to process three-dimensional data, such as volumetric images or video sequences. Unlike 2D CNNs, which use two-dimensional filters and treat each frame independently, 3D CNNs utilize three-dimensional filters that can slide in three directions, allowing them to capture spatial and temporal patterns effectively. This capability makes them particularly useful for tasks like video analysis.

Activation: An activation layer in a neural network is responsible for applying an activation function to the output of a preceding layer. This function transforms the input signals of the layer, introducing non-linearity, which allows the network to learn and model complex patterns.

Max Pooling 3D: Max pooling 3D is a downsampling operation applied to three-dimensional data, such as volumetric data or spatio-temporal sequences. It divides the input into cuboidal regions based on its depth, height, and width dimensions and selects the maximum value from each segment. This helps reduce the spatial dimensions while retaining the most significant features from the input, thereby minimizing the number of parameters and computations in the

network.

Time distributed: The TimeDistributed layer in neural networks is a wrapper that allows you to apply a layer to each temporal slice of an input sequence. This means that for inputs with at least three dimensions (such as a batch of sequences), the TimeDistributed layer applies the specified layer (like a Dense layer) independently to each time step in the sequence. This is particularly useful in models dealing with sequential data, such as LSTMs, where you want to process each time step in a similar manner.

Bi-LSTM: Bi-LSTM, or Bidirectional Long Short-Term Memory, is a type of recurrent neural network (RNN) designed for processing sequential data. It consists of two LSTM layers: one processes the input sequence in the forward direction, while the other processes it in the backward direction. This architecture allows the model to capture information from both past and future contexts, making it particularly effective for tasks such as time series analysis.

Dropout: A Dropout layer is used to prevent overfitting by randomly setting a fraction of the input units to 0 during training. This means that at each training step, a specified percentage of neurons are "dropped out," which helps the model generalize better by not relying too heavily on any single neuron or connection.

Dense layers: A dense layer, also known as a fully connected layer, is a type of layer in neural networks where each neuron is connected to every neuron in the preceding layer. This structure allows for comprehensive interaction among features learned from the previous layer, facilitating complex computations and learning.

CTC Loss: CTC (Connectionist Temporal Classification) loss is a loss function used primarily in sequence-to-sequence tasks where the input and output sequences may have different lengths, such as in speech recognition. It allows for the alignment of unsegmented input sequences (like audio) to target sequences (like text) by summing over the probabilities of all possible alignments between the two.

Callback functions: The callback functions are used to save the model's training checkpoints after a specified number of iterations.

Adam Optimizer: Adam (Adaptive Moment Estimation) is an optimization algorithm for training neural networks. It is a variant of stochastic gradient descent that adjusts the learning rate based on the first and second moments of the gradients, making it efficient for various tasks in deep learning.

HTML: HTML (HyperText Markup Language) is the standard markup language used to create and structure content on the web. It consists of a series of elements represented by tags that define the structure of web pages, such as headings, paragraphs, links, images, and other media. HTML provides the basic building blocks for websites, focusing on content organization rather than styling or functionality, and works in conjunction with CSS and JavaScript to create complete web experiences.

CSS: CSS (Cascading Style Sheets) is a style sheet language used to describe the presentation and visual formatting of HTML documents. It controls layout, colors, fonts, spacing, and responsive design aspects, effectively separating content from presentation. CSS uses selectors to target HTML elements and apply style declarations, supporting features like media queries for responsive design, animations, and transformations, allowing developers to create visually appealing and consistent user interfaces across multiple devices.

JavaScript (JS): JavaScript is a versatile, high-level programming language that enables interactive and dynamic content on websites. As one of the core technologies of web development alongside HTML and CSS, JavaScript runs on the client side within web browsers and increasingly on servers through environments like Node.js. It allows developers to implement complex features such as form validation, animations, data manipulation, asynchronous requests, and full web applications, making static web pages responsive to user interactions without requiring page reloads.

Flask: Flask is a lightweight, minimalist web framework for Python that provides the essential

tools for building web applications while maintaining simplicity and flexibility. Unlike more comprehensive frameworks, Flask follows a "micro-framework" philosophy, offering a solid core with routing, request handling, and templating, while allowing developers to choose which components to add for databases, forms, and authentication. Its simplicity, extensive documentation, and compatibility with various extensions make it particularly popular for small to medium-sized projects, APIs, and prototyping, though it scales effectively for larger applications as well.

Googletrans: Googletrans is a free, unofficial Python library that provides a simple API for accessing Google Translate's neural machine translation service. It allows developers to programmatically translate text between over 100 supported languages, detect the language of a given text, and perform batch translations without requiring an API key. Though not officially supported by Google, it has become popular due to its ease of use and functionality, making it valuable for applications requiring multilingual support, content localization, and natural language processing tasks where translation capabilities are needed.

4.2 Algorithms

LipVision employs a sophisticated multi-stage algorithmic architecture designed to process video inputs and generate accurate transcriptions. At the core of the visual feature extraction process are 3D Convolutional Neural Networks (3D-CNN), which analyze both spatial and temporal dimensions simultaneously across consecutive video frames. Unlike standard 2D CNNs that would process individual frames independently, the 3D-CNN architecture captures motion dynamics across time, essential for distinguishing subtle differences in lip movements. These convolutional layers, activated through Rectified Linear Units (ReLU), extract increasingly abstract features while MaxPooling3D layers reduce dimensionality while preserving critical information.

57 37 Following visual feature extraction, the system employs Bi-directional Long Short-Term Memory (Bi-LSTM) networks, which process the sequence in both forward and backward directions. This bidirectional approach allows the algorithm to understand context from both past and future frames, crucial for accurate sentence-level interpretation. The Bi-LSTM layers are

implemented through Time Distributed wrappers to maintain temporal coherence across the sequence processing. To prevent overfitting, strategic Dropout layers are incorporated throughout the network architecture. The entire training process is optimized using the Connectionist Temporal Classification (CTC) loss function—a specialized algorithm designed for sequence prediction without requiring explicit alignment between input and output sequences—paired with the Adam optimizer for efficient gradient updates.

4.3 Dataset Description

GRID is a large multitalker audiovisual sentence dataset which consists of audio and video (facial) recordings of 1000 sentences spoken by each of 34 talkers (18 male, 16 female). Sentences are of the form "put red at G9 now".

Audio files are available in two formats: the original raw file with 50 kHz signals, and the downsampled files with 25 kHz signals. Video files are also available in two formats: normal quality (360x288; ~1kbit/s) and high quality (720x576; ~6kbit/s).

Due to a technical oversight, video for speaker 21 is not available. The transcriptions (specifying the coordinates and words being spoken) are also available for each audio and video file.

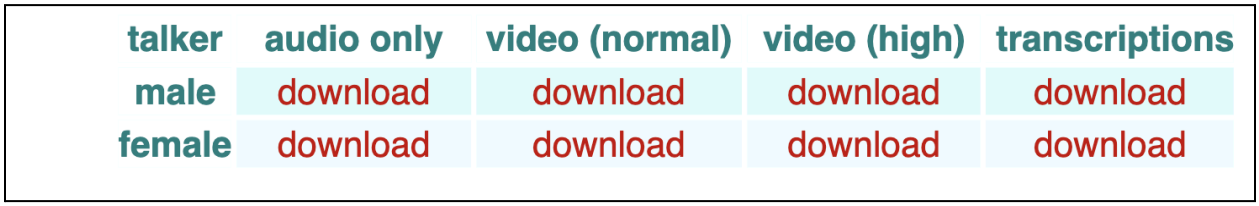

Figure 4.1: An example screenshot of the GRID dataset files.

Figure 4.1 shows a sample of the video and audio files available to download from the GRID dataset. The file options include an audio only file, video files with normal and high qualities, and their corresponding transcriptions. All the files are available for both male and female talkers.

Figure 4.2: Screenshot of a sample video (high resolution) and its corresponding transcription downloaded from the GRID dataset.

Figure 4.2 shows the screenshot of a high resolution video downloaded from the GRID dataset on the left and its corresponding transcription on the right. The transcription contains the words spoken in the video as well as indications of silence in the video using the word “sil”.

4.4 Testing Process

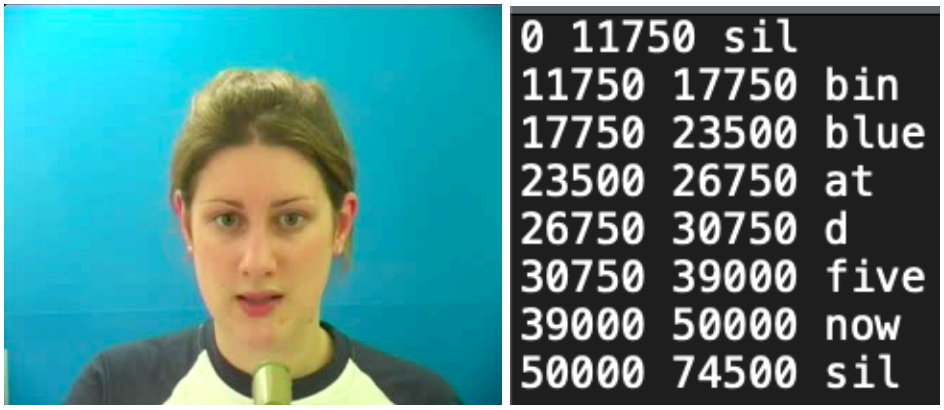

Figure 4.3: Screenshot of a video and its corresponding transcription downloaded from the GRID dataset.

Figure 4.3 shows the screenshot of a high resolution video downloaded from the GRID dataset on the left and its corresponding transcription on the right. The transcription contains the words spoken in the video as well as indications of silence in the video using the word “sil”.

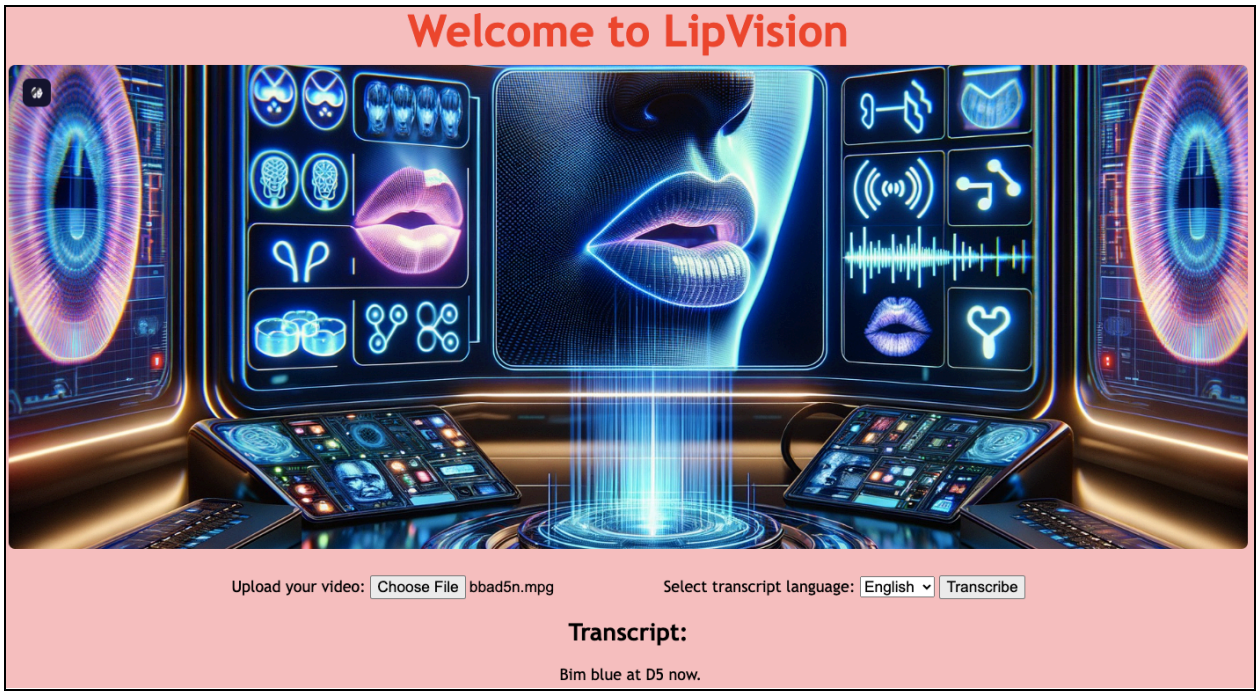

Figure 4.4: Screenshot of the LipVision interface with the generated transcription in English.

Figure 4.4 shows the screenshot of the LipVision interface with the transcript generated in the English language when the video in Figure 4.3 is chosen by the user.

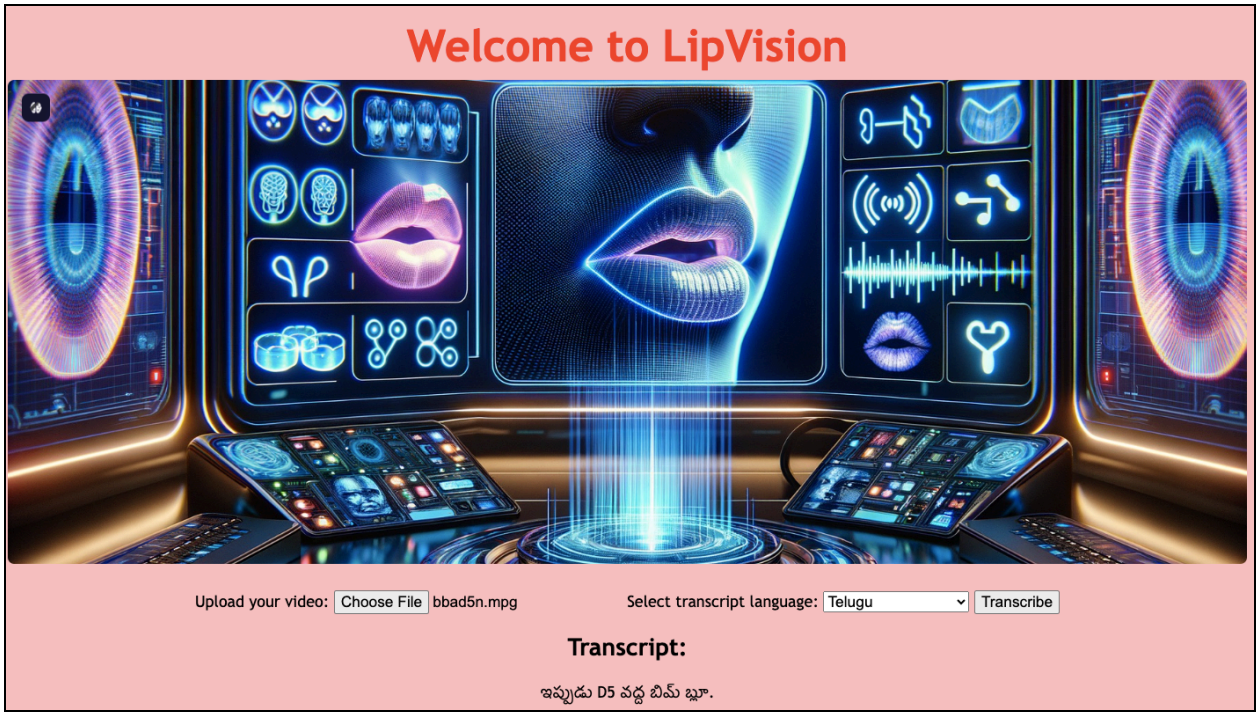

Figure 4.5: Screenshot of the LipVision interface with the generated transcription in Telugu.

Figure 4.5 shows the Telugu language translation of the transcript obtained in Figure 4.4. The transcript language can be chosen from the dropdown menu available on the interface.

Figure 4.6: Screenshot of the custom video.

Figure 4.6 shows the screenshot of a completely new, custom video of a previously unseen speaker.

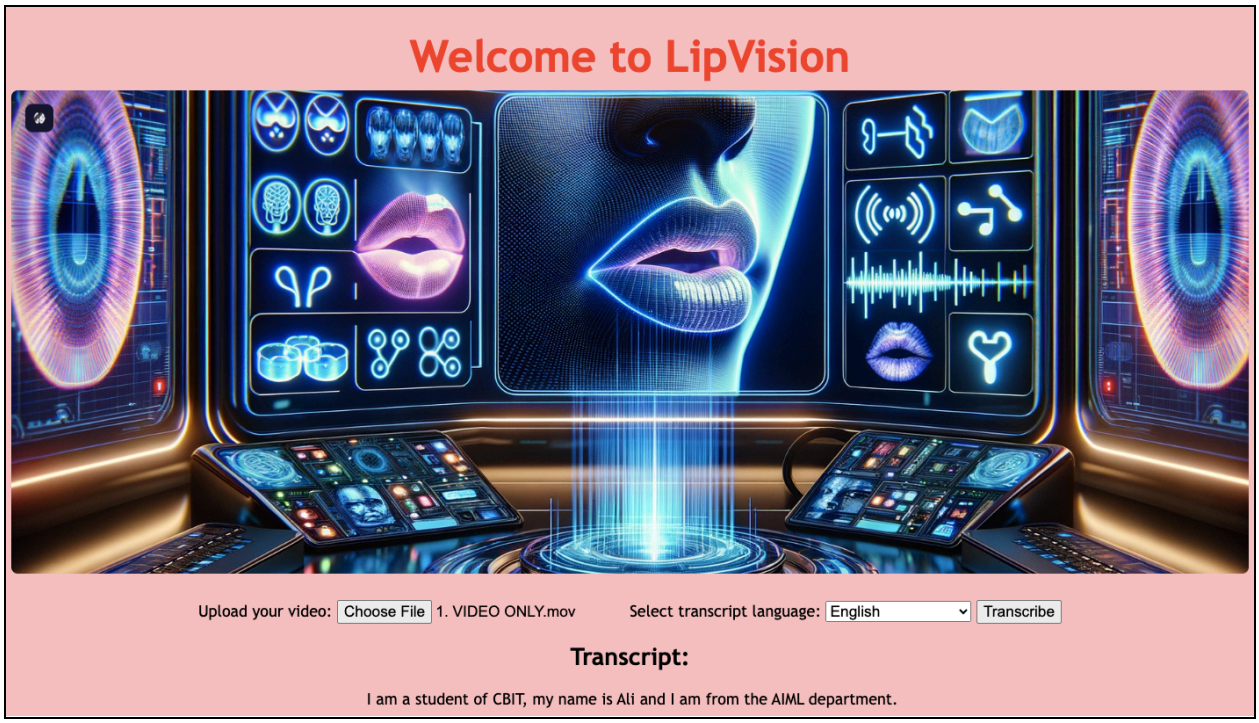

Figure 4.7: Screenshot of the LipVision interface with the generated transcription in English.

Figure 4.7 shows the screenshot of the LipVision interface with the transcript generated in the English language when the video in Figure 4.6 is chosen by the user.

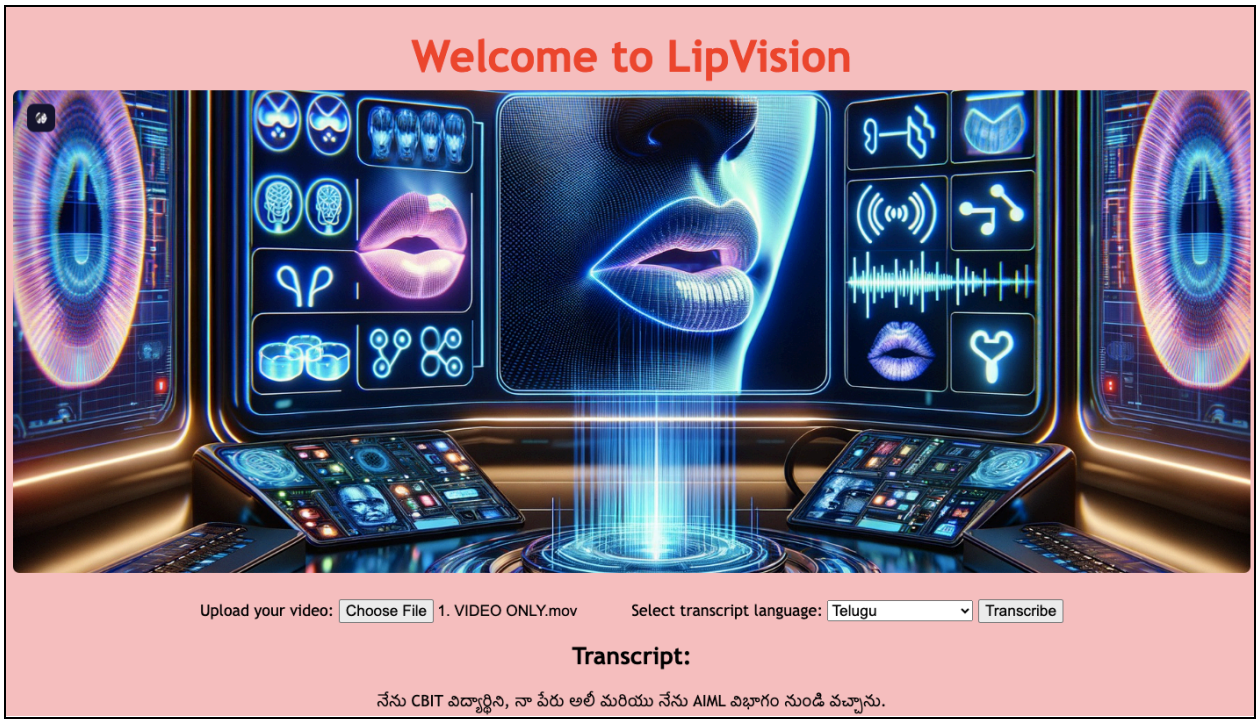

Figure 4.8: Screenshot of the LipVision interface with the generated transcription in Telugu.

Figure 4.8 shows the translation of the transcript obtained in Figure 4.7 in the Telugu language. The user can select their preferred transcript language from the dropdown menu available.

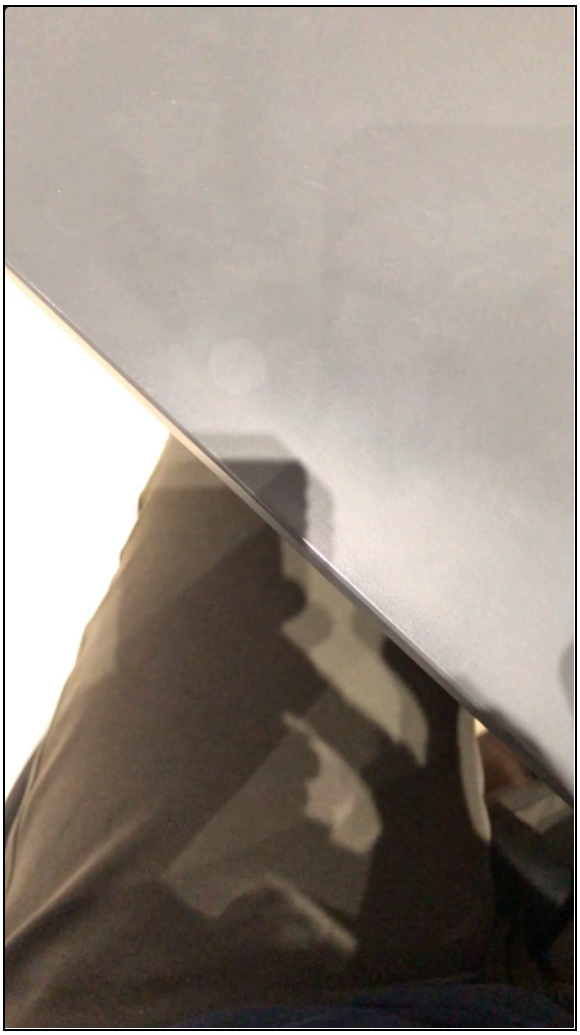

Figure 4.9: Screenshot of an erroneous video with no visible speaker.

Figure 4.9 shows the screenshot of an erroneous video, that is, a video with no visible speaker, created for error handling purposes.

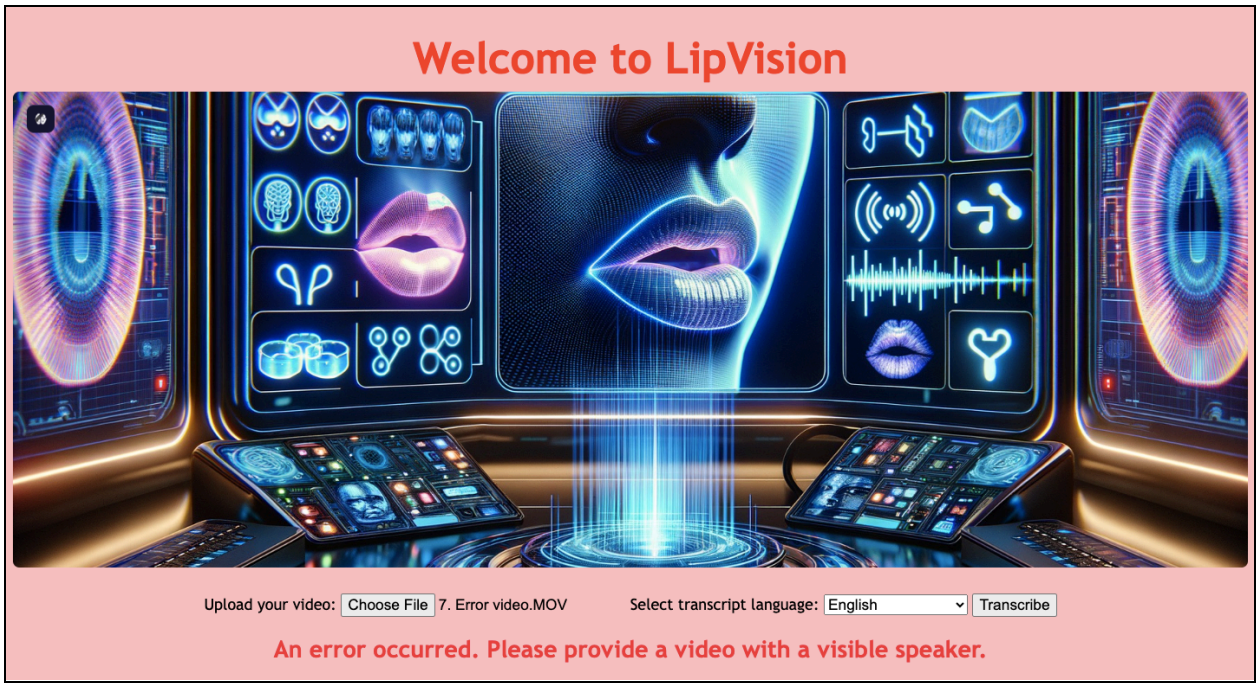

Figure 4.10: Screenshot of the LipVision interface displaying an error message in response to giving an erroneous video as input.

Figure 4.10 shows the LipVision interface with an error message generated in response to the user uploading the erroneous video in Figure 4.9.

5. RESULTS AND DISCUSSIONS

Based on the formula of accuracy as defined in chapter 2.1, the accuracy of LipVision in generating accurate transcriptions is measured on 34,000 videos from the GRID dataset, 50 custom videos, as well as 50 erroneous videos.

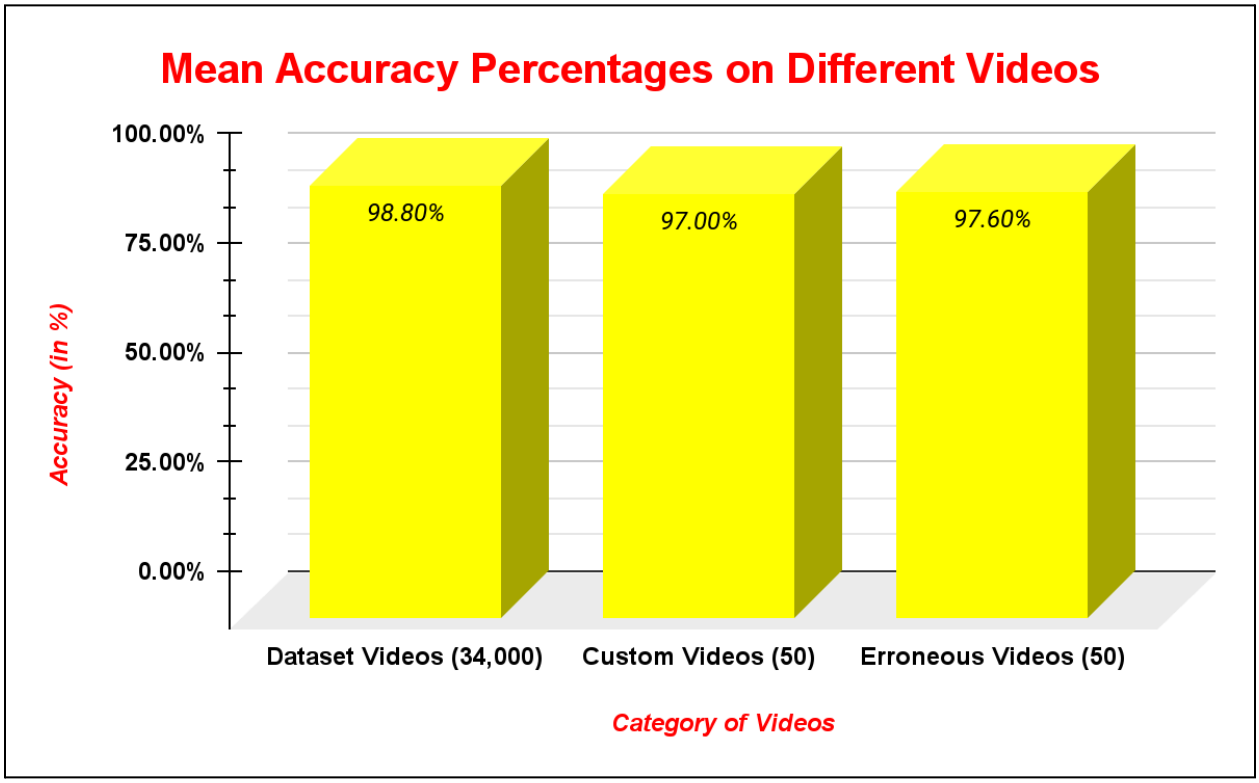

Figure 5.1: Bar graph showing the mean accuracy percentages obtained after testing LipVision on the GRID dataset videos, custom videos, and erroneous videos.

Figure 5.1 shows the mean accuracy percentages of LipVision on dataset, custom, and erroneous videos. LipVision is observed to perform the best on the GRID dataset with an accuracy of 98.80%, a significant improvement over the previous model's 96.7% accuracy on the same dataset [10]. LipVision can distinguish between erroneous videos with no speakers and valid videos with a clear visible speaker with 97.60% accuracy. On custom videos, LipVision is able to generate the correct transcripts with an impressive accuracy of 97%. Therefore, the final average accuracy of the LipVision model in generating correct transcripts and differentiating between valid and invalid input videos is estimated to be 97.8%.

6. CONCLUSIONS

6.1 Conclusions

This project successfully developed "LipVision," a deep-learning web application designed for accurate video lip-reading and translation. By integrating a robust architecture comprising 3D Convolutional Neural Networks (3D-CNN) and Bi-directional Long Short-Term Memory (Bi-LSTM) networks, trained rigorously on the GRID Audiovisual Sentence Corpus using TensorFlow, the system effectively interprets lip movements. The project achieved a notable average accuracy of 97.8% across dataset, custom, and erroneous video inputs, demonstrating proficiency in converting visual lip data into coherent English sentences and distinguishing valid inputs from those lacking a visible speaker.

The implementation extends beyond mere transcription, incorporating a user-friendly web interface built with Flask, HTML, CSS, and JavaScript, and leveraging Googletrans for multilingual translation capabilities. Testing confirmed LipVision's ability to handle both controlled dataset videos and novel custom videos with high fidelity, showcasing its potential for real-world applications. These applications span enhancing communication accessibility for the hearing-impaired, providing tools for silent communication in security contexts, and aiding content creators with automated subtitling, thereby addressing significant societal and industry needs.

6.2 Future Scope

While LipVision demonstrates substantial success, future enhancements could further expand its capabilities. One significant area for advancement is enabling the system to process videos featuring multiple speakers. Future iterations could incorporate speaker diarization techniques to identify different individuals in a video and generate separate, speaker-identified transcripts. This would expand LipVision's utility for applications such as meeting transcription, movie subtitling, and complex interview documentation, where distinguishing between speakers is essential for meaningful content interpretation.

Another crucial enhancement would involve training the model on a more diverse, multilingual

dataset. While LipVision currently processes videos with English speakers and offers translation capabilities, training the underlying model on multilingual lip movements would allow it to directly interpret and transcribe speech from videos in languages beyond English.

7. CONTRIBUTIONS OF THE WORK

7.1. Paper-1

CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY
An Autonomous Institute | Affiliated to Osmania University
Kokapet Village, Gandipet Mandal, Hyderabad, Telangana-500075, www.cbit.ac.in

COMMITTED TO RESEARCH, INNOVATION AND EDUCATION 46 years

BENTHAM SCIENCE

Certificate of Presentation

This is to certify that Mr./Ms./Dr. **Ali Hasan** affiliated with **Chaitanya Bharathi Institute of Technology** has presented the paper titled **"Study on Various Techniques of Video Lip Reading Using Deep Learning"** at the 1st International conference on Innovative Computing Technologies (ICICT-2024) held during 13-14 December 2024 organized at Chaitanya Bharathi Institute of Technology, Hyderabad-India, by the Departments of CSE, IT, AI&ML, AI&DS, MCA and Mathematics.

Prof. Prabhakar Kandukuri
Conference - Coordinator
Dept. of AI&ML

Prof. Y. Ramadevi
Conference - General Chair
and Head - Dept. of AI&ML

Prof. C. V. Narasimhulu
Principal