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

Approved by | Affiliated to | UGC Autonomous | 10 Programs Accredited by | Grade A++ in | All India Ranking 151-200 Band

COMMITTED TO RESEARCH, INNOVATION AND EDUCATION
46 years

Department of Artificial Intelligence and Machine Learning

Capstone Project-II (Part 2) - 20AMC18

Project Title:

“LipVision: Video Lip-Reading Using Deep-Learning”

By: PID-12

160121729020: Ali Hasan

160121729053: Sadwale Shivaji

Guided By:

Ms. Ramya Thavva, M. Tech (CSE)

Assistant Professor, Dept. of AIML

LIST OF CONTENTS

- Abstract
- Problem Statement
- Literature Survey
- Objectives
- Real-World Applications
- Dataset Used
- Architecture & Modules
- Outputs
- Conclusion
- Literature Paper Status
- Final Paper Status
- References

ABSTRACT

- LipVision is a deep-learning web application designed to accurately interpret and translate spoken sentences by analyzing lip movements in videos.

- The system leverages advanced frameworks including 3D-Convolutional Neural Networks, TensorFlow, OpenCV, NumPy, and Bi-LSTM to deliver robust video lip-reading capabilities.

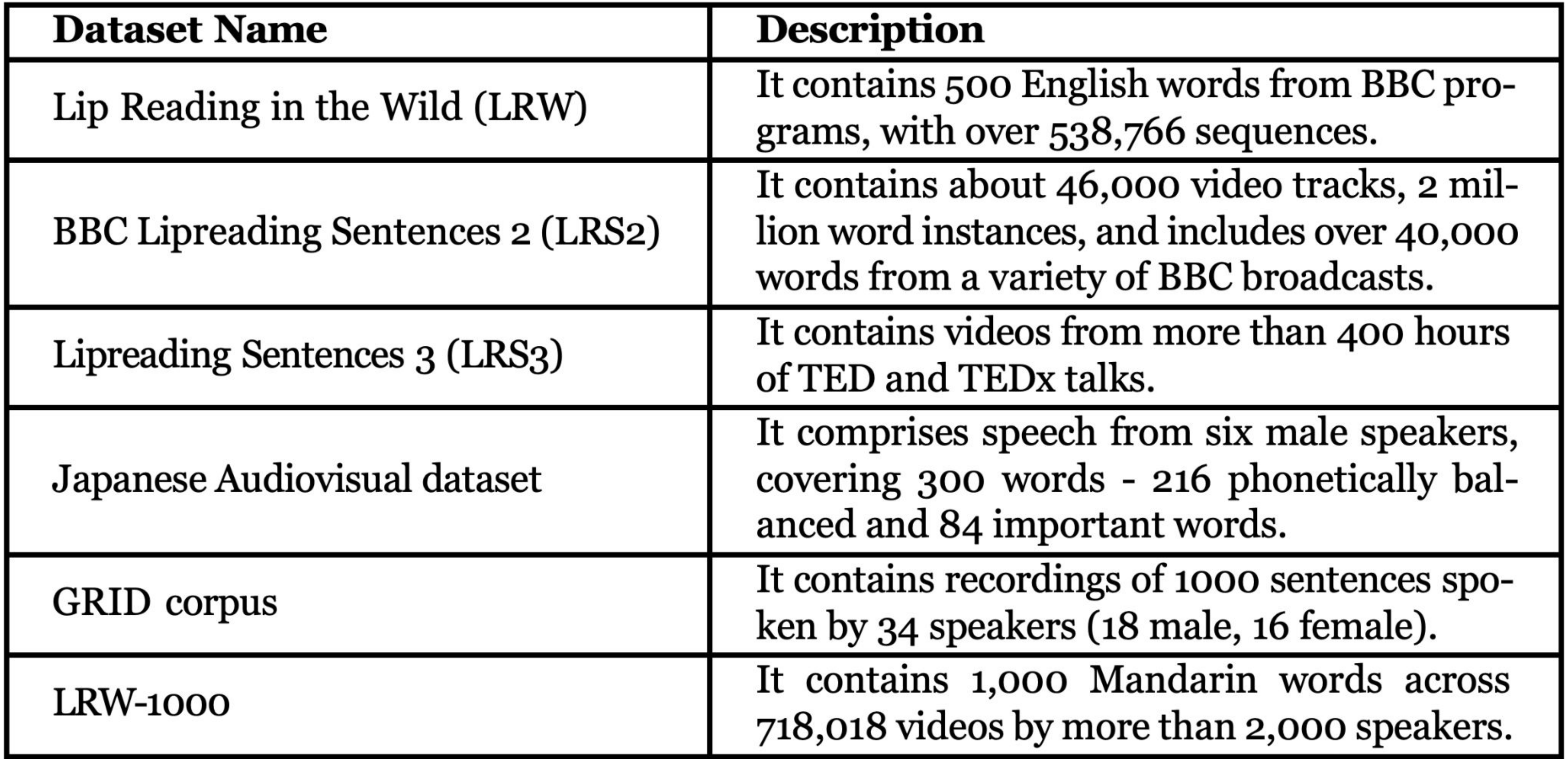
- Built using HTML, CSS, JavaScript, Flask, and Googletrans, LipVision was rigorously trained and evaluated on the GRID Audiovisual Sentence Corpus, demonstrating high accuracy in deciphering spoken sentences.

- The technology shows significant potential for applications in defense and security sectors where silent communication is crucial, and for assisting individuals with hearing impairments by providing alternative means of understanding spoken language.

PROBLEM STATEMENT

The inadequacy of existing communication aids for silent communication among the deaf and hard-of-hearing necessitates innovative solutions like "LipVision," which leverages deep learning and computer vision to interpret lip movements into sentences, thereby enhancing communication accessibility.

Literature Survey



Dataset Name	Description
Lip Reading in the Wild (LRW)	It contains 500 English words from BBC programs, with over 538,766 sequences.
BBC Lipreading Sentences 2 (LRS2)	It contains about 46,000 video tracks, 2 million word instances, and includes over 40,000 words from a variety of BBC broadcasts.
Lipreading Sentences 3 (LRS3)	It contains videos from more than 400 hours of TED and TEDx talks.
Japanese Audiovisual dataset	It comprises speech from six male speakers, covering 300 words - 216 phonetically balanced and 84 important words.
GRID corpus	It contains recordings of 1000 sentences spoken by 34 speakers (18 male, 16 female).
LRW-1000	It contains 1,000 Mandarin words across 718,018 videos by more than 2,000 speakers.

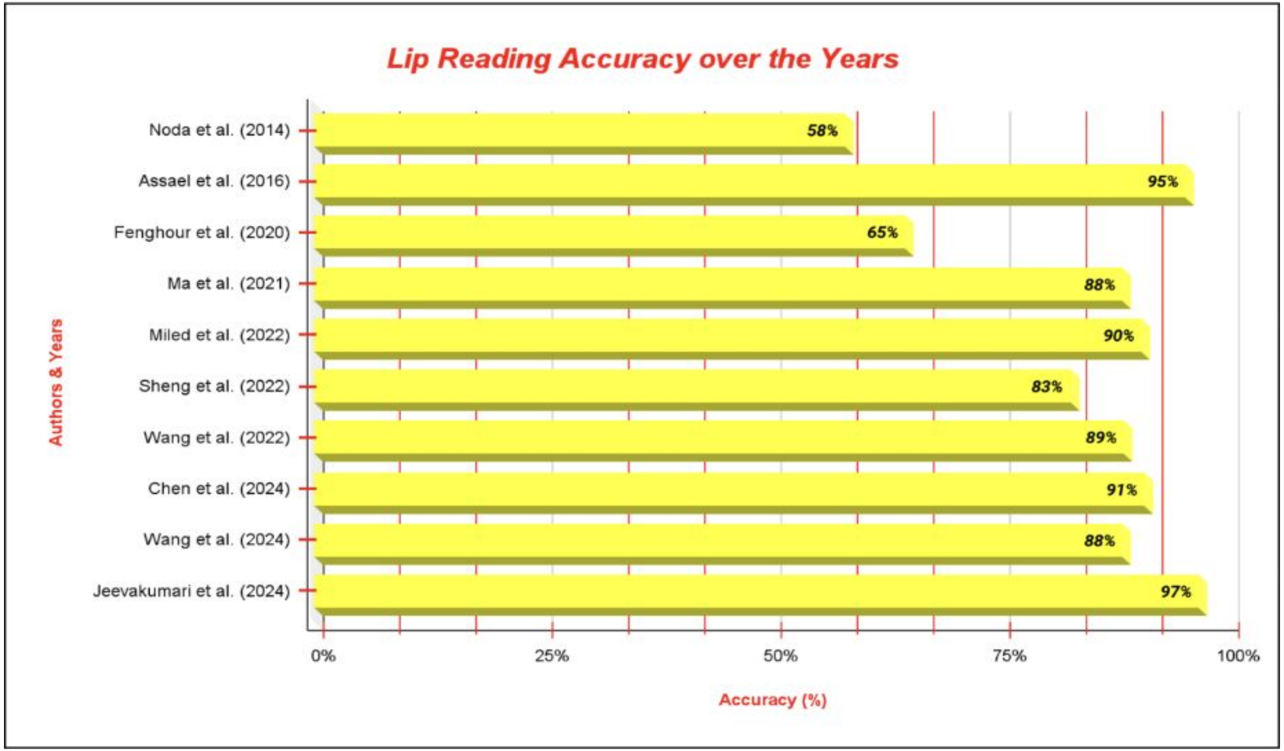
Descriptions of the most common datasets used in lip reading architectures.

Authors	Approach	Strengths	Limitations	Performance
Noda, Ku- niaki, et al. (2014)	Visual Speech Recognition (VSR) system with CNN for extracting features from mouth area images.	Application of deep learning with reasonable phoneme recognition and robustness against image variances.	Use of a small dataset affecting generalizability, reliance on speaker-dependent models, limited consonant recognition.	Achieved a phoneme recognition rate of 58% and word recognition rate of 37%.
Assael, Yannis M., et al. (2016)	Complete sentence level lip reading using STCNNs, Bi-GRUs, and CTC loss.	High accuracy on the GRID dataset, surpassing human lip reading performance and former models.	Use of a limited dataset, uncertain performance in uncontrolled, noisy environments.	The model achieved 6.4% CER and 11.4% WER for unseen speakers and 1.9% CER and 4.8% WER for overlapped speakers.
Fenghour, Souheil, et al. (2020)	Predicting spoken sentences from silent videos using 3D convolutional network and 2D ResNet.	Significant decrease in error rate and increase in accuracy from previous models, no need for audio input.	Significant gap between Viseme and Word error rates, indicating problems in converting viseme to words.	Decreased the WER to 35.4% compared to the past models.
Ma, Pingchuan, et al. (2021)	DC-TCN for lip reading of words that are isolated using 3D convolutional layers and 2D ResNet-18 for refinement.	High accuracy on LRW-1000 and LRW datasets, surpassing previous models.	Increased computational complexity, scope for improving accuracy on the LRW-1000 dataset.	Achieved 43.65% and 88.36% accuracies on the LRW-1000 and the LRW dataset respectively.

Ma, Pingchuan, et al. (2021)	DC-TCN for lip reading of words that are isolated using 3D convolutional layers and 2D ResNet-18 for refinement.	High accuracy on LRW-1000 and LRW datasets, surpassing previous models.	Increased computational complexity, scope for improving accuracy on the LRW-1000 dataset.	Achieved 43.65% and 88.36% accuracies on the LRW-1000 and the LRW dataset respectively.
Miled, Malek, Mohammed Anouar Ben Messaoud, and Aicha Bouzid. (2022)	Lip segmentation and Lip reading using Haar Cascade classifier (for segmentation) and CNN and BiGRU (for lip reading).	Development of an accurate and robust lip reading system that uses lip segmentation and deep learning to achieve high accuracy on the LRW dataset.	Potential challenges in noisy environments and reliance on a specific dataset, affecting generalizability.	Achieved an accuracy of 90.38%, OL of 91%, and SE of 7.8%.
Sheng, Chang- chong, et al. (2022)	Adaptive Semantic-Spatio-Temporal Graph Convolutional Network (ASST-GCN), Transformers, and Convolutional Networks.	Enhancement of dynamic and automatic lip reading using novel frameworks, high experimental validation.	Over-reliance on precise facial landmark detection leads to challenges in complicated real-world scenarios.	Achieved an accuracy of 82.6% on the LRW dataset.

Wang, Huijuan, Gangqiang Pu, and Tingyu Chen. (2022)	3D Convolutional Vision Transformer and BiGRU (bidirectional gated recurrent unit).	Advancement of lip-reading technology using Transformers, effective feature extraction, comprehensive evaluation on LRW dataset.	Complexity of the model leading to high computational demands, and limiting real-world applicability.	Achieved 57.5% and 88.5% accuracies on the LRW-1000 and the LRW dataset respectively.
Chen, Hang, et al. (2024)	VSM (Hybrid and Collaborative) and End to End Modeling.	Integration of temporal mask module in CVSEM, filtering noise and improving model's decision making.	Computational complexity and reliance on large dataset limiting real-world applicability.	Achieved 58.89% and 90.75% accuracies on the LRW-1000 and the LRW dataset respectively.
Wang, Huijuan, et al. (2024)	3D Convolution and Visual Transformer techniques such as BiGRU.	Innovative use of weight sharing and distillation techniques, enhancing model efficiency without sacrificing performance.	Challenges in accurately capturing subtle lip movements and potential performance degradation during compressing of models.	Achieved 57.1% and 88.3% accuracies on the LRW-1000 and the LRW dataset respectively.

Jeevakumari, SAAmutha, and Koushik Dey. (2024)	3D Convolutional Neural Networks (3D CNN), Bi-Directional Long Short- Term Memory (BiLSTM), and Connectionist Temporal Classification (CTC) loss.	Comprehensive dataset utilization, successful integration of temporal and spatial data processing.	Reliance on a specific dataset affecting generalizability, need for larger datasets and alternative architectures to enhance real-world performance.	Achieved an accuracy of 96.7% and a WER of 8.2%.



OBJECTIVES

1. Developing a better foundational deep learning model for lip-reading on the diverse GRID Audiovisual Sentence dataset using TensorFlow, 3DCNN, Bi-LSTM, and OpenCV.

2. Extending the current technology of word-level lip-reading to sentence-level lip-reading, and sentence-level translation of the transcript into different languages.

3. Increasing the robustness of the model by rigorously training the model to generate accurate predictions of the sentences spoken in the dataset videos as well as in custom videos, and

4. Developing an interactive web application using Python frameworks to increase the accessibility of the technology for the general public.

REAL-WORLD APPLICATIONS

- **Court Proceedings:** Accurate court records are essential to the legal system but creating them is labor-intensive. LipVision can supplement traditional court reporting with automated transcription of proceedings, ensuring comprehensive documentation.
- **Law Enforcement:** Police interviews and witness statements must be meticulously documented for evidentiary purposes. LipVision provides an additional layer of documentation by capturing spoken testimony directly from video recordings, reducing potential discrepancies in written reports.
- **Surveillance:** Security agencies sometimes need to understand conversations occurring at a distance without audio. LipVision can analyze surveillance footage to transcribe conversations from lip movements alone, providing intelligence that would otherwise be inaccessible.
- **Silent Communication:** In scenarios where speaking aloud is impossible or dangerous, such as military operations or emergency situations, lip reading becomes critical. LipVision enables silent communication to be accurately captured and transmitted as text, potentially saving lives in critical situations.
- **Video Content Accessibility:** Many online videos lack proper subtitles, creating barriers for viewers with hearing impairments. LipVision can automatically generate accurate captions for videos across platforms, making digital content more accessible to millions of users worldwide.

- **Public Space Announcements:** Important announcements in airports, train stations, and other public venues often go unheard by those with hearing difficulties. LipVision can convert these announcements to text on digital displays, ensuring critical information reaches everyone regardless of hearing ability.
- **Content Creation:** Creating subtitles for videos traditionally requires significant manual effort and expense. LipVision streamlines this process with automated, accurate transcription that reduces production costs and turnaround time for creators of all sizes.
- **Multilingual Individuals:** People navigating multilingual environments frequently encounter language barriers that hinder effective communication. LipVision bridges these gaps by not only transcribing speech but also offering translation capabilities, enabling cross-language understanding.
- **Remote Education:** As online education becomes increasingly common, LipVision ensures all students can access lecture content through accurate transcription.
- **Linguistic Research:** Researchers studying language patterns need extensive samples of natural speech with accurate transcription. LipVision facilitates this research by automatically processing video interviews and conversations, allowing linguists to analyze larger datasets than would be feasible with manual transcription.

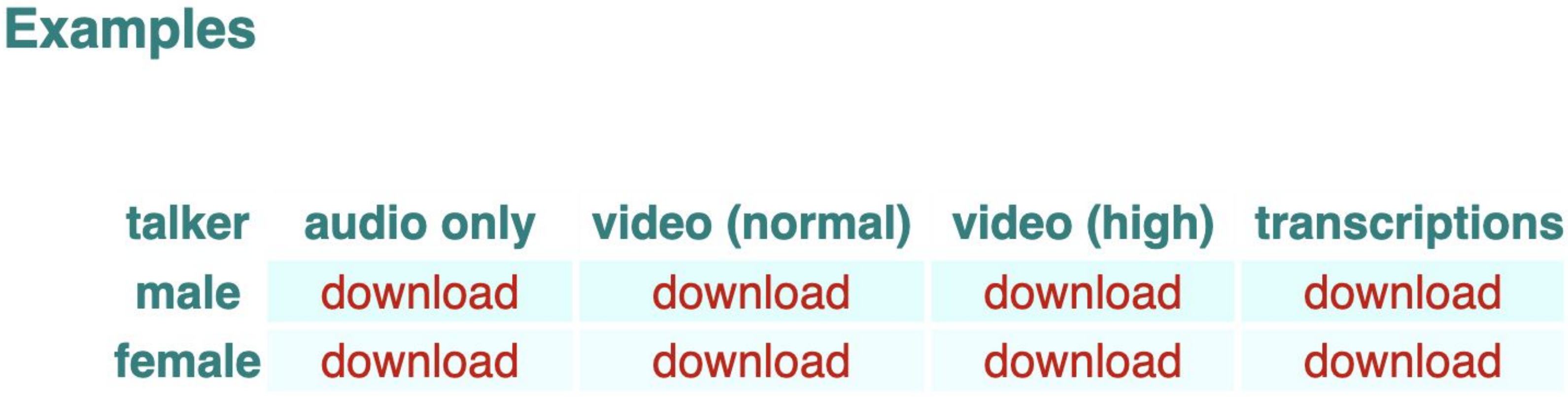
DATASET USED

- 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.

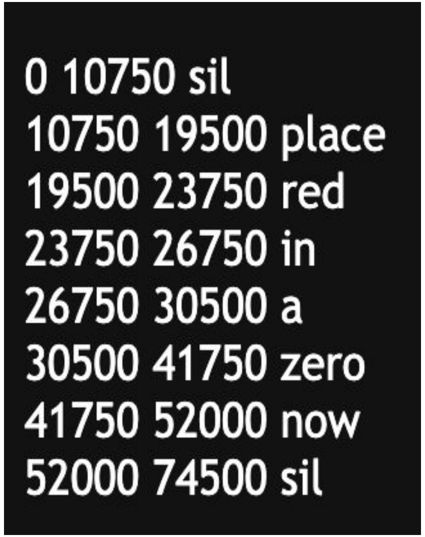


Examples

talker	audio only	video (normal)	video (high)	transcriptions
male	download	download	download	download
female	download	download	download	download

Video (normal)

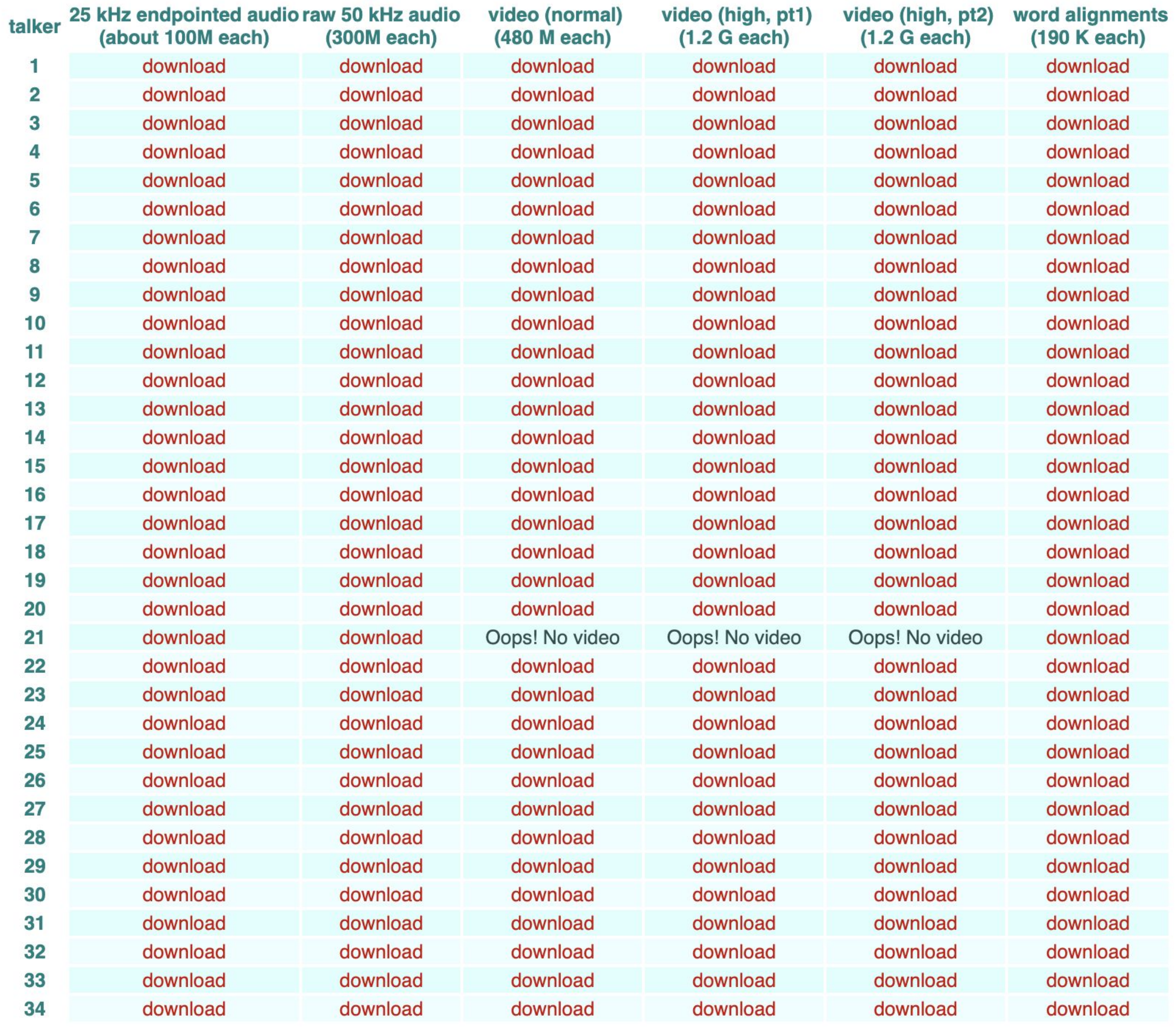
Video (high)



```
0 10750 sil
10750 19500 place
19500 23750 red
23750 26750 in
26750 30500 a
30500 41750 zero
41750 52000 now
52000 74500 sil
```

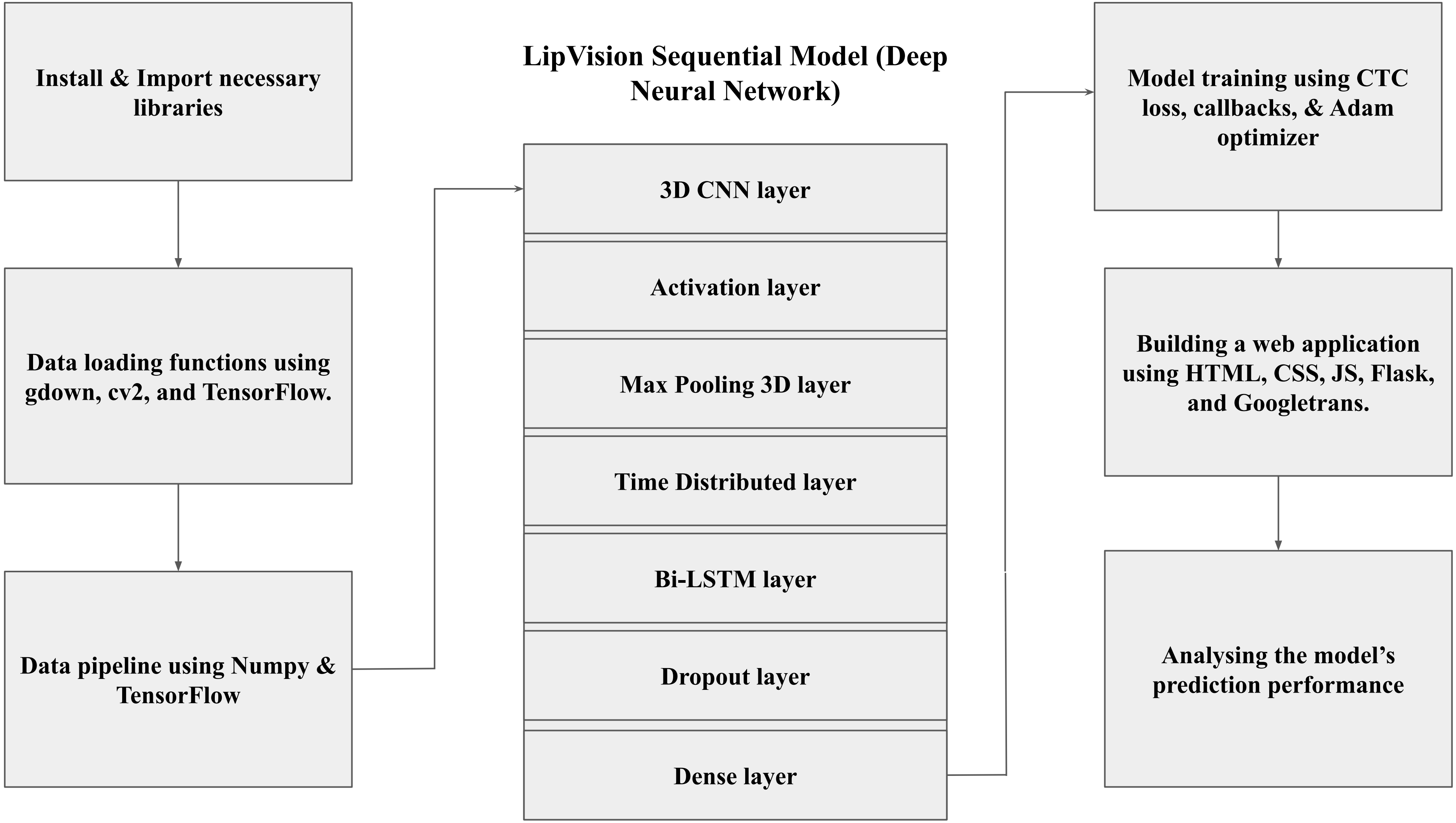
Transcription

Audio



talker	25 kHz endpointed audio (about 100M each)	raw 50 kHz audio (300M each)	video (normal) (480 M each)	video (high, pt1) (1.2 G each)	video (high, pt2) (1.2 G each)	word alignments (190 K each)
1	download	download	download	download	download	download
2	download	download	download	download	download	download
3	download	download	download	download	download	download
4	download	download	download	download	download	download
5	download	download	download	download	download	download
6	download	download	download	download	download	download
7	download	download	download	download	download	download
8	download	download	download	download	download	download
9	download	download	download	download	download	download
10	download	download	download	download	download	download
11	download	download	download	download	download	download
12	download	download	download	download	download	download
13	download	download	download	download	download	download
14	download	download	download	download	download	download
15	download	download	download	download	download	download
16	download	download	download	download	download	download
17	download	download	download	download	download	download
18	download	download	download	download	download	download
19	download	download	download	download	download	download
20	download	download	download	download	download	download
21	download	download	Oops! No video	Oops! No video	Oops! No video	download
22	download	download	download	download	download	download
23	download	download	download	download	download	download
24	download	download	download	download	download	download
25	download	download	download	download	download	download
26	download	download	download	download	download	download
27	download	download	download	download	download	download
28	download	download	download	download	download	download
29	download	download	download	download	download	download
30	download	download	download	download	download	download
31	download	download	download	download	download	download
32	download	download	download	download	download	download
33	download	download	download	download	download	download
34	download	download	download	download	download	download

Architecture & Modules



Install & Import necessary libraries

Data loading functions using gdown, cv2, and TensorFlow.

Data pipeline using Numpy & TensorFlow

LipVision Sequential Model (Deep Neural Network)

3D CNN layer

Activation layer

Max Pooling 3D layer

Time Distributed layer

Bi-LSTM layer

Dropout layer

Dense layer

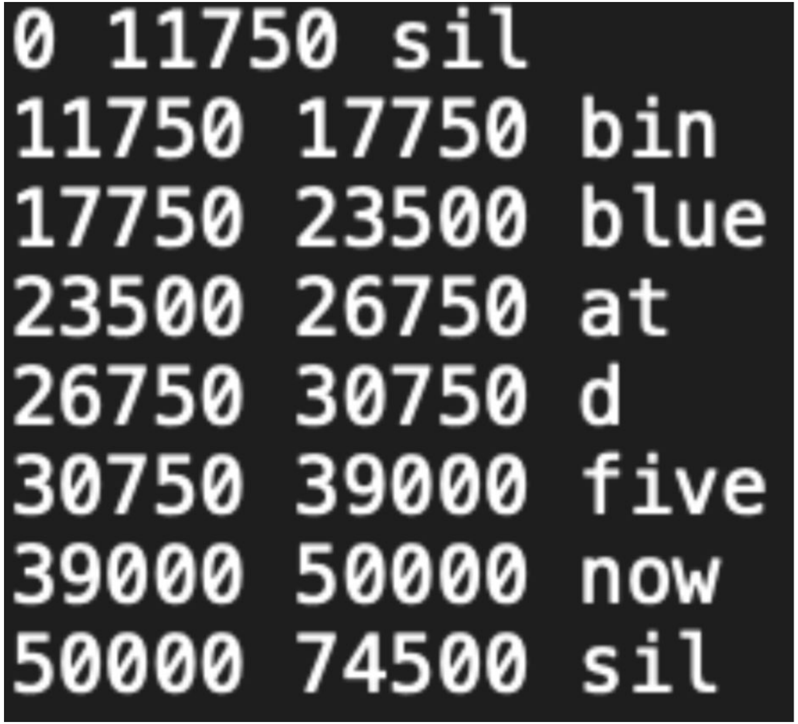
Model training using CTC loss, callbacks, & Adam optimizer

Building a web application using HTML, CSS, JS, Flask, and Googletrans.

Analysing the model's prediction performance

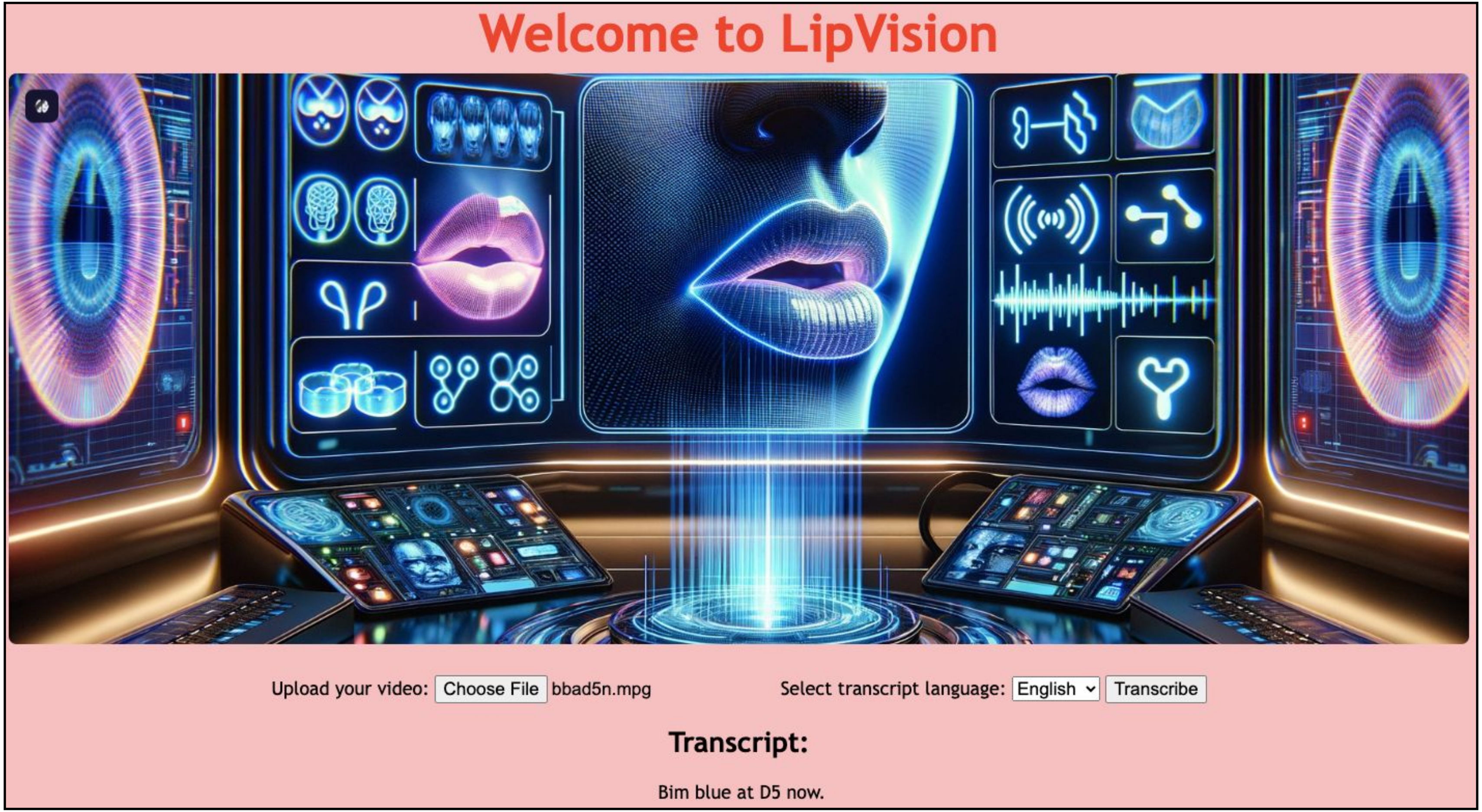
Outputs

Testing on Dataset Video



```
0 11750 sil
11750 17750 bin
17750 23500 blue
23500 26750 at
26750 30750 d
30750 39000 five
39000 50000 now
50000 74500 sil
```

A video and its corresponding transcription downloaded from the GRID dataset.



Welcome to LipVision

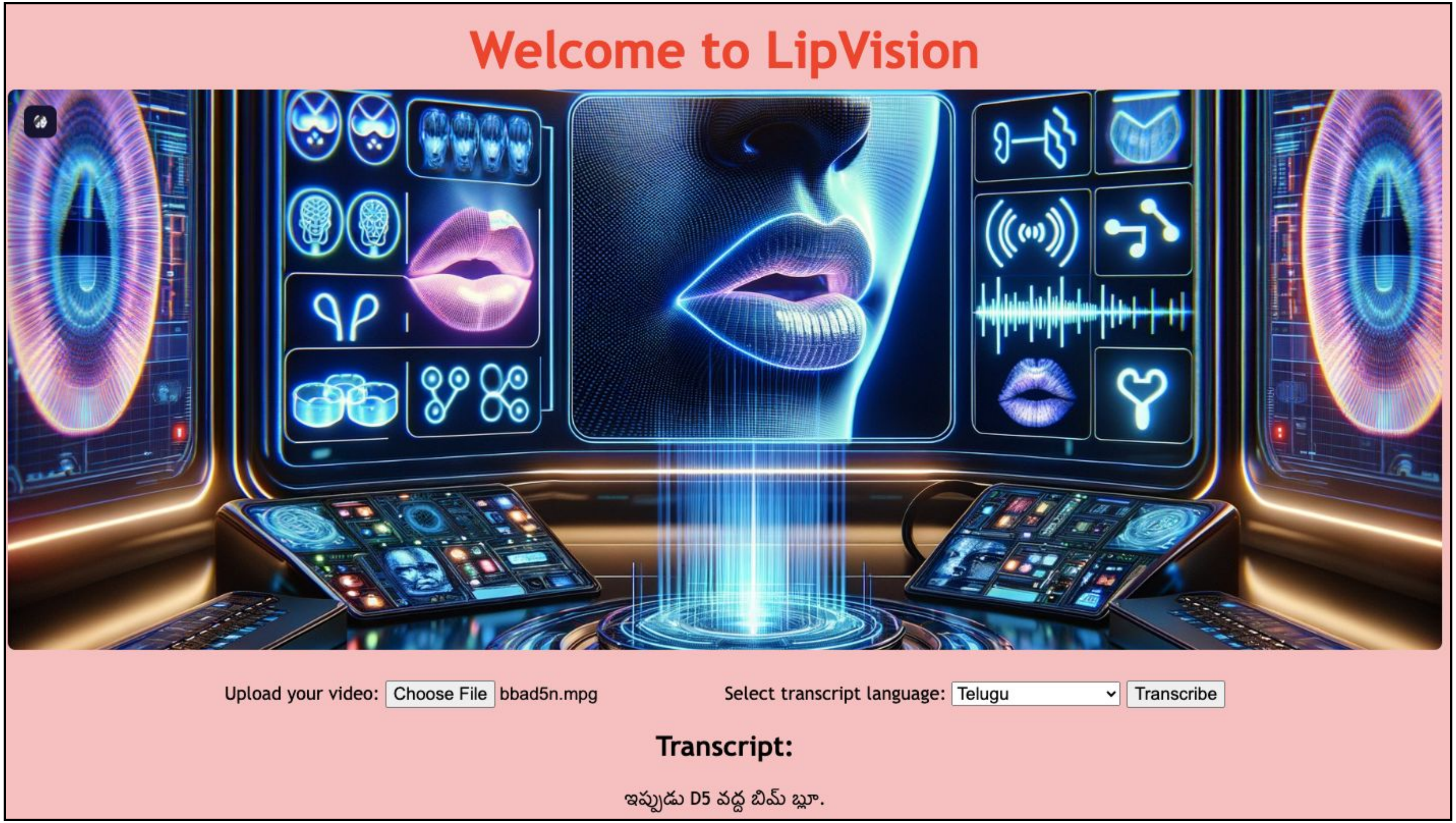
Upload your video: Choose File bbad5n.mpg

Select transcript language: English

Transcript:

Bim blue at D5 now.

LipVision interface with the generated transcription in English.



Welcome to LipVision

Upload your video: Choose File bbad5n.mpg

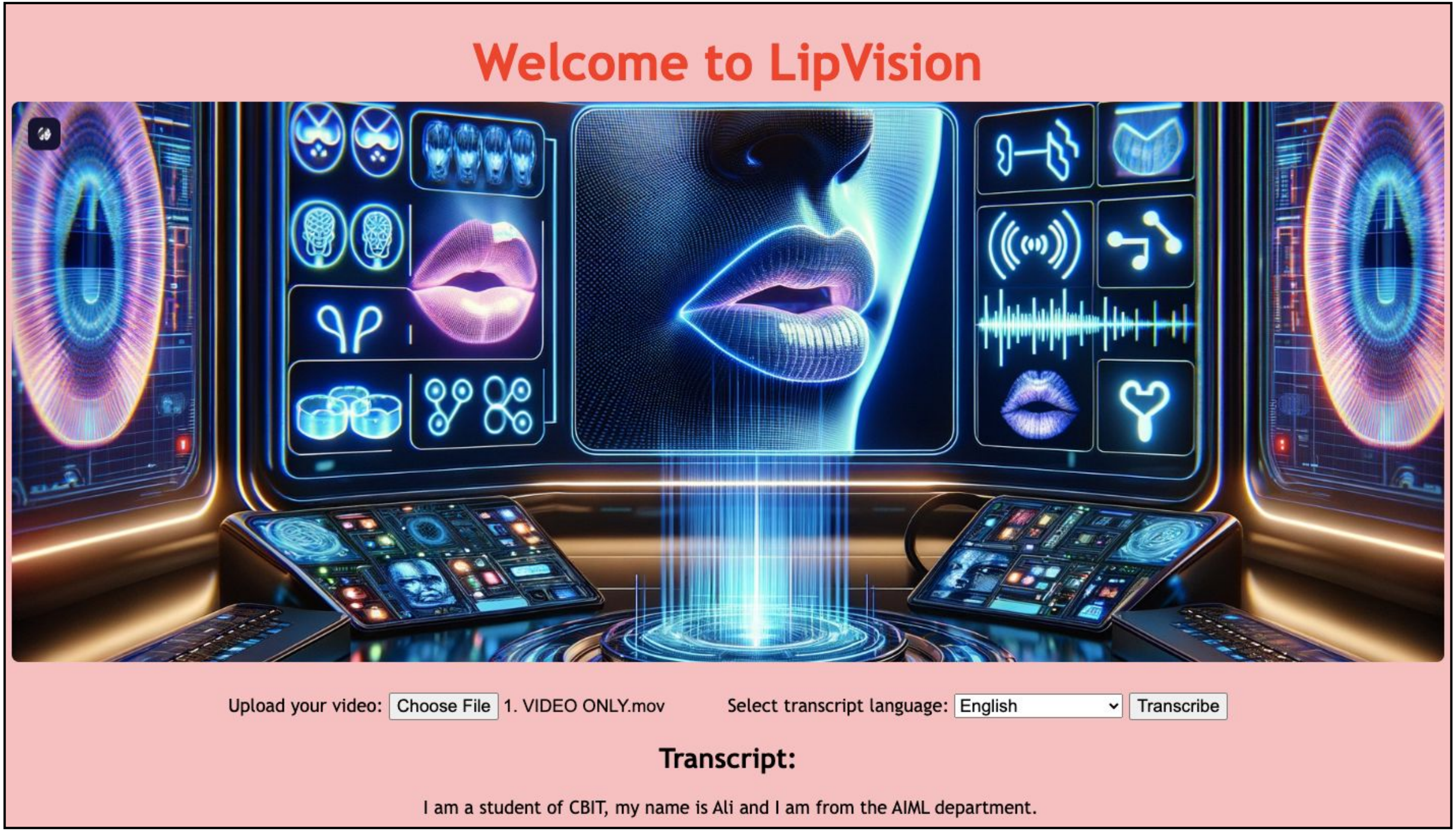
Select transcript language: Telugu

Transcript:

ఇప్పుడు D5 వద్ద బిమ్ బ్లూ.

LipVision interface with the generated transcription in Telugu.

Testing on Custom Video



Welcome to LipVision

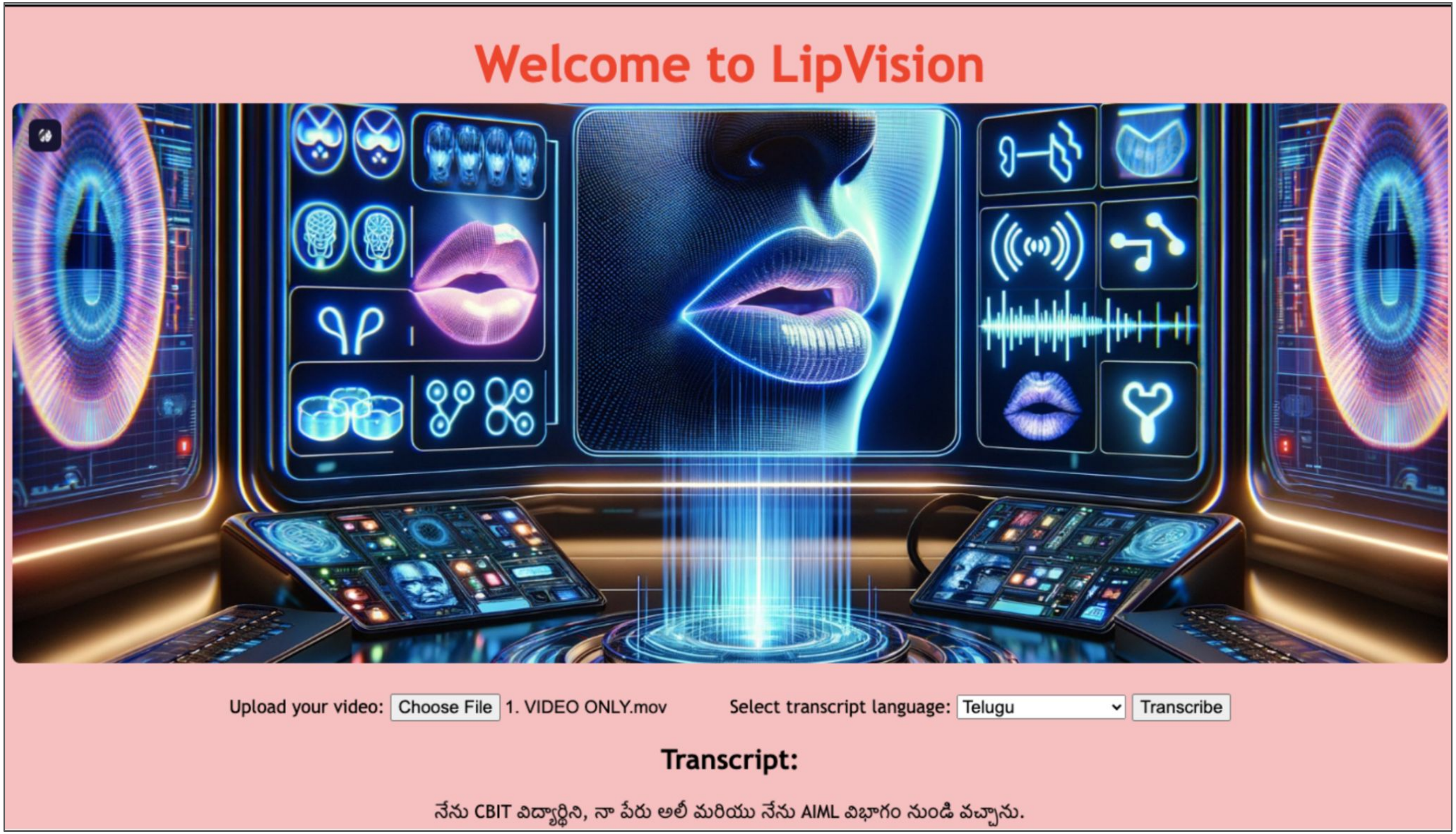
Upload your video: 1. VIDEO ONLY.mov

Select transcript language: English

Transcript:

I am a student of CBIT, my name is Ali and I am from the AIML department.

LipVision interface with the generated transcription in English.



Welcome to LipVision

Upload your video: Choose File 1. VIDEO ONLY.mov

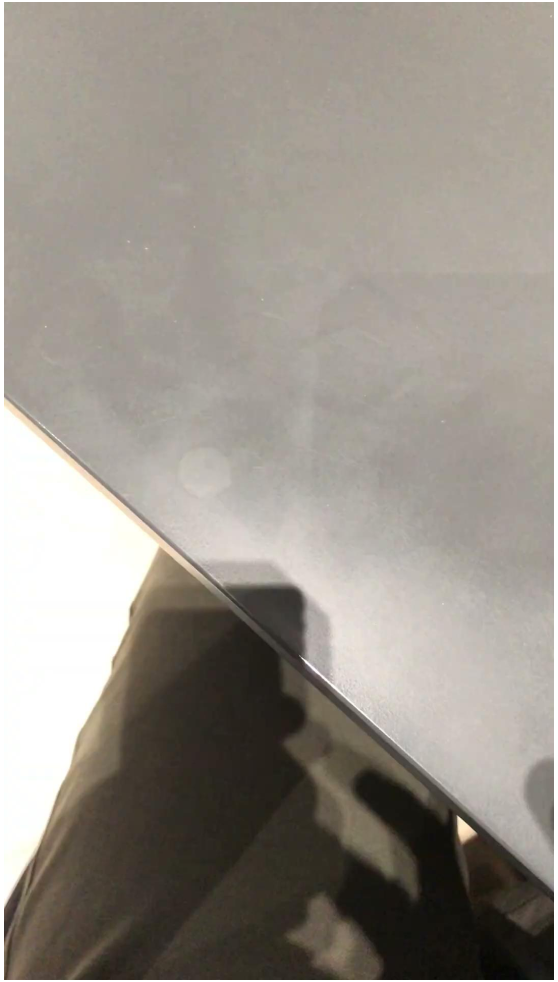
Select transcript language: Telugu

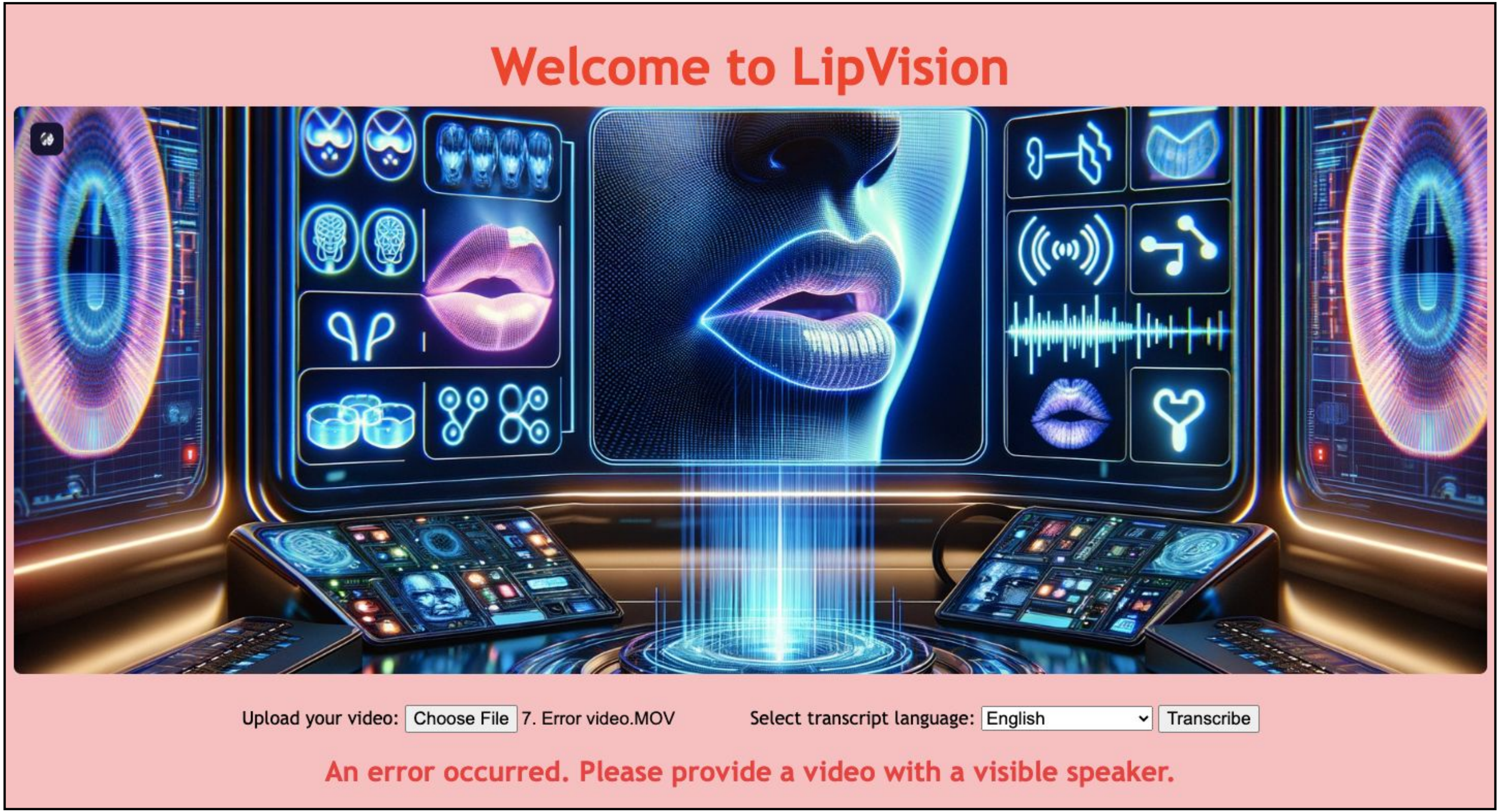
Transcript:

నేను CBIT విద్యార్థిని, నా పేరు అలీ మరియు నేను AIML విభాగం నుండి వచ్చాను.

LipVision interface with the generated transcription in Telugu.

Testing on Erroneous Video





Welcome to LipVision

Upload your video: Choose File 7. Error video.MOV

Select transcript language: English

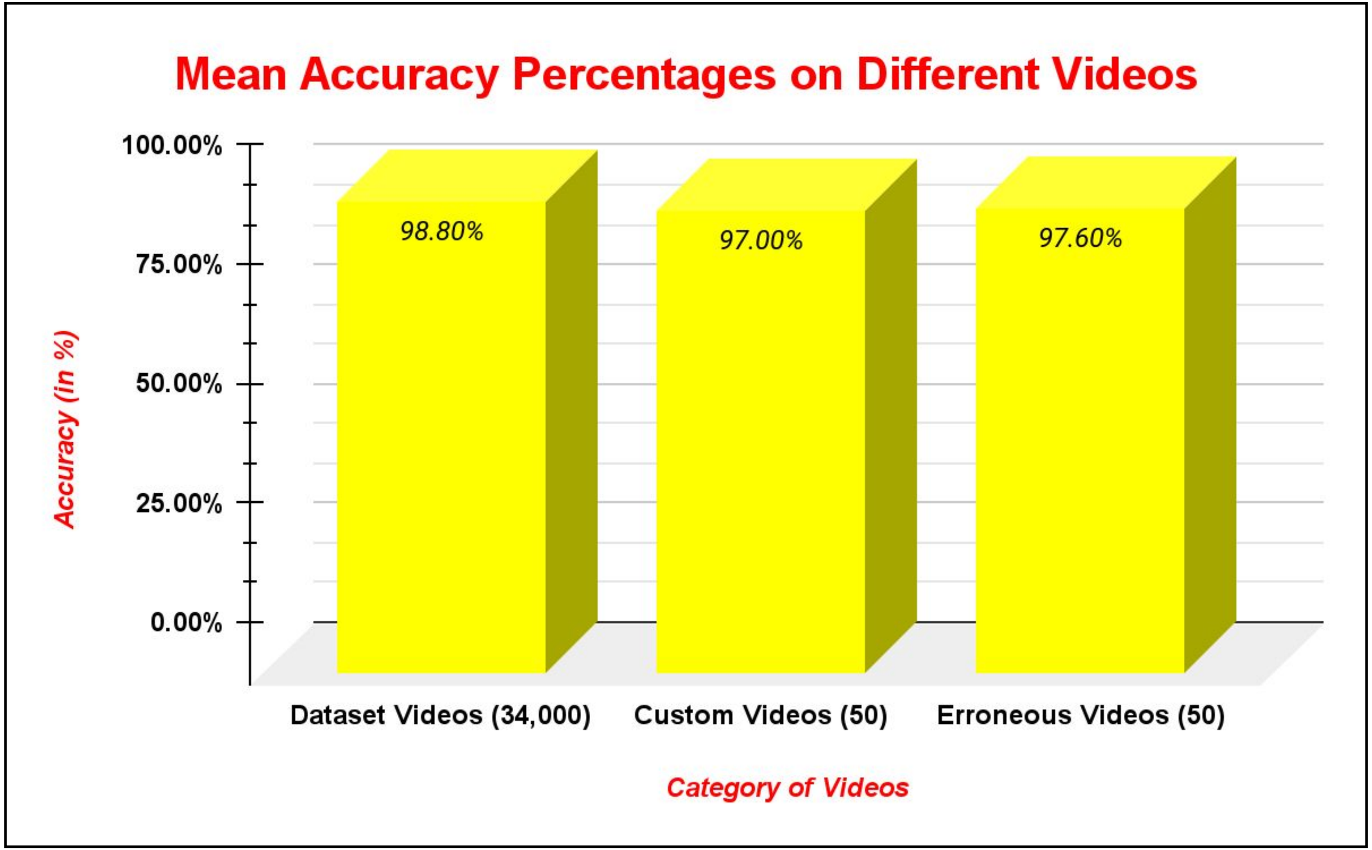
An error occurred. Please provide a video with a visible speaker.

Error message displayed on the LipVision interface in response to giving an erroneous video as input.

Evaluation Metric: Accuracy

It is calculated as the ratio of correct predictions (C) to the total number of words (T) multiplied by 100.

Accuracy = (C/T) x 100



CONCLUSION

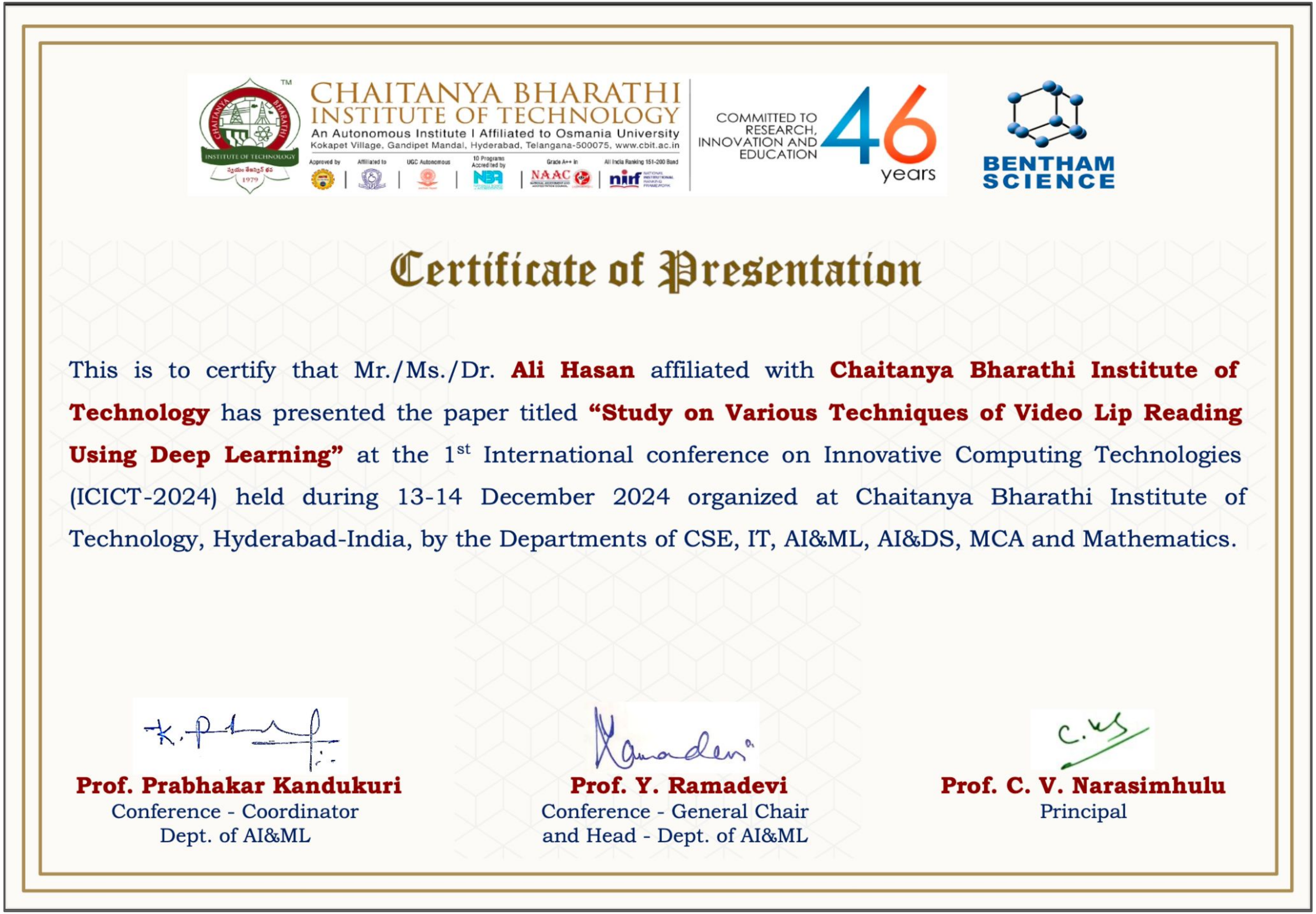
- LipVision, a deep-learning web application for video lip-reading and translation, achieved 97.8% accuracy by integrating 3D-CNN and Bi-LSTM networks trained on the GRID Audiovisual Sentence Corpus.

- The system features a user-friendly interface built with Flask, HTML, CSS, and JavaScript, with Googletrans enabling multilingual translation capabilities beyond the core transcription functions.

- Real-world applications include enhancing communication for the hearing-impaired, enabling silent communication in security contexts, and providing automated subtitling for content creators.

- Future enhancements could include speaker diarization for handling multiple speakers, training on diverse multilingual datasets, and implementing advanced architectures like 3D Convolutional Vision Transformers.

Literature Paper Status



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

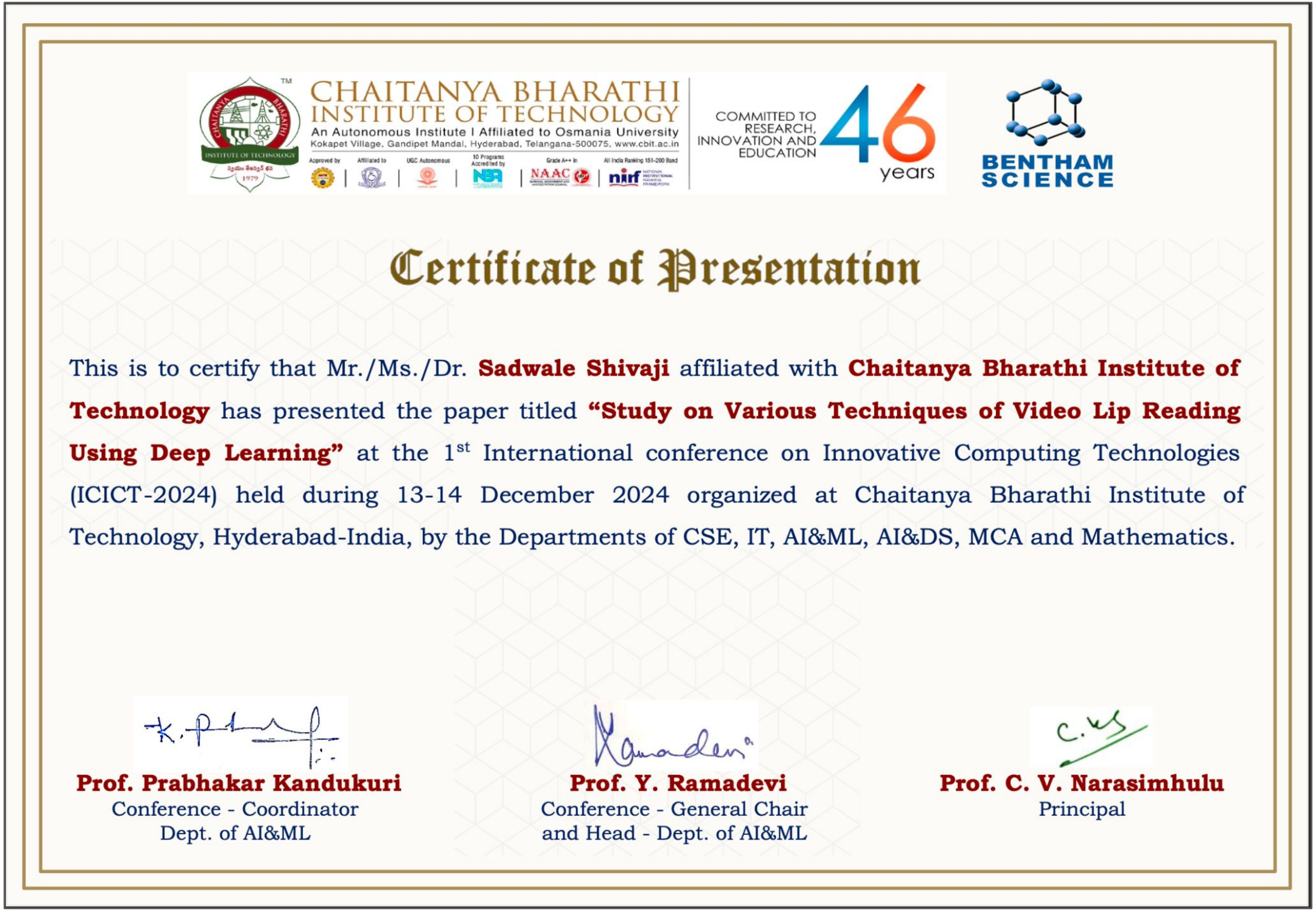
Certificate of Presentation

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



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

Approved by | Affiliated to | UGC Autonomous | 10 Programs Accredited by NBA | Grade A++ in NAAC | All India Ranking 151-200 Band nirf

COMMITTED TO RESEARCH, INNOVATION AND EDUCATION 46 years

BENTHAM SCIENCE

Certificate of Presentation

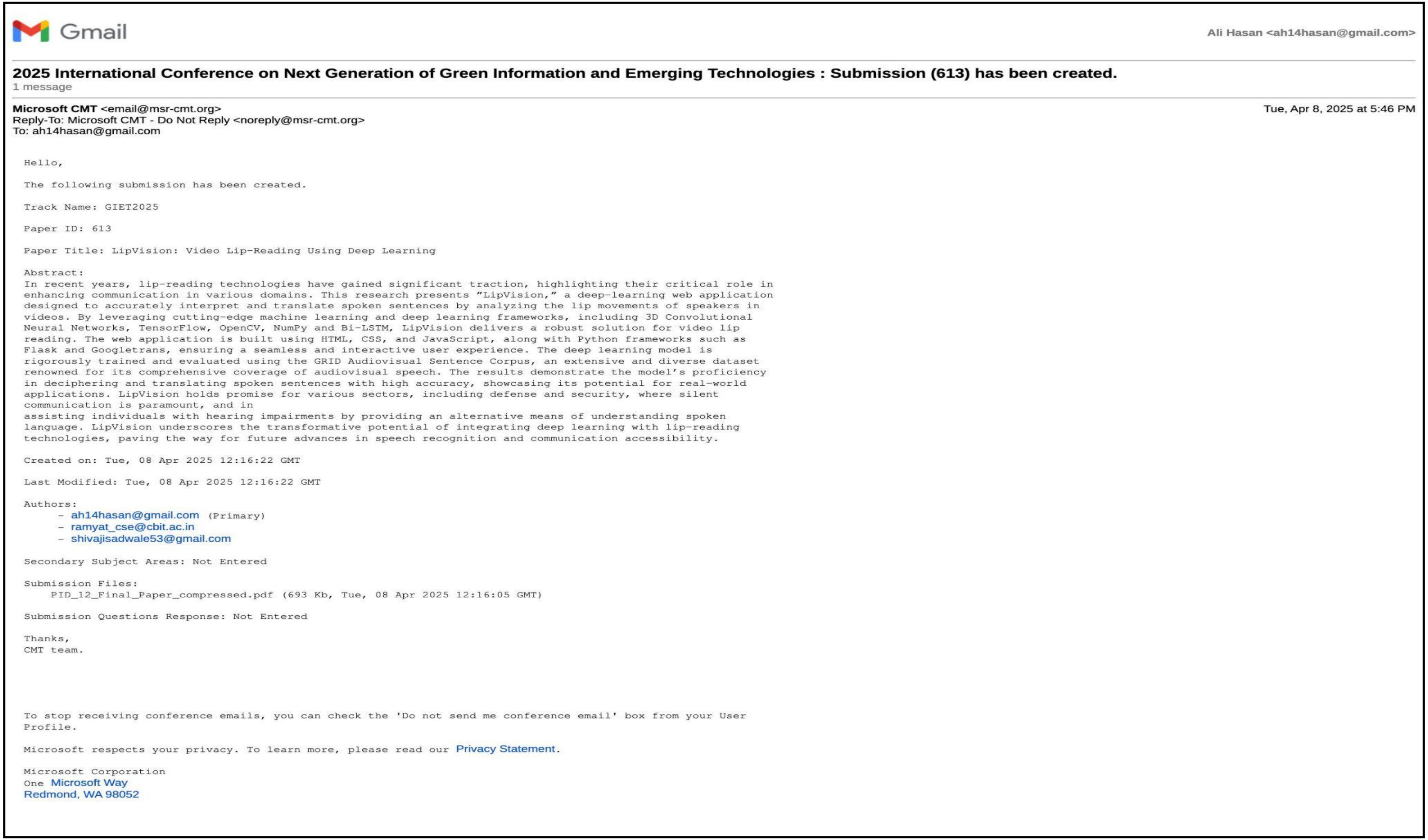
This is to certify that Mr./Ms./Dr. **Sadwale Shivaji** affiliated with **Chaitanya Bharathi Institute of Technology** has presented the paper titled **“Study on Various Techniques of Video Lip Reading Using Deep Learning”** at the 1st International conference on Innovative Computing Technologies (ICICT-2024) 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

Final Paper Status



Ali Hasan <ah14hasan@gmail.com>

2025 International Conference on Next Generation of Green Information and Emerging Technologies : Submission (613) has been created.

1 message

Microsoft CMT <email@msr-cmt.org>
Reply-To: Microsoft CMT - Do Not Reply <noreply@msr-cmt.org>
To: ah14hasan@gmail.com

Tue, Apr 8, 2025 at 5:46 PM

Hello,

The following submission has been created.

Track Name: GIET2025

Paper ID: 613

Paper Title: LipVision: Video Lip-Reading Using Deep Learning

Abstract:
In recent years, lip-reading technologies have gained significant traction, highlighting their critical role in enhancing communication in various domains. This research presents "LipVision," a deep-learning web application designed to accurately interpret and translate spoken sentences by analyzing the lip movements of speakers in videos. By leveraging cutting-edge machine learning and deep learning frameworks, including 3D Convolutional Neural Networks, TensorFlow, OpenCV, NumPy and Bi-LSTM, LipVision delivers a robust solution for video lip reading. The web application is built using HTML, CSS, and JavaScript, along with Python frameworks such as Flask and Googletrans, ensuring a seamless and interactive user experience. The deep learning model is rigorously trained and evaluated using the GRID Audiovisual Sentence Corpus, an extensive and diverse dataset renowned for its comprehensive coverage of audiovisual speech. The results demonstrate the model's proficiency in deciphering and translating spoken sentences with high accuracy, showcasing its potential for real-world applications. LipVision holds promise for various sectors, including defense and security, where silent communication is paramount, and in
assisting individuals with hearing impairments by providing an alternative means of understanding spoken language. LipVision underscores the transformative potential of integrating deep learning with lip-reading technologies, paving the way for future advances in speech recognition and communication accessibility.

Created on: Tue, 08 Apr 2025 12:16:22 GMT

Last Modified: Tue, 08 Apr 2025 12:16:22 GMT

Authors:
- ah14hasan@gmail.com (Primary)
- ramyat_cse@cbit.ac.in
- shivajisadwale53@gmail.com

Secondary Subject Areas: Not Entered

Submission Files:
PID_12_Final_Paper_compressed.pdf (693 Kb, Tue, 08 Apr 2025 12:16:05 GMT)

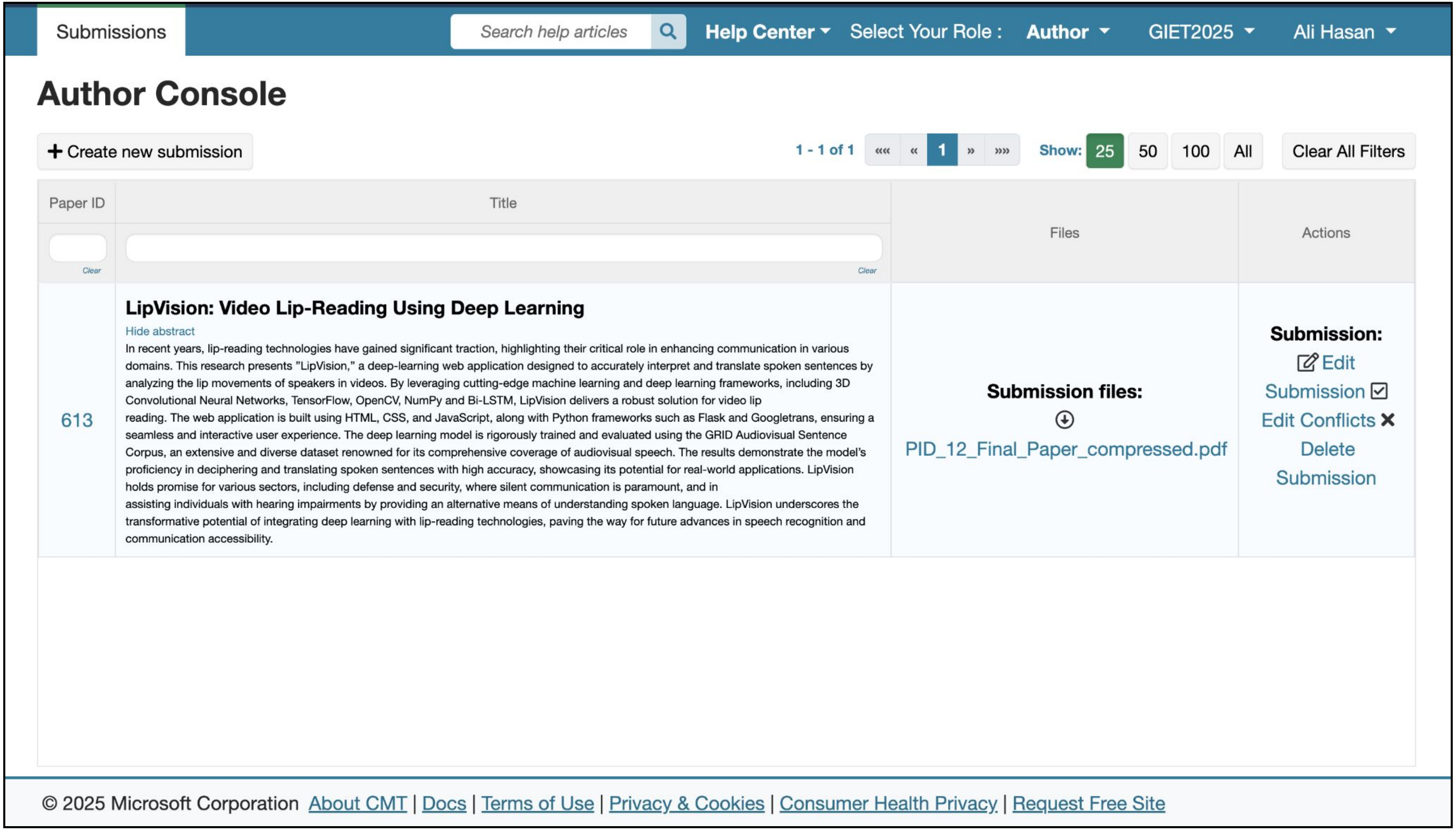
Submission Questions Response: Not Entered

Thanks,
CMT team.

To stop receiving conference emails, you can check the 'Do not send me conference email' box from your User Profile.

Microsoft respects your privacy. To learn more, please read our Privacy Statement.

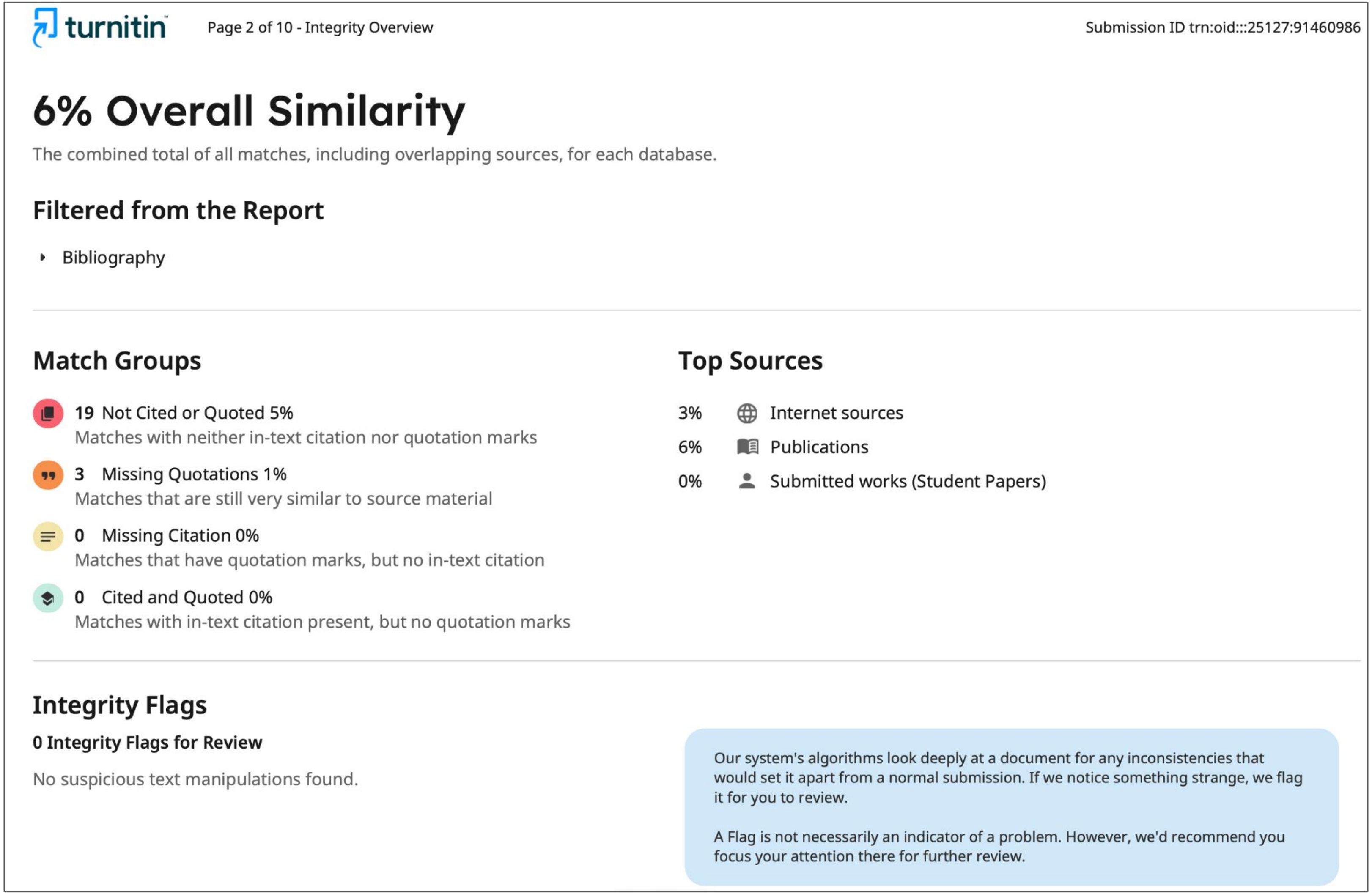
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052



Author Console

Paper ID	Title	Files	Actions
613	**LipVision: Video Lip-Reading Using Deep Learning** Hide abstract In recent years, lip-reading technologies have gained significant traction, highlighting their critical role in enhancing communication in various domains. This research presents ”LipVision,” a deep-learning web application designed to accurately interpret and translate spoken sentences by analyzing the lip movements of speakers in videos. By leveraging cutting-edge machine learning and deep learning frameworks, including 3D Convolutional Neural Networks, TensorFlow, OpenCV, NumPy and Bi-LSTM, LipVision delivers a robust solution for video lip reading. The web application is built using HTML, CSS, and JavaScript, along with Python frameworks such as Flask and Googletrans, ensuring a seamless and interactive user experience. The deep learning model is rigorously trained and evaluated using the GRID Audiovisual Sentence Corpus, an extensive and diverse dataset renowned for its comprehensive coverage of audiovisual speech. The results demonstrate the model’s proficiency in deciphering and translating spoken sentences with high accuracy, showcasing its potential for real-world applications. LipVision holds promise for various sectors, including defense and security, where silent communication is paramount, and in assisting individuals with hearing impairments by providing an alternative means of understanding spoken language. LipVision underscores the transformative potential of integrating deep learning with lip-reading technologies, paving the way for future advances in speech recognition and communication accessibility.	**Submission files:** PID_12_Final_Paper_compressed.pdf	**Submission:** Edit Submission ☑ Edit Conflicts ✖ Delete Submission

© 2025 Microsoft Corporation About CMT \| Docs \| Terms of Use \| Privacy & Cookies \| Consumer Health Privacy \| Request Free Site



6% Overall Similarity

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

Filtered from the Report

- Bibliography

Match Groups

- **19** Not Cited or Quoted 5%
  Matches with neither in-text citation nor quotation marks
- **3** Missing Quotations 1%
  Matches that are still very similar to source material
- **0** 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

3%	Internet sources
6%	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.

REFERENCES

[1] B. Martinez, P. Ma, S. Petridis and M. Pantic, "Lipreading Using Temporal Convolutional Networks," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 6319-6323, doi: 10.1109/ICASSP40776.2020.9053841.

[2] N. K. Mudaliar, K. Hegde, A. Ramesh and V. Patil, "Visual Speech Recognition: A Deep Learning Approach," *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, India, 2020, pp. 1218-1221, doi: 10.1109/ICCES48766.2020.9137926.

[3] M. Wand, J. Koutník and J. Schmidhuber, "Lipreading with long short-term memory," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 6115-6119, doi: 10.1109/ICASSP.2016.7472852.

[4] U. Saeed *et al*., "Extracting Visual Micro-Doppler Signatures From Human Lips Motion Using UoG Radar Sensing Data for Hearing Aid Applications," in *IEEE Sensors Journal*, vol. 23, no. 19, pp. 22111-22118, 1 Oct.1, 2023, doi: 10.1109/JSEN.2023.3308972.

[5] Li, Dengshi, et al. "Improving speech recognition performance in noisy environments by enhancing lip reading accuracy." *Sensors* 23.4 (2023): 2053.

[6] Fenghour, Souheil, et al. "An effective conversion of visemes to words for high-performance automatic lipreading." *Sensors* 21.23 (2021): 7890.

[7] Mudaliar, Navin Kumar, et al. "Visual speech recognition: a deep learning approach." *2020 5th international conference on communication and electronics systems (ICCES)*. IEEE, 2020.

[8] Howell, Dominic, Stephen Cox, and Barry Theobald. "Visual units and confusion modelling for automatic lip-reading." *Image and Vision Computing* 51 (2016): 1-12.

[9] Petridis, Stavros, Zuwei Li, and Maja Pantic. "End-to-end visual speech recognition with LSTMs." *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017.

[10] Hilder, Sarah, Richard W. Harvey, and Barry-John Theobald. "Comparison of human and machine-based lip-reading." *AVSP*. 2009.

[11] Shillingford, Brendan, et al. "Large-scale visual speech recognition." *arXiv preprint arXiv:1807.05162* (2018).

[12] Xu, Kai, et al. "LCANet: End-to-end lipreading with cascaded attention-CTC." *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018.

[13] Noda, Kuniaki, et al. "Lipreading using convolutional neural network." Interspeech. Vol. 1. 2014.

[14] Miled, Malek, Mohammed Anouar Ben Messaoud, and Aicha Bouzid. "Lip reading of words with lip segmentation and deep learning." Multimedia Tools and Applications 82.1 (2022): 551-571.

[15] Wang, Huijuan, et al. "Mini-3DCvT: a lightweight lip-reading method based on 3D convolution visual transformer." The Visual Computer (2024): 1-13.

[16] Ma, Pingchuan, et al. "Lip-reading with densely connected temporal convolutional networks." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021.

[17] Assael, Yannis M., et al. "Lipnet: End-to-end sentence-level lipreading." arXiv preprint arXiv:1611.01599 (2016).

Thank you