LipVision: Video Lip-Reading Using Deep Learning

1st Ramya Thavva
Department of AIML
CBIT, Gandipet
Hyderabad, India
ramyat_cse@cbit.ac.in

2nd Ali Hasan
Department of AIML
CBIT, Gandipet
Hyderabad, India
ah14hasan@gmail.com

3rd Sadwale Shivaji
Department of AIML
CBIT, Gandipet
Hyderabad, India
shivajisadwale53@gmail.com

***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.**

***Index Terms*—Lip reading, Deep Learning, 3D-Convolutional Neural Networks, TensorFlow, Bi-LSTM, GRID Audiovisual Sentence corpus, Flask, NumPy, OpenCV.**

I. Introduction

Video lip reading technology has evolved from expert-dependent interpretation to automated deep learning systems that address hearing loss, which affects 1.4 billion people worldwide. LipVision advances this field by transforming lip movements into comprehensible sentences through an innovative architecture utilizing the GRID Audiovisual Sentence dataset, TensorFlow, 3DCNN, and LSTM networks. Going beyond word-level recognition to sentence-level interpretation with multilingual translation capabilities, LipVision delivers these advancements through an accessible web application, offering a comprehensive solution for the hearing-impaired community's communication needs.

LipVision serves various accessibility purposes while providing substantial benefits in all professional sectors. For people with hearing impairments, it breaks communication barriers through automatic caption generation for videos, transcription of public announcements, and support for remote education accessibility. The technology also offers significant advantages to content creators through cost-effective subtitle generation, legal environments via enhanced documentation, security operations by enabling silent communication and surveillance analysis, and linguistic research through automated processing of video datasets. This dual focus on individual accessibility and professional applications positions LipVision as a transformative tool for both personal communication and specialized operational contexts.

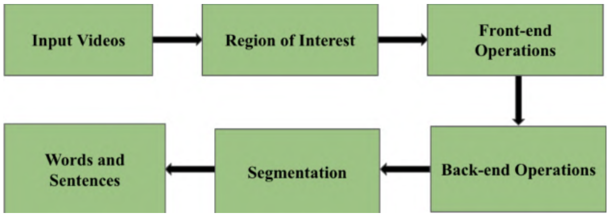


Fig. 1: Block diagram of a typical Lip-Reading architecture.

A. Methodology

The broad structure of lip-reading systems contains a front-end to extract features, a back-end to classify, and some pre-processing steps [3]. A typical automated lip reading process consists of the following steps shown in Figure 1 [3].

- **Input videos:** Videos capturing different speakers are divided into individual frames that must be interpreted and decoded.
- **Region of Interest:** This pre-processing step involves identifying and extracting the Region of Interest (ROI). It includes the lips from the image data. It involves face detection, lip localization, and lip extraction from the frames. Basic computer vision manipulations like cropping are applied to the ROI to streamline the process of training and validation [13].
- **Front-End operations:** This step focuses on extracting pertinent features from redundant information and converting large-dimensional data into a more compact representation.
- **Back-end operations:** Speech is assigned to movements on the face that have been transformed to smaller-

dimensional feature vectors.

- **Words and Sentences:** The interpreted speech is classified into segments or categories, eventually being converted into spoken words or sentences.

B. Evaluation Metric: Accuracy

The benchmark metric through which the performance of a lip reading model is evaluated is called 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 [5]. It is given by the formula:

$$Accuracy = \frac{C}{T} \times 100 \tag{1}$$

C. Datasets

The various datasets used for implementation of lip reading using deep learning are described in Table I.

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

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.

II. Literature Review

A visual speech recognition (VSR) system using a convolutional neural network (CNN) was developed 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 [9].

LipNet, an advanced neural network model for complete sentence level lipreading, was proposed 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 [1].

A novel method to predict spoken sentences from silent videos using lip movements from the LRS2 dataset was proposed. 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 horizonatal 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 [4].

A DC-TCN (Densely Connected Temporal Convolutional Network) model was proposed 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 [7].

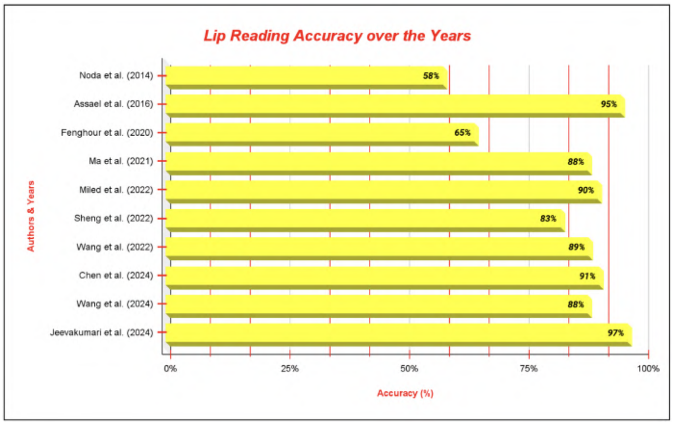


Fig. 2: Bar graph illustrating the trends of accuracies from 2014 to 2024 in various lip reading architectures.

A two-block lip reading system was 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 that 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 [8].

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 [10].

Another method using a 3D Convolutional Vision Transformer (3DCvT) was proposed, 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 [12].

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) were also proposed. 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 [2].

Another innovative approach called Mini-3DCvT, which integrated 3D convolution with visual transformers to effectively extract spatiotemporal features from videos, was proposed. 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 [11].

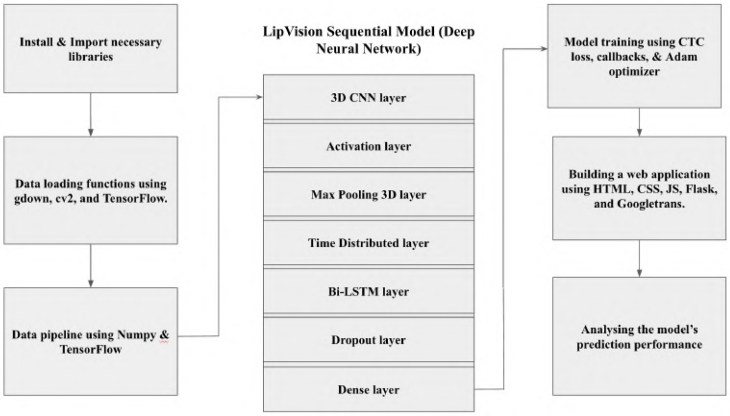

Install & Import necessary libraries
Data loading functions using gdown, cv2, and TensorFlow.
Data pipeline using Numpy & TensorFlow
LipVision Sequential Model (Deep Neural Network)
3D CNN layer
Activation layer
Max Pooling 3D layer
Time Distributed layer
Bi-LSTM layer
Dropout layer
Dense layer
Model training using CTC loss, callbacks, & Adam optimizer
Building a web application using HTML, CSS, JS, Flask, and Googletrans.
Analysing the model's prediction performance

Fig. 3: Architecture diagram of the proposed system.

Another approach for VSR in noisy environments was proposed 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 [6].

Figure 2 illustrates the accuracies achieved by different lip reading approaches. Table II summarizes these approaches and highlights their key limitations.

III. Proposed System

A. Architecture Diagram

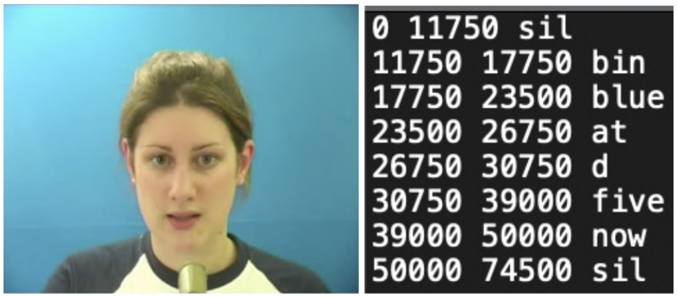
The proposed architecture along with the flow of events and required modules and frameworks is illustrated in Figure 3.

B. Implementation Procedure

The first step is to import and install all the required deep learning dependencies that will be used to implement the model. The second step is to create data loading functions using gdown, cv2, and TensorFlow Python frameworks to extract videos, frames, and alignments (transcriptions) from the GRID dataset. Then, NumPy and Tensorflow Python libraries are used to create a data pipeline to feed the data into the model through a standardized array format. Then, the deep

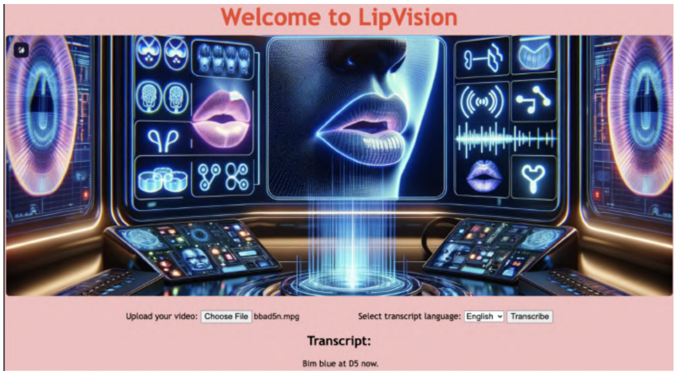
TABLE II: Overview of the various lip reading approaches and their limitations.

Authors	Approach	Limitations
Noda, Kuniaki, et al. (2014)	Visual Speech Recognition (VSR) system with CNN for extracting features from mouth area images.	Use of a small dataset affecting generalizability, reliance on speaker-dependent models, limited consonant recognition.
Assael, Yannis M., et al. (2016)	Complete sentence level lip reading using STCNNs, Bi-GRUs, and CTC loss.	Use of a limited dataset, uncertain performance in uncontrolled, noisy environments.
Fenghour, Souheil, et al. (2020)	Predicting spoken sentences from silent videos using 3D convolutional network and 2D ResNet.	Significant gap between Viseme and Word error rates, indicating problems in converting viseme to words.
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.	Increased computational complexity, scope for improving accuracy on the LRW-1000 dataset.
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).	Potential challenges in noisy environments and reliance on a specific dataset, affecting generalizability.
Sheng, Changchong, et al. (2022)	Adaptive Semantic-Spatio-Temporal Graph Convolutional Network (ASST-GCN), Transformers, and Convolutional Networks.	Over-reliance on precise facial landmark detection leads to challenges in complicated real-world scenarios.
Wang, Huijuan, Gangqiang Pu, and Tingyu Chen. (2022)	3D Convolutional Vision Transformer and BiGRU (bidirectional gated recurrent unit).	Complexity of the model leading to high computational demands, and limiting real-world applicability.
Chen, Hang, et al. (2024)	VSM (Hybrid and Collaborative) and End to End Modeling.	Computational complexity and reliance on large dataset limiting real-world applicability.
Wang, Huijuan, et al. (2024)	3D Convolution and Visual Transformer techniques such as BiGRU.	Challenges in accurately capturing subtle lip movements and potential performance degradation during compressing of models.
Jeevakumari, SA Amutha, and Koushik Dey. (2024)	3D Convolutional Neural Networks (3D CNN), Bi-Directional Long Short-Term Memory (BiLSTM), and Connectionist Temporal Classification (CTC) loss.	Reliance on a specific dataset affecting generalizability, need for larger datasets and alternate architectures to enhance real-world performance.



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

Fig. 4: A video and its corresponding transcription from the GRID dataset.



Upload your video: bbad5n.mpg

Select transcript language: English

Transcript:

Bim blue at D5 now.

Fig. 5: LipVision interface with the generated transcription in English.

neural network of the model is designed using TensorFlow's sequential model having various deep learning layers such as 3D CNN layer, activation layer, Max Pooling 3D layer, Time Distributed Layer, Bi-LSTM layer, Dropout layer, and Dense Layer. The CTC (Connectionist Temporal Classification) loss function, callback functions, and Adam optimizer are then used to train the model on the alignments and videos of the GRID dataset. An interactive web application is then built using HTML, CSS, JS, and Python Frameworks such as Flask and Googletrans. The model's performance is then analyzed and its ability in making accurate predictions and translations of the sentences being spoken in the videos is evaluated.

IV. Testing and Results

A. *Testing on a Dataset Video*

One of the videos available in the GRID dataset and its transcription is first downloaded. The screenshot of the video and its corresponding transcription is shown in Figure 4

The video in Figure 4 is then chosen in the LipVision interface and a transcription is generated in the English language as shown in Figure 5.

B. *Testing on a Custom Video*

A completely new, custom video of a previously unseen speaker is created, the screenshot of which is shown in Figure 6.

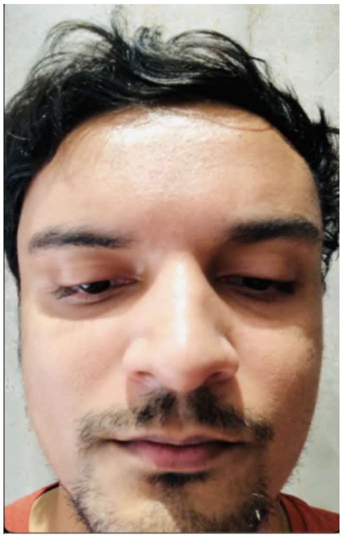
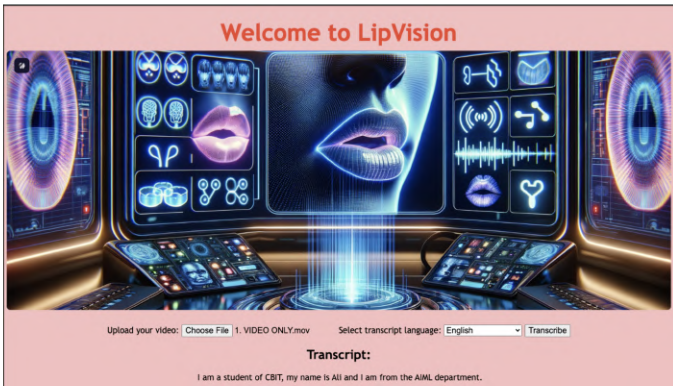


Fig. 6: Screenshot of the custom video.



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.

Fig. 7: LipVision interface with the generated transcription in English.

The video in Figure 6 is then chosen in the LipVision interface and a transcription is generated in the English language as shown in Figure 7.

The choice of language is then changed to Telugu and the transcript for the same video is generated again as shown in Figure 8.

C. Testing on an Erroneous Video

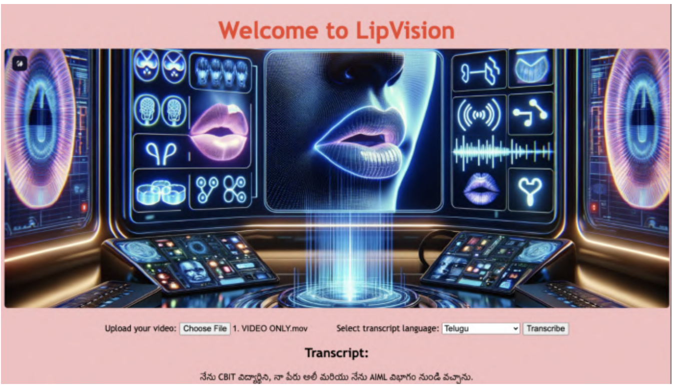
An erroneous video, that is, a video with no visible speaker, is created, the screenshot of which is shown in Figure 9.

The video in Figure 9 is then chosen in the LipVision interface and an attempt is made to generate the transcription of the video in the English language. The result is an error message as shown in Figure 10.

V. Discussions

A. Accuracies Achieved

Based on the formula of accuracy as defined in section I-B, the accuracy of LipVision in generating accurate transcriptions is measured on the GRID dataset videos, custom videos, as



Welcome to LipVision

Upload your video: 1. VIDEO ONLY.mov

Select transcript language: Telugu

Transcript:

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

Fig. 8: LipVision interface with the generated transcription in Telugu.

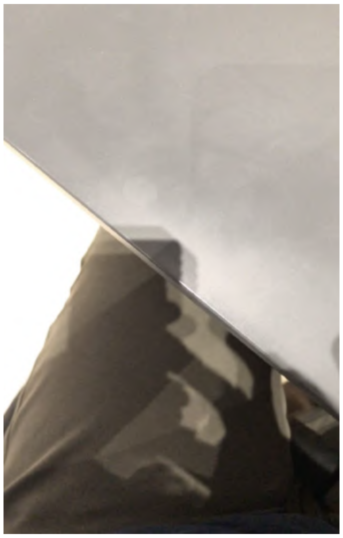
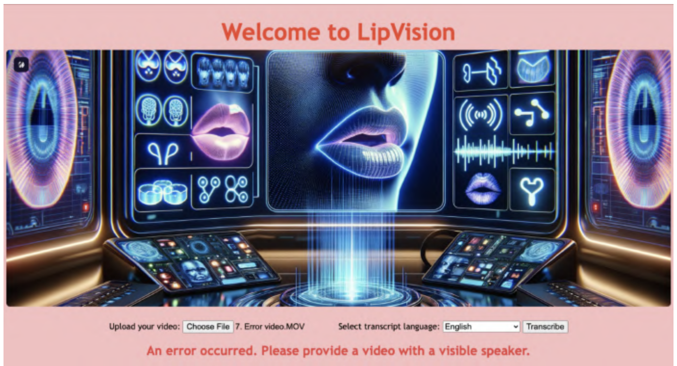


Fig. 9: An erroneous video with no visible speaker.



Welcome to LipVision

Upload your video: 7. Error video.MOV

Select transcript language: English

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

Fig. 10: Error message displayed on the LipVision interface in response to giving an erroneous video as input.

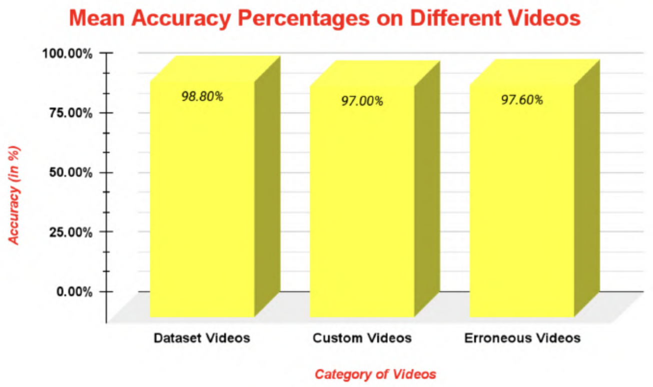


Fig. 11: Bar graph showing the mean accuracy percentages obtained after testing LipVision on the GRID dataset videos, custom videos, and erroneous videos.

well as erroneous videos. The ability of LipVision to accurately categorize and distinguish between erroneous videos and valid input videos is observed, recorded, and summarized in the bar graph shown in Figure 11.

LipVision is observed to perform the best on the GRID dataset with an accuracy of 98.80%, a significant improvement over the base paper's 96.7% accuracy on the same dataset [6]. 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%.

B. Future Scope

LipVision has demonstrated exceptional performance in video lip reading technology with 97.8% accuracy, successfully converting lip movements into text transcriptions. Despite this achievement, several promising development paths could significantly expand its capabilities and applications. The most notable enhancements include implementing speaker diarization to differentiate between multiple individuals in videos, which would prove valuable for meeting transcriptions, movie subtitling, and interview documentation where speaker identification is crucial.

Further advancements would involve training the system on diverse multilingual datasets to enable direct interpretation of languages beyond English, eliminating the current two-step process of English transcription followed by translation. Additionally, exploring advanced architectural approaches like the 3D Convolutional Vision Transformer (3DCvT) could potentially surpass the current neural network structure's performance, especially when processing complex sentences or challenging visual conditions. These enhancements would collectively transform LipVision from an already impressive technology into a more versatile and comprehensive communication solution applicable across a broader range of real-world scenarios.

VI. Conclusions

LipVision, a deep-learning web application that performs video lip-reading and translation with 97.8% accuracy, was successfully developed. The system combined 3D Convolutional Neural Networks and Bi-directional LSTM networks trained on the GRID Audiovisual Sentence Corpus using TensorFlow. It successfully converted lip movements into coherent English sentences while identifying invalid inputs without visible speakers. The application featured a user-friendly interface built with Flask, HTML, CSS, and JavaScript, with Googletrans providing multilingual translation capabilities. Testing confirmed its effectiveness with both dataset videos and custom recordings, demonstrating readiness for real-world implementation.

LipVision addresses key societal needs by enhancing accessibility for the hearing-impaired, enabling silent communication in security contexts, and providing automated subtitling for content creators. Future enhancements could include speaker diarization for multiple speakers, training on multilingual datasets, and implementing advanced architectures like 3D Convolutional Vision Transformers to further advance lip-reading technology.

References

[1] Assael, Y.M., Shillingford, B., Whiteson, S., De Freitas, N.: Lipnet: End-to-end sentence-level lipreading. arXiv preprint arXiv:1611.01599 (2016)

[2] Chen, H., Wang, Q., Du, J., Wan, G.S., Xiong, S.F., Yin, B.C., Pan, J., Lee, C.H.: Collaborative viseme subword and end-to-end modeling for word-level lip reading. IEEE Transactions on Multimedia (2024)

[3] Fenghour, S., Chen, D., Guo, K., Li, B., Xiao, P.: Deep learning-based automated lip-reading: A survey. IEEE Access **9**, 121184–121205 (2021)

[4] Fenghour, S., Chen, D., Guo, K., Xiao, P.: Lip reading sentences using deep learning with only visual cues. IEEE Access **8**, 215516–215530 (2020)

[5] Ferraro, A., Galli, A., La Gatta, V., Postiglione, M.: Benchmarking open source and paid services for speech to text: an analysis of quality and input variety. Frontiers in big Data **6**, 1210559 (2023)

[6] Jeevakumari, S.A., Dey, K.: Lipsyncnet: A novel deep learning approach for visual speech recognition in audio-challenged situations. IEEE Access (2024)

[7] Ma, P., Wang, Y., Shen, J., Petridis, S., Pantic, M.: Lip-reading with densely connected temporal convolutional networks. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2857–2866 (2021)

[8] Miled, M., Messaoud, M.A.B., Bouzid, A.: Lip reading of words with lip segmentation and deep learning. Multimedia Tools and Applications **82**(1), 551–571 (2023)

[9] Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H.G., Ogata, T., et al.: Lipreading using convolutional neural network. In: Interspeech. vol. 1, p. 3 (2014)

[10] Sheng, C., Zhu, X., Xu, H., Pietikäinen, M., Liu, L.: Adaptive semantic-spatio-temporal graph convolutional network for lip reading. IEEE Transactions on Multimedia **24**, 3545–3557 (2021)

[11] Wang, H., Cui, B., Yuan, Q., Pu, G., Liu, X., Zhu, J.: Mini-3dcvt: a lightweight lip-reading method based on 3d convolution visual transformer. The Visual Computer pp. 1–13 (2024)

[12] Wang, H., Pu, G., Chen, T.: A lip reading method based on 3d convolutional vision transformer. IEEE Access **10**, 77205–77212 (2022)

[13] Zhang, J.X., Wan, G., Pan, J.: Is lip region-of-interest sufficient for lipreading? In: Proceedings of the 2022 International Conference on Multimodal Interaction. pp. 368–372 (2022)