WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Speak Sync AI: Transforming Lip Movements into Text with AI Precision

[1]Prof. Sagar Birje, [2] Prasad Nandeshwar, [3]Neelambika Fatakal, [4]Bhavana Kalloli, [5]Danappa Dayappanavar

[1]Professor, Department of Artificial Intelligence & Data Science, Angadi Institute of Technology and Management, Belagavi, Karnataka, India.
[2,3,4,5]Student, Department of Artificial Intelligence & Data Science, Angadi Institute of Technology and Management, Belagavi, Karnataka, India.

ABSTRACT:

The Speak Sync Prediction System is an innovative deep learning project aimed at improving human-computer interaction, enabling assistive communication, and delivering entertainment via quiet voice detection. The device interprets spoken language by examining video frames and recognizing visual lip movements, facilitating silent communication. This is particularly advantageous for individuals with speech impairments or in contexts where verbal communication is unfeasible, such as noise-sensitive or isolated environments. The system utilizes a hybrid of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to effectively extract spatial and temporal information from the lip area. It incorporates components such as facial landmark detection and lip movement tracking to convert these visual inputs into corresponding textual or audio outputs. The model has been evaluated in real-time scenarios and has demonstrated high accuracy and robust performance, making it suitable for practical use. Moreover, visualization techniques are utilized to display the sequences of lip movements along with their expected interpretations, hence enhancing user understanding and system transparency. This experiment illustrates the substantial relevance of deep learning and computer vision in bridging communication gaps and promoting inclusive technologies.

Keywords: Silent speech recognition, CNN-RNN based analysis, real-time accuracy, lip movement extraction, visual output display, and support for accessibility.

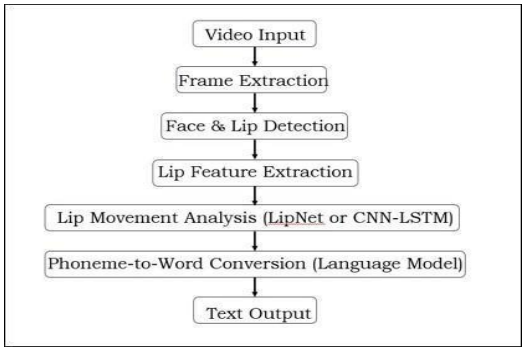
Introduction:

Introduction:The block diagram offers a comprehensive depiction of the system's process, illustrating the progression from user input to final text output. The process commences with the user supplying a video clip that exhibits discernible lip movements to the system. This input functions as the principal data for the system's analysis. Upon upload, the video enters a pre-processing stage. During pre-processing, the system executes tasks like frame extraction, resizing, normalization, and maybe noise reduction to prepare the video data for analysis. The polished frames are subsequently directed to the feature extraction phase, where notable visual characteristics associated with lip movement are discerned utilizing methods such as convolutional neural networks (CNNs). Subsequent to feature extraction, the system employs pattern matching to juxtapose the extracted features with predefined patterns of identified speech motions. This stage is crucial for comprehending lip motions within the framework of language. Upon completing the matching process, the system executes classification to associate the identified patterns with specific text characters or phrases, utilizing previously trained models. The categorized output is subsequently compiled and shown in textual form.

Methodology:

This project utilizes a process that includes the design and construction of a Speak-sync system, developed through deep learning techniques. This system can interpret spoken words exclusively from visual signals (lip movements) without utilizing auditory data. The methodology encompasses the following stages:

Figure 2: Methodology of Speak Sync AI



2.1 System Workflow

2.1.1 Video Input and Preprocessing

- **Video Input**: The system accepts a silent video of a person speaking.
- **Frame Extraction**: The input video is divided into individual frames for detailed analysis.
- **Face & Lip Detection**: Each frame is processed to detect and crop the lip region using techniques such as OpenCV or Haar cascades.

2.1.2 Feature Extraction

- **3D-CNN Layers**: These layers are used to extract spatiotemporal features from the video sequence, capturing motion and appearance across multiple frames.
- **EfficientNetB0**: Utilized for extracting high-level spatial features from individual video frames to understand abstract facial characteristics.
- **Feature Fusion**: The features obtained from 3D-CNN and EfficientNetB0 are fused to create a rich representation of lip movements.

2.1.3 Sequence Modelling and Transcription

- **Bi-LSTM**: A bidirectional Long Short-Term Memory network is used to capture forward and backward dependencies in the temporal sequence of features.
- **CTC Loss (Connectionist Temporal Classification)**: Enables sequence-to-sequence mapping without requiring exact alignment between input frames and output labels.

2.1.4 Training Phase

- **Model Training**: The model is trained using **TensorFlow**, utilizing GPU acceleration to optimize training speed and efficiency.
- **Input Data**: The model is trained on labelled datasets (e.g., GRID corpus), which provide frame sequences and their corresponding textual outputs.

2.1.5 Text Decoding and Output

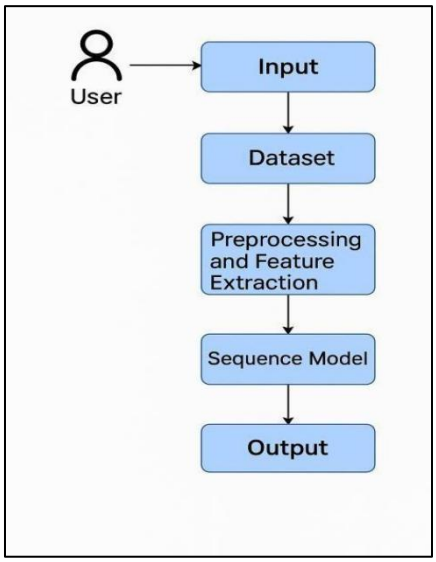
- **Phoneme-to-Word Conversion**: A language model is used to convert predicted phonemes or characters into valid words and sentences.
- **Text Output**: The decoded text is presented as the system's final interpretation of the lip movements from the silent video.

SYSTEM DESIGN

3.1 Architecture Diagram

The architectural diagram illustrates the complete pipeline of a machine learning or deep learning system designed for sequence-based tasks, such as speech recognition, lip reading, or natural language processing. The process begins with User Input, when unprocessed data (e.g., video, audio, or text) is provided to the system. The input is then matched or stored in the Dataset, which serves as the primary repository of data samples used for training or inference. The data undergoes preprocessing and feature extraction, wherein it is cleansed, normalized, and transformed into a suitable format that retains the essential characteristics required for learning. The processed data is relayed to the Sequence Model—usually a Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), or Transformer—that recognizes temporal relationships and detects patterns from sequential inputs. Ultimately, the model produces the Output, which may include a prediction, classification, or translated text, depending on the application. This design ensures a methodical, thorough advancement from raw input to meaningful output, enabling discerning decision-making from complex sequential data.

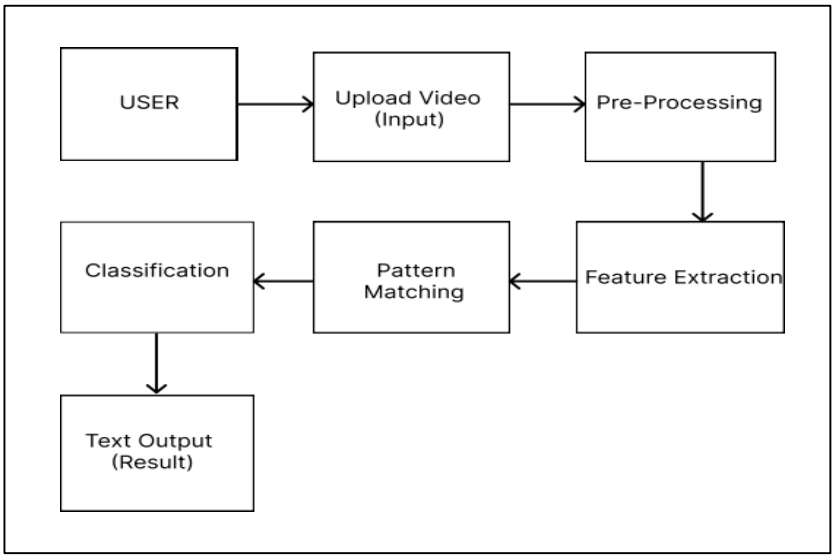
Figure 3.1.1: 3D view of building



3.1 Block Diagram

The block diagram provides a detailed representation of the system's workflow, encompassing the transition from user input to final text output. The process begins with the user sending a video clip featuring identifiable lip movements to the system. This input serves as the primary data for the system's analysis. Upon upload, the video undergoes a pre-processing phase. During pre-processing, the system performs activities such as frame extraction, resizing, normalization, and maybe noise reduction to prepare the video data for analysis. The refined frames are then routed to the feature extraction step, where significant visual attributes related to lip movement are identified using techniques such as convolutional neural networks (CNNs). Following feature extraction, the system utilizes pattern matching to compare the extracted features with established patterns of recognized speech movements. This phase is essential for understanding lip movements in the context of language. Following the matching process, the system performs classification to link the recognized patterns with particular text characters or words, employing pre-trained models. The classified output is ultimately aggregated and presented in textual format.

Fig. 3.2.1 Block Diagram



RESULTS AND DISCUSSION

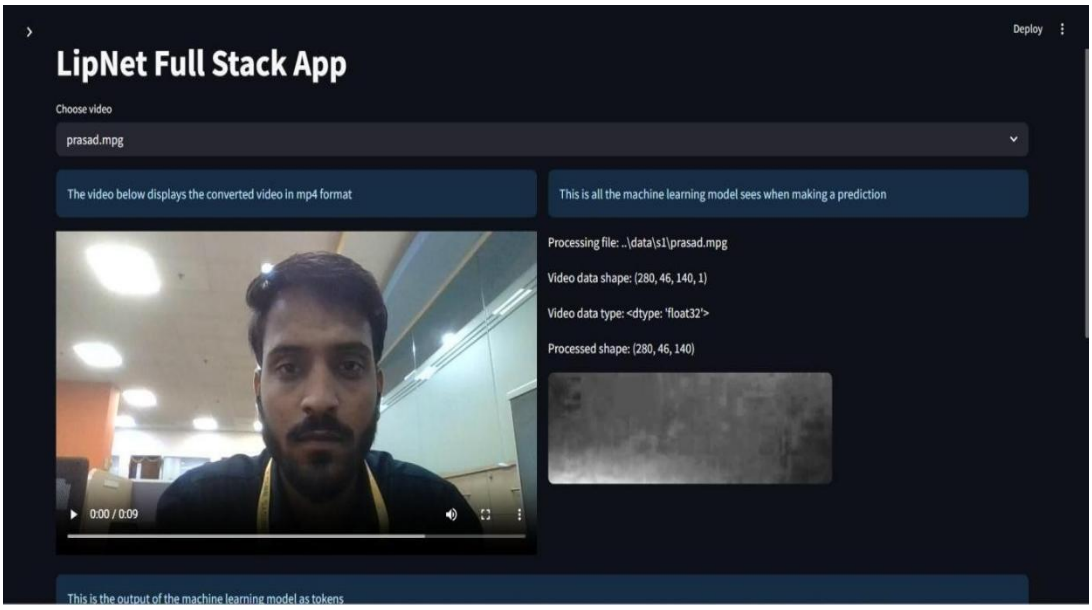
The **Speak Sync** application was successfully deployed on localhost:8501.
The user selected the video bba2fn.mpg, which was displayed after being converted to .mp4 format. The application processed the video with the following properties:

- Shape: (75, 46, 140, 1)
- Data Type: float32
- Final processed shape: (75, 46, 140)

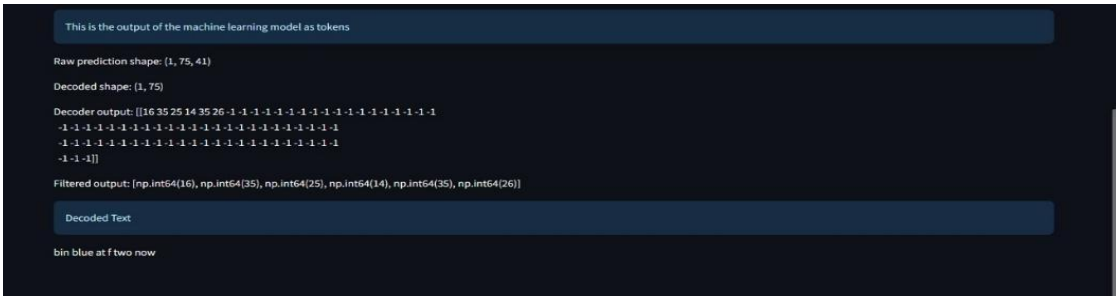
The system correctly handled video upload, preprocessing (grayscale conversion, resizing), and made the data ready for prediction. The user interface and backend processing worked without any errors.
The machine learning model successfully produced the output from the video input. The following results were observed:

- Raw prediction shape: (1, 75, 41)
- Decoded shape: (1, 75)
- Decoder output: Long sequence, mostly blanks (-1) except for token 29.
- Filtered

output: [29] Final
Decoded Text:
➔ "bin blue at f two now"
The system correctly decoded the video into understandable text, demonstrating the model's prediction capability.



Snapshot 7.1 Homepage



Snapshot 7.2 Final output

Conclusion

The block diagram offers a comprehensive depiction of the system's process, illustrating the progression from user input to final text output. The procedure commences with the user supplying a video clip that exhibits discernible lip movements to the system. This input constitutes the fundamental data for the system's analysis. Upon upload, the video enters a pre-processing phase. During pre-processing, the system executes tasks like frame extraction, resizing, normalization, and maybe noise reduction to prepare the video data for analysis. The polished frames are subsequently directed to the feature extraction phase, where notable visual characteristics associated with lip movement are discerned utilizing methods such as convolutional neural networks (CNNs). Subsequent to feature extraction, the system employs pattern matching to juxtapose the extracted features with predefined patterns of identified speech motions. This phase is crucial for comprehending lip motions within the framework of language. Upon completing the matching process, the system executes classification to associate the identified patterns with specific text characters or phrases, utilizing previously trained models. The classified output is ultimately compiled and provided in textual format.

REFERENCES

Research Papers:

1. D. Li, Y. Gao, C. Zhu, Q. Wang, and R. Wang, ''Improving speech recognition performance in noisy environments by enhancing lip reading accuracy,'' Sensors, vol. 23, no. 4, p. 2053, Feb. 2023, doi: 10.3390/s23042053.
2. S. Jeon and M. S. Kim, End-to-end sentence-level multi-view lipreading architecture with spatial attention module integrated multiple CNNs and cascaded local self-attention-CTC,' Sensors, vol. 22, no. 9, p. 3597, May 2022, doi: 10.3390/s22093597.
3. C. Sheng, X. Zhu, H. Xu, M. Pietikäinen, and L. Liu, ''Adaptive semantic-spatio-temporal graph convolutional network for lip reading,'' IEEE Trans. Multimedia, vol. 24, pp. 3545–3557, 2022, doi: 10.1109/TMM.2021.3102433.
4. H. Wang, G. Pu, andT.Chen, ''Alipreading method based on 3Dconvolutional vision transformer,'' IEEE Access, vol. 10, pp. 77205– 77212, 2022, doi: 10.1109/ACCESS.2022.3193231.
5. M. A. Haq, S.-J. Ruan, W.-J. Cai, and L. P. Li, ''Using lip reading recognition to predict daily Mandarin conversation,'' IEEE Access,

vol. 10, pp. 53481–53489, 2022, doi: 10.1109/ACCESS.2022.3175867.

6. Y. Xiao, L. Teng, A. Zhu, X. Liu, and P. Tian, ‘‘Lip reading in Cantonese,’’ IEEE Access, vol. 10, pp. 95020–95029, 2022, doi: 10.1109/ACCESS.2022.3204677.
7. S. Feng hour, D. Chen, K. Guo, B. Li, and P. Xiao, ‘‘Deep learning based automated lip-reading: A survey,’’ IEEE Access, vol. 9, pp. 121184–121205, 2021, doi: 10.1109/ACCESS.2021.3107946.
8. R. A. Ramadan, ‘‘Detecting adversarial attacks on audio-visual speech recognition using deep learning method,’’ Int. J. Speech Technol., vol. 25, pp. 625–631, Jun. 2021, doi: 10.1007/s10772-021-09859-3.
9. S. Feng hour, D.Chen,K.Guo,B.Li,andP.Xiao, ‘‘An effective conversion of visemes to words for high-performance automatic lipreading,’’ Sensors, vol. 21, no. 23, p. 7890, Nov. 2021, doi: 10.3390/s21237890.