LIP READING: TRANSFORMING SPEECHTO TEXT

Ankitha Bekal[1], Abhay Shetty M[1], Gourav[1], Maneesh Shetty[1]

[1]Department of Computer Science and Engineering, P. A. College of Engineering, Mangaluru, Karnataka, India.

*Corresponding Author: Ankitha Bekal Email: ankitha_cs@pace.edu.in

Abstract:

Lip reading, the ability to interpret spoken language by observing lip movements, is a valuable skill that can aid in various applications, particularly in enhancing speech recognition systems. This project explores the implementation of a deep learning-based lip reading model to improve the accuracy and robustness of speech recognition in challenging environments, such as noisy or audio-limited settings. The proposed lip reading system leverages Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to effectively capture temporal and spatial features from lip movement sequences. A comprehensive dataset comprising diverse speakers and linguistic contexts is used to train and evaluate the model, ensuring its generalization across different scenarios. The key steps of the project include data preprocessing, feature extraction, model training, and integration with existing speech recognition systems. The model is trained on synchronized audio-visual datasets to learn the correlation between speech signals and corresponding lip movements. To address real-world challenges, the system is designed to handle variations in lighting conditions, facial expressions, and speaker accents the performance of the lip reading-enhanced speech recognition system is evaluated through rigorous testing, comparing its accuracy and efficiency against traditional audio- only systems. The results demonstrate the potential of incorporating lip reading as a supplementary modality to improve the overall robustness and accuracy of speech recognition, especially in scenarios where audio signals alone may be insufficient. This research contributes to the growing field of multimodal deep learning and

ISBN-978-81-19905-39-3

highlights the practical applications of lip reading in enhancing human-computer interaction, accessibility, and communication systems. The findings open avenues for future research in developing more advanced and context-aware models that can further bridge the gap between visual and auditory information processing in intelligent systems.

Key Words: Lip reading, Recurrent Neural Networks, Convolutional Neural Networks, speech recognition

1. Introduction

In today's interconnected world, communication plays a vital role in every aspect of human interaction. While verbal communication is often taken for granted, there exists a significant population for whom traditional auditory communication is challenging or inaccessible. This includes individuals with hearing impairments as well as scenarios where environmental noise or distance inhibits clear communication. In such contexts, visual cues, particularly lip movements, serve as a crucial supplement or alternative to auditory cues. Lip reading, the skill of understanding speech by observing the movements of the lips, has garnered substantial interest to facilitate effective communication for individuals facing auditory challenges. Traditionally, lip reading has relied heavily on human interpretation, often with limited accuracy and efficiency. The advent of deep learning, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), has revolutionized the field of computer vision, enabling machines to surpass human capabilities in various visual recognition tasks. Leveraging the power of TensorFlow, a popular deep learning framework, researchers have increasingly explored automated lip reading systems as a promising solution to enhance communication accessibility. This project focuses on the development and implementation of a lip reading system using a deep learning approach within the TensorFlow framework. By harnessing the capabilities of neural networks, we aim to create a robust and accurate model capable of transcribing spoken words solely from visual lip movements captured in video footage. The system's functionality extends to both browsing pre-recorded video files and capturing real-time video through a webcam interface. Upon processing the video input, the system outputs the transcribed text, providing a seamless and intuitive means of communication for individuals with hearing impairments or in noisy environments.

ISBN-978-81-19905-39-3

2. Literature Review

Wand et al. presents a novel approach to improve speaker-independent lip reading using domain-adversarial training. The proposed model, a speaker-independent lip-reading system, achieves end-to-end sentence-level lip reading by employing a connectionist temporal classification loss and leveraging spatiotemporal convolutions and recurrent operations. The study emphasizes the importance of integrating multimodal features to enhance the performance of audio-visual automatic speech recognition (AV-ASR). The researchers introduce MobiLipNet, a computationally efficient lip reading model that utilizes depth wise and point wise convolutions, making it suitable for real-world applications on resource- constrained devices. The significance of lip reading lies in its ability to aid individuals with hearing impairments in better communication, as human perception of phonemes relies on both auditory information and visual cues from lip and facial movements. Overall, Wand et al.'s [1] research represents a significant advancement in the field of speaker-independent lip reading, contributing to more accurate and adaptable speech recognition technologies that rely on visual features for improved communication and accessibility. Chung and Zisserman [2] present a groundbreaking study on lip reading in the wild, aiming to recognize phrases and sentences spoken by a talking face solely based on visual information. The key contributions include the development of a 'Watch, Listen, Attend and Spell' (WLAS) network that transcribes mouth motion to characters, a curriculum learning strategy for training acceleration, and the creation of the 'Lip Reading Sentences' (LRS) dataset with over 100,000 natural sentences from British television. The WLAS model, trained on the LRS dataset, outperforms previous works on standard lip reading benchmarks significantly, even surpassing professional lip readers on BBC television videos. The study demonstrates the importance of visual information in enhancing speech recognition performance, even when audio is available. The model operates at the character level, incorporates a dual attention mechanism, and can process visual and audio inputs independently or jointly, showcasing remarkable advancements in open-world lip reading. Fenghour et al. [3] present a comprehensive survey on automated lip-reading approaches, with a focus on deep learning methodologies. The authors compare various components of lip-reading systems, including audio-visual databases, feature extraction techniques, classification networks,

ISBN-978-81-19905-39-3

and schemas. One of the key contributions of the survey is a comparison of Convolution Neural Networks (CNNs) with other neural network architectures for feature extraction, highlighting the advantages of CNNs in capturing spatial and temporal information from lip movements. The authors also provide a critical review of the advantages of Attention-Transformers and Temporal Convolution Networks over Recurrent Neural Networks for classification, emphasizing their ability to capture long- range dependencies and parallelize computations. Additionally, the survey compares different classification schemas used for lip-reading, such as ASCII characters, phonemes, and visemes, discussing the trade-offs between accuracy and interpretability. The authors also review the most up-to-date lip-reading systems up until early 2021, showcasing the evolution of these systems from recognizing isolated speech units to decoding entire sentences, thanks to advancements in deep neural networks and the availability of large-scale databases. The survey outlines the stages of automated lip-reading, including preprocessing, feature extraction, and classification, providing valuable insights into the current state of deep learning-based automated lip-reading and highlighting areas for future research and development. Stafylakis and Tzimiropoulos [4] proposed a novel approach for lip reading by combining Residual Networks with Long Short-Term Memory (LSTM) networks. Their end-to-end deep learning architecture for word-level visual speech recognition integrates spatiotemporal convolutional, residual, and bidirectional LSTM networks. The system was trained and evaluated on the challenging Lip reading In-The-Wild benchmark, achieving a word accuracy of 83.0, which is a significant improvement of 6.8 absolute points over the existing state-of-the-art methods. Notably, this performance enhancement was achieved without utilizing information about word boundaries during training or testing. The study showcases the effectiveness of leveraging deep learning techniques, specifically the combination of Residual Networks and LSTMs, in advancing visual speech recognition systems. By pushing the boundaries of audiovisual word recognition, Stafylakis and Tzimiropoulos demonstrate the potential of their approach to enhance the accuracy and efficiency of lip reading technologies, contributing to the evolution of automated speech recognition systems that rely on visual cues. Zhao et al. (2020) introduced a novel approach, Lip by Speech (LIBS), to enhance lip reading by leveraging knowledge from pre-trained speech recognizers. The method distils multi-granularity knowledge, including sequence-level, context-

ISBN-978-81-19905-39-3

level, and frame-level information, from speech recognizers to lip readers. This cross-modal knowledge distillation addresses the challenge of inconsistent lengths of audio and video sequences and refines the speech recognizer's predictions. Experiments on the CMLR and LRS2 datasets demonstrated that LIBS achieves state-of-the-art performance, outperforming baselines by 7.66% and 2.75% in character error rate, respectively. Saliency visualization showed that LIBS improves the lip reader's ability to extract discriminative visual features compared to the baseline. By transferring complementary information from speech recognizers, LIBS significantly advances the accuracy of lip reading, contributing to the development of more effective automated speech recognition systems that rely on visual cues. This work highlights the potential of knowledge distillation in improving lip reading and has implications for applications in speech recognition, particularly in noisy environments or for individuals with hearing impairments. Assael et al.(5) introduced LipNet, the first end-to-end sentence-level lip reading model that simultaneously learns spatiotemporal visual features and a sequence model. LipNet maps a variable-length sequence of video frames to text using spatiotemporal convolutions, a recurrent network, and the connectionist temporal classification loss, trained entirely end-to-end. On the GRID corpus, LipNet achieves 95.2% accuracy in a sentence-level, overlapped speaker split task, outperforming experienced human lip readers and the previous 86.4% word-level state-of-the-art accuracy. The authors argue that human lip reading performance increases for longer words, indicating the importance of features capturing temporal context in an ambiguous communication channel. LipNet's end-to-end architecture allows it to learn visual features directly from pixels, eliminating the need for hand-crafted features. The model's strong performance highlights the potential of deep learning techniques in advancing lip reading technology, which has applications in speech recognition, particularly in noisy environments or for individuals with hearing impairments. By demonstrating the feasibility of sentence-level lip reading, LipNet represents a significant step forward in the field and paves the way for further research and development in this area.

3.0. Proposed Methodology

Lip reading system architectures are crucial for accurately interpreting visual cues from lip movements to recognize spoken words or sentences. These systems typically consist of several key

ISBN-978-81-19905-39-3

components that work together to process visual information effectively. The first step in a lip reading system architecture involves face detection and lip localization, where the system identifies the speaker's face and isolates the lip region from video frames. This initial stage is essential for focusing on the area where lip movements occur, enabling precise analysis of visual cues. Following face detection and lip localization, the system moves on to feature extraction, a critical process where the movements and shapes of the lips are analyzed to extract discriminative visual features. Techniques such as spatiotemporal convolutions, recurrent networks, and attention mechanisms are commonly used to extract relevant information from the visual input. These features play a vital role in capturing the nuances of lip movements that convey speech information. Once the visual features are extracted, the system proceeds to the classification stage, where the extracted features are mapped to text using deep learning models trained end-to-end on extensive datasets. This mapping process is crucial for translating visual information from lip movements into textual representations of spoken words or sentences accurately. The use of deep learning models allows for the system to learn complex patterns and relationships within the visual data, enhancing the accuracy of the lip reading process. In addition to feature extraction and classification, modern lip reading systems often incorporate sequence modelling techniques to capture temporal context and predict entire sentences rather than isolated words. This sequential modelling approach enables the system to understand the flow and structure of spoken language, improving the overall accuracy and robustness of the lip reading process. By integrating these components cohesively, lip reading system architectures can achieve remarkable performance, even surpassing human lip reading capabilities on challenging benchmarks. The combination of advanced feature extraction, classification, and sequence modelling techniques in these architectures demonstrates the potential of deep learning and artificial intelligence in advancing visual speech recognition systems. These systems have significant applications in aiding individuals with hearing impairments, improving speech recognition in noisy environments, and enhancing communication accessibility for a wide range of users.

ISBN-978-81-19905-39-3

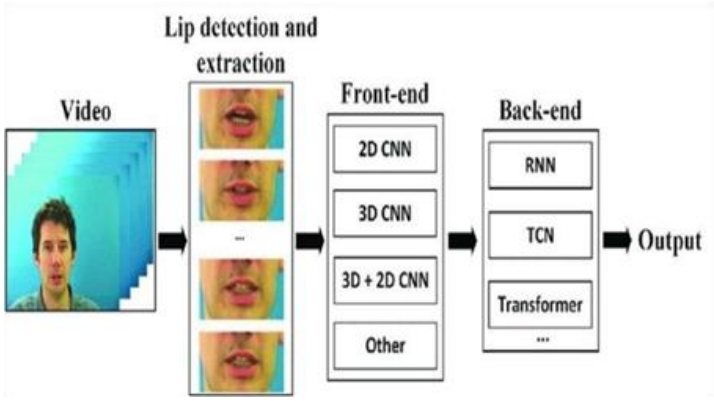


Figure 1: System Architecture

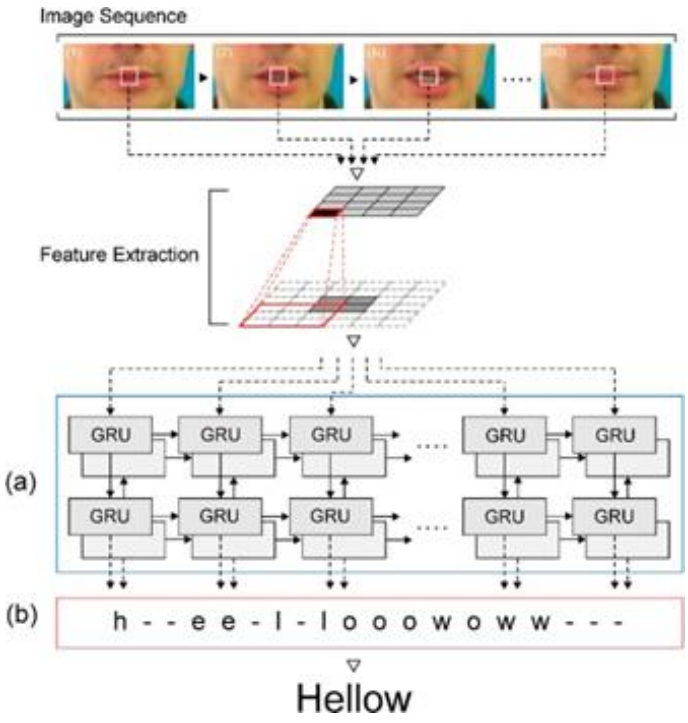


Figure 2: Workflow of facial recognition

3.1. Databases

There are several prerecorded video datasets available for lip reading research and development. The LRW, LRS2, and LRS3 datasets, collected from "in the wild" videos,consist of over 6 million word instances, 800+ hours of footage, and 5,000+ identities. These large-scale datasets provide

ISBN-978-81-19905-39-3

cropped face tracks and corresponding subtitles, with no overlap between the versions. The GRID audio-visual sentence corpus and MIRACL-VC1 dataset are also commonly used for lip reading experiments. During preprocessing, the videos are converted to frames using Open CV, and the lip region is isolated using facial landmark detection with the Dlib library. The cropped lip images are then used as training data for end-to-end lip reading models like LipNet, which uses spatiotemporal convolutions, recurrent networks, and connectionist temporal classification to predict entire sentences from visual input. These prerecorded datasets enable researchers to develop and evaluate advanced lip reading systems, pushing the boundaries of visual speech recognition and aiding applications such as speech recognition for hearing-impaired individuals in noisy environments.

Fig. 3: Dataset

3.2 Lip Detection and Localization

Face detection is typically performed using well-established computer vision techniques, such as Viola-Jones object detection or deep learning-based methods like Faster R-CNN and YOLO. These algorithms scan the input video frames and identify the location of the speaker's face, providing a bounding box that encompasses the facial region. This step ensures that the system can isolate the relevant area of interest and discard irrelevant background information, improving the overall efficiency and accuracy of the lip reading process. Following face detection, the next critical step is lip localization, where the system identifies the specific region of the lips within the detected facial area. This is often

ISBN-978-81-19905-39-3

achieved using facial landmark detection algorithms, which can precisely locate key facial features, including the lips. Commonly used tools for this purpose include the Dlib library, which provides a pre-trained facial landmark detector, and Open CV's built-in facial landmark estimation functionality. By combining face detection and lip localization, the system can accurately isolate the lip region from the video frames, ensuring that the subsequent feature extraction and classification stages focus solely on the visual information that is most relevant for lip reading. This targeted approach helps to minimize the impact of irrelevant visual cues and background noise, enhancing the overall performance of the lip reading system. The accuracy of face detection and lip localization is crucial, as errors in these initial steps can propagate through the entire system, leading to suboptimal feature extraction and classification. Therefore, researchers and developers often invest significant effort in optimizing these components, leveraging the latest advancements in computer vision and deep learning to achieve robust and reliable lip region isolation.

3.3. Feature Extraction

Feature extraction is a crucial step in the lip reading system architecture, as it involves analyzing the movements and shapes of the lips to extract discriminative visual features that can be effectively mapped to textual representations of speech .One of the key techniques used for feature extraction in lip reading is spatiotemporal convolutions. These convolutional neural network (CNN) architectures are designed to capture both the spatial and temporal information present in the lip movements. By applying 3D convolutions, the system can learn to extract features that encode the dynamic changes in the lip shape and position over time, rather than just static image-based features. This allows the model to better understand the temporal context and sequence of lip movements, which is essential for accurate lip reading. In addition to spatiotemporal convolutions, recurrent neural networks (RNNs), such as Long Short-Term Memory (LSTMs) and Gated Recurrent Units (GRUs), have also been widely employed for feature extraction in lip reading systems. These networks are adept at modeling the sequential nature of speech and can effectively capture the long-term dependencies in lip movements. By processing the video frames in a sequential manner, RNNs can learn to extract features that represent the temporal evolution

ISBN-978-81-19905-39-3

of the lip shapes, enabling the system to better understand the context and flow of the spoken language. Furthermore, attention mechanisms have been introduced to enhance the feature extraction process in lip reading. Attention-based models can selectively focus on the most relevant regions of the lips, dynamically weighting the importance of different spatial and temporal features during the extraction process. This allows the system to concentrate on the most informative visual cues, improving the overall accuracy and robustness of the lip reading system. The combination of these techniques, such as spatiotemporal convolutions, recurrent networks, and attention mechanisms, has led to significant advancements in the field of lip reading. By effectively capturing the complex spatiotemporal patterns and dynamics of lip movements, these feature extraction methods have enabled lip reading systems to achieve human parity on challenging benchmarks, demonstrating their effectiveness in translating visual speech information into textual representations .Ongoing research in this area continues to explore novel feature extraction techniques, leveraging the latest developments in deep learning and computer vision to further enhance the performance and capabilities of lip reading systems. As these systems become more accurate and robust, they hold great promise for applications in speech recognition, accessibility for individuals with hearing impairments, and various other domains where visual speech information can complement or supplement traditional audio-based approaches.

ISBN-978-81-19905-39-3

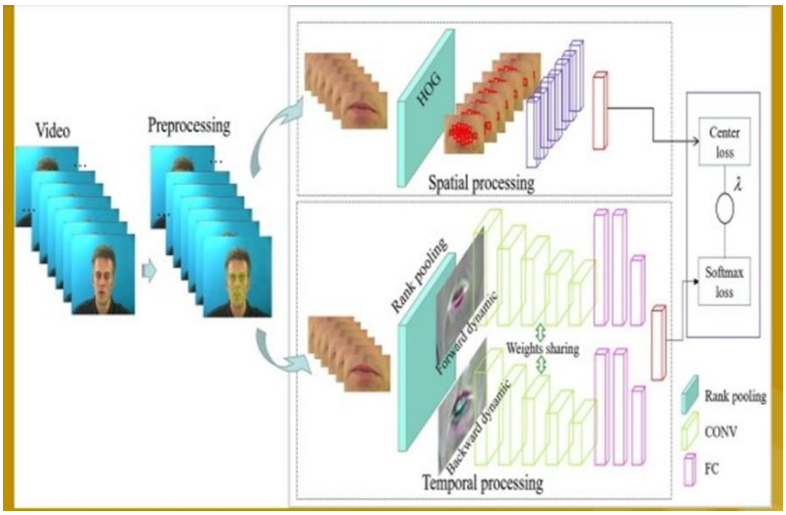


Figure 4: 4: Video Frame Processing

3.4. Classification

The classification stage in a lip reading system involves mapping the extracted visual features to text using deep learning models trained end-to-end on large-scale datasets. Recent advancements in this area leverage knowledge distillation from pre-trained speech recognizers to further improve accuracy. Deep learning models like convolution neural networks (CNNs), recurrent neural networks (RNNs), and transformers have been widely employed for this task. CNNs are effective in capturing spatial and temporal information from lip movements, while RNNs excel at modeling the sequential nature of speech. Attention mechanisms have also been introduced to selectively focus on the most relevant visual cues during classification. By training these models end-to-end on large-scale datasets like LRW, LRS2, and LRS3, which contain millions of word instances and hundreds of hours of footage, lip reading systems can achieve state-of-the-art performance. The use of knowledge distillation from pre-trained speech recognizers further enhances the accuracy of these models by transferring complementary information from the audio domain. The combination of advanced deep learning architectures and large-scale training datasets has led to significant breakthroughs in lip reading, with systems now approaching human parity on challenging benchmarks. As these techniques continue to evolve, they hold great promise for applications in speech recognition,

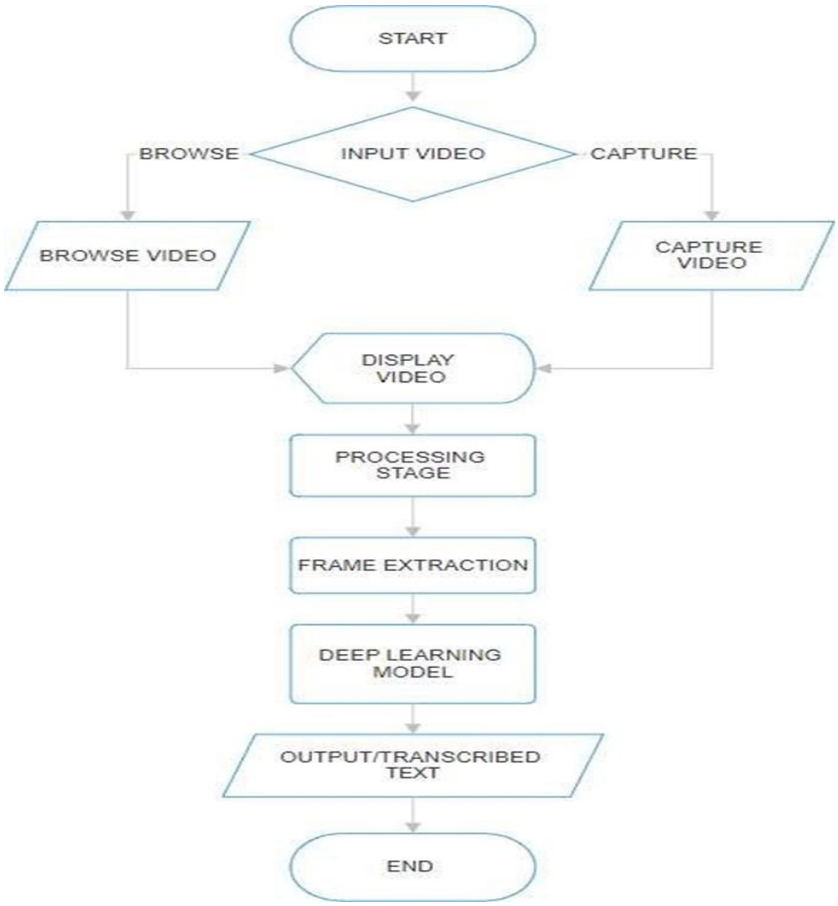
ISBN-978-81-19905-39-3

accessibility for individuals with hearing impairments, and various other domains where visual speech information can complement or supplement traditional audio-based approaches. Training these deep learning models end-to-end on extensive datasets like LRW, LRS2, and LRS3, which contain millions of word instances and hundreds of hours of video footage, enables lip reading systems to achieve state-of-the-art performance levels. Recent advancements in the field have introduced knowledge distillation techniques, where information from pre-trained speech recognizers is transferred to the lip reading models to further enhance accuracy and improve performance. This knowledge distillation process allows the lip reading system to benefit from the complementary information present in the audio domain, leading to more robust and accurate transcription of visual speech cues. By combining sophisticated deep learning with large-scale training datasets and knowledge distillation techniques, lip reading systems have made significant strides towards achieving human-level performance on challenging benchmarks. These advancements not only improve the accuracy and efficiency of lip reading technology but also have broader implications for applications in speech recognition, accessibility for individuals with hearing impairments and various domains where visual speech information can supplement traditional audio-based approach. Lip reading has various applications, including speech recognition for the hearing im- paired, silent speech interfaces, and audio-visual speech enhancement in noisy environments.

3.5. Flow diagram

Flow diagram can be used to understand the flow of the process Flow chart for the lip reading process shows how the flow of operations works. This diagram contains various processes which applied on the set of data sets such as video where we read the lip movements of the person in that video. The diagram can be given as:

ISBN-978-81-19905-39-3



The system flow diagram depicts the sequential flow of actions within the lip reading system. It begins with the user initiating an action, such as capturing a new video or browsing for an existing one. This interaction is facilitated through the User Interface module, which serves as the primary point of interaction between the user and the system. Upon selecting the desired action, the User Interface module directs the request to either the Video Capture or Video Browse component, depending on the user's choice. If the user opts to capture a new video, the Video Capture component is activated, whereas selecting to browse for an existing video trigger the Video Browse component. Both the Video Capture and Video Browse components forward the video data to the Input Module, which serves as the entry point for processing within the system. The Input Module is responsible for handling the received video data, regardless of its source, and prepares it for

ISBN-978-81-19905-39-3

further processing. Once the video data is received and processed by the Input Module, it is passed on to the Processing Module. This module performs the core lip reading task, which involves analyzing the visual information, particularly the movements of the lips, along with any accompanying audio, to transcribe the spoken content in the video. After processing the video data, the Processing Module forwards the results to the Output Module. Here, the processed data is formatted and presented in a suitable manner for output. In the context of lip reading, this typically involves converting the transcribed speech into text format. Finally, the formatted output, such as the transcribed text, is sent back to the User Interface module for display to the user The User Interface module presents the final output to the user, completing the sequential flow of actions within the lip reading system.

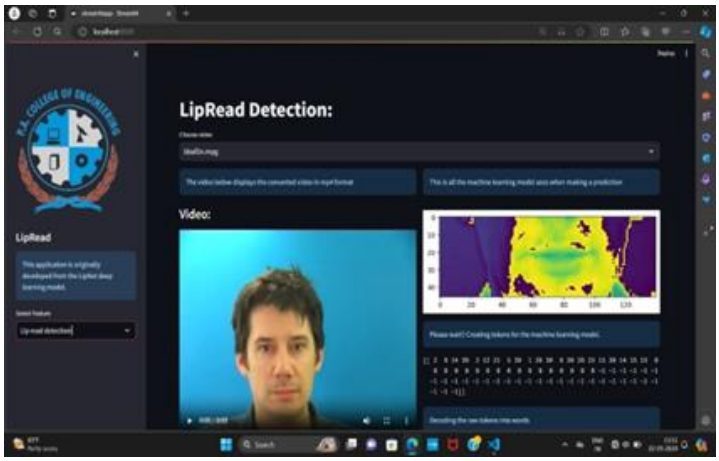
4.0 Results and Discussions

Lip reading, also known as speech reading, is the process of interpreting speech by visually interpreting the movements of the lips, face, and tongue. It is a valuable skill for individuals with hearing impairments or in noisy environments where auditory cues may be difficult to discern. Lip reading systems typically follow a flow diagram that includes several key steps: face detection, lip localization, feature extraction, classification, language modeling, and output. The face detection step involves identifying the speaker's face and locating the lip region. Next, the lip region is isolated from the rest of the face through techniques like lip contour tracking and key point detection. Visual features related to the movements and shapes of the lips are then extracted from the lip region, which can include analyzing lip geometry and appearance-based features. These features are classified using machine learning models like hidden Markov models, neural networks, or convolution neural networks to identify the spoken phonemes or visemes (visual counterparts of phonemes). To improve accuracy, the classified visemes are processed through a language model that incorporates linguistic and contextual information to predict the most likely words or sentences. Finally, the output of the lip reading system is the transcribed text of the spoken words. Recent advancements in deep learning have led to significant improvements in lip reading accuracy. Techniques like efficient-Ghost Net for feature extraction and gated recurrent units (GRUs) for temporal modeling have shown promising results. However, lip reading remains a

ISBN-978-81-19905-39-3

challenging task due to factors such as speaker variability, occlusions, and the inherent ambiguity of some visemes. Lip reading has various applications, including speech recognition for the hearing impaired, silent speech interfaces, and audio-visual speech enhancement in noisy environments. It can also be used for biometric identification based on the unique characteristics of an individual's lips. Despite the progress made in automatic lip reading, human lip readers still outperform current systems in many scenarios. Ongoing research aims to further improve lip reading accuracy, robustness, and real-time performance to make it a more practical and widely adopted technology. Lip reading, also known as speech reading, is a technique that involves understanding speech by visually interpreting the movements of the lips, face, and tongue without sound. It is particularly valuable for individuals with hearing impairments or in noisy environments

where auditory cues may be challenging to discern. Lip reading accuracy varies, with estimates suggesting that it can range from as low as 30%. While lip reading is commonly used by deaf and hard-of-hearing individuals, even those with normal hearing can process some speech information visually. The ability to lip read can be influenced by factors such as context, language knowledge, and residual hearing. Additionally, lip reading can play a crucial role in communication strategies for those with hearing impairment and can support the hearing aid fitting process, especially in improving speech recognition accuracy when combined with hearing aids.



ISBN-978-81-19905-39-3

5. Conclusion

In conclusion, lip reading, or speech reading, is a valuable skill that enables individuals with hearing impairments to better understand spoken language through visual cues. While lip reading accuracy can vary and may not capture all nuances of speech, it remains a crucial communication tool, especially in noisy environments or when auditory information is limited. Combining lip reading with other communication strategies, such as hearing aids, can enhance overall speech recognition and improve communication effectiveness for individuals with hearing challenges. Continued research and advancements in technology, such as automatic lip reading systems using deep learning, hold promise for further improving the accuracy and applicability of lip reading in various settings. Overall, lip reading plays a significant role in empowering individuals with hearing impairments, enhancing their independence, confidence, and communication abilities.

References

[1] Wand, Michael & Schmidhuber, Juergen. (2017). Improving Speaker-Independent Lipreading with Domain-Adversarial Training. 10.48550/arXiv.1708.01565.

[2] Chung, Joon Son & Zisserman, Andrew. (2017). Lip Reading in the Wild. 87-103. 10.1007/978-3-319-54184-6_6.

[3] Fenghour, Souheil & Chen, Daqing & Guo, Kun & Li, Bo & Xiao, Perry. (2021). Deep Learning-Based Automated Lip-Reading: A Survey. IEEE Access. 9. 121184 - 121205. 10.1109/ACCESS.2021.3107946.

[4] Stafylakis, Themos & Tzimiropoulos, Georgios. (2017). Combining Residual Networks with LSTMs for Lipreading. 10.21437/Interspeech.2017-85.

[5] Assael, Yannis & Shillingford, Brendan & Whiteson, Shimon & Freitas, Nando. (2016). LipNet: Sentence-level Lipreading. 10.48550/arXiv.1611.01599

ISBN-978-81-19905-39-3