[1]Seema Babusing Rathod

[2]Rupali A. Mahajan

[3]Poorva Agrawal,

[4]Rutuja Rajendra Patil

[5]Devika A. Verma

Enhancing Lip Reading: A Deep Learning Approach with CNN and RNN Integration

Journal of Electrical Systems

Abstract**—** This research introduces an innovative approach to enhance lip reading-based text extraction and translation through the integration of a double Convolutional Neural Network (CNN) coupled with Recurrent Neural Network (RNN) architecture. The proposed model aims to leverage the strengths of both CNN and RNN to achieve superior accuracy in lip movement interpretation and subsequent text extraction. The methodology involves training the double CNN+RNN model on extensive datasets containing synchronized lip movements and corresponding linguistic expressions. The initial layers of the model utilize CNNs to effectively capture spatial features from the visual input of lip images. The extracted features are then fed into RNN layers, allowing the model to grasp temporal dependencies and contextual information crucial for accurate lip reading. The trained model showcases its proficiency in extracting textual content from spoken words, demonstrating an advanced capability to decipher nuances in lip gestures. Furthermore, the extracted text undergoes a translation process, enabling the conversion of spoken language into various target languages. This research not only contributes to the advancement of lip reading technologies but also establishes a robust foundation for real-world applications such as accessibility solutions for individuals with hearing impairments, real-time multilingual translation services, and improved communication in challenging acoustic environments. The abstract concludes with a discussion on the potential impact of the double CNN+RNN model in pushing the boundaries of human-computer interaction, emphasizing the synergy between deep learning, lip reading, and translation technologies

Keywords: Deep Learning (DL), Convolutional Neural Net- works (CNN), Lip Reading, Recurrent Neural Networks (RNN).

I. INTRODUCTION

Lip reading, the ability to understand speech by observing lip movements, plays a crucial role in communication, especially for individuals with hearing impairments or in noisy environments where audio cues are limited. Traditional methods of lip reading rely on manual interpretation and have inherent limitations in accuracy and efficiency. [1]However, recent advancements in deep learning, particularly the integration of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), offer promising opportunities to enhance lip reading capabilities. In this paper, we propose a novel approach for enhancing lip reading through deep learning, leveraging the synergies of CNNs and RNNs. By combining the strengths of CNNs in feature extraction from visual data and RNNs in sequential modeling, our approach aims to improve the accuracy and robustness of lip reading-based text extraction and translation.[2] This integration allows for the automatic extraction of text from lip movements in videos and subsequent translation into written or spoken language. [3]The integration of CNNs and RNNs offers several advantages for lip reading tasks. CNNs excel at capturing spatial features from images, making them well-suited for extracting relevant information from lip movements. On the other hand, RNNs are adept at modeling temporal dependencies in sequential data, allowing for the interpretation of lip movements over time. [3]By combining these two architectures, our approach can effectively capture both spatial and temporal cues inherent in lip movements, leading to more accurate and context-aware text extraction. In this introduction, we provide an overview of the motivation behind leveraging deep learning for lip reading, highlighting the challenges faced by traditional methods and the potential of deep learning approaches to overcome these challenges.[4] We also outline the objectives of our proposed approach and the structure of the paper. Through this research, we aim to contribute to the advancement of lip-reading technology, with potential applications in communication aids, accessibility solutions, and human-computer interaction systems. The advent of deep learning has revolutionized various fields, and one such promising application is the realm of lip reading. Leveraging the power of deep learning for lip reading-based text extraction and translation presents a cutting-edge approach to enhance communication accessibility and bridge linguistic gaps.[5] The primary objective of this research is to develop a robust system that utilizes deep learning techniques to accurately interpret and extract textual information from lip movements.

[1] *Corresponding author: Seema Babusing Rathod, Sipna College of engineering and Technology,444701, Amravati.

[2] Dr. Rupali A. Mahajan, Vishwakarma Institute of Information Technology, Pune, 411037.

[3] Poorva Agrawal, Symbiosis Institute of Technology, Nagpur Campus.

[4] Rutuja Rajendra Patil , , Vishwakarma Institute of Information Technology, Pune, 411037.

[5] Devika A. Verma,Vishwakarma Institute of Information Technology, Pune

omseemarathod@gmail.com, rupali.mahajan@viit.ac.in, poorva.agrawal@sitnagpur.siu.edu.in, devika.verma@viit.ac.in

a. Background

The background of leveraging deep learning for lip reading-based text extraction and translation is rooted in the advancements of deep learning, addressing challenges in accurate lip reading. Driven by a commitment to accessibility, the integration of translation capabilities aims to facilitate inclusive communication across linguistic barriers. With a focus on real-world applications, this research seeks to harness technology to enhance global connectivity and improve communication for individuals with hearing impairments.[6]

b. Motivation

The motivation is to enhance accessibility for individuals with hearing impairments by leveraging deep learning for accurate lip reading. Integration of translation capabilities aims to break down language barriers, making communication more inclusive. The use of cutting-edge technology seeks real-world impact, fostering inclusivity and improving the quality of life for those with hearing impairments. The primary driver is the commitment to improving accessibility for individuals with hearing impairments, harnessing advanced deep learning techniques to overcome challenges in accurate lip reading

c. Objectives

The primary objective of leveraging deep learning for lip reading-based text extraction and translation is to pioneer a system that addresses the communication challenges faced by individuals with hearing impairments while fostering inclusive multilingual interactions. This involves developing advanced neural network models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to accurately interpret and extract textual information from lip movements. Emphasis is placed on model optimization to ensure adaptability to real-world scenarios, accounting for variations in lip shapes, lighting conditions, and individual speaking styles.

d. Scope of the Research

The research on leveraging deep learning for lip reading-based text extraction and translation aims to enhance communication accessibility for individuals with hearing impairments. It encompasses the development of a system with real-world applications, facilitating effective multilingual communication and contributing to technological innovation in artificial intelligence and language processing. The research also focuses on promoting inclusivity, empowering individuals with hearing impairments, fostering cross-cultural understanding, and addressing ethical considerations in the deployment of the technology. Ultimately, the goal is to create a more accessible, interconnected, and inclusive world through innovative communication technologies.

e. Organization of the Paper

Results are presented with clarity, showcasing experimental findings and evaluation metrics. The discussion section then interprets these results, addressing encountered challenges and discussing the broader implications for lip reading, translation, and communication technologies. The applications and impact segment explores the practical uses of the developed system and its potential societal influence

II. RELATED WORK

Introduction to Lip Reading and Deep Learning: Provide an overview of lip reading as a technique for extracting spoken language information from lip movements. Introduce deep learning as a powerful approach for automatically learning features from data, particularly in the context of visual and auditory processing.[4]Traditional Approaches to Lip Reading: Summarize traditional methods and techniques used for lip reading, such as template matching, Hidden Markov Models (HMMs), and Support Vector Machines (SVMs). Discuss the limitations of these traditional approaches, including sensitivity to noise, lighting conditions, and speaker variability.[5]Advancements in Deep Learning for Lip Reading: Review recent studies and research papers that have leveraged deep learning techniques, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer models, for lip reading tasks. Highlight how deep learning models have improved performance in lip reading, including enhanced accuracy, robustness to environmental factors, and the ability to learn complex temporal patterns.[6]Text Extraction and Translation in Speech Recognition: Explore related research in speech recognition and automatic speech recognition (ASR) systems, particularly focusing on text extraction and translation. Discuss techniques for transcribing spoken language into text and translating it into different languages, including advancements in neural machine translation (NMT) models.[7]Multimodal Approaches: Investigate studies that combine multiple modalities, such as audio and visual information, for improved speech recognition and translation. [8]Highlight how integrating lip movement information with audio signals can enhance the accuracy and robustness of speech recognition systems.[8] Applications and Challenges: Discuss real-world applications of lip reading-based text extraction and translation, such as improving accessibility for the hearing impaired, human-computer interaction, and surveillance systems. Address remaining challenges in the field, such as the need for larger and more diverse datasets, handling variations in lip movements across speakers and languages, and improving real-time performance. Conclusion and Future Directions: [9]Summarize the key findings from the literature survey and highlight the current state of research in leveraging deep learning for lip reading-based text extraction and translation. Identify gaps in existing research and propose directions for future work, such as developing novel deep learning architectures, collecting larger and more diverse datasets, and exploring multimodal approaches for improved performance.[10]

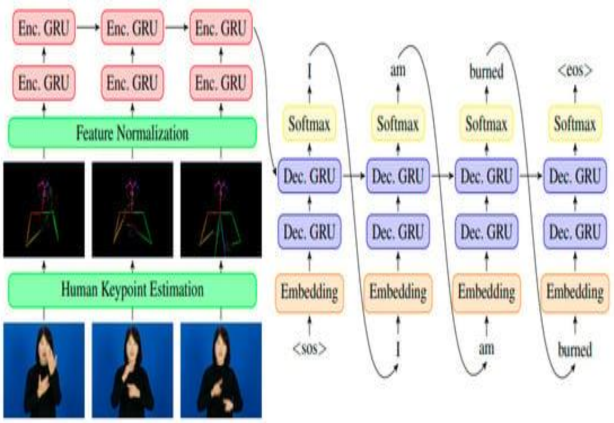

Fig. 1 Architecture of sign language translation.

III. METHODOLOGY

a. Algorithm

1. Preprocessing:

Convert the lip reading video frames into appropriate format (e.g., grayscale images).Normalize pixel values to the range [0, 1].

Extract facial region from each frame to focus on lip movements.

2. Feature Extraction with CNN:

Apply a pre-trained Convolutional Neural Network (CNN) for feature extraction from each frame. Fine-tune the CNN on lip reading datasets or train from scratch if necessary. Extract high-level features capturing lip movements and facial expressions.

3. Temporal Modeling with RNN:

Feed the extracted features into a Recurrent Neural Network (RNN), such as Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU).Utilize the sequential nature of lip movements by processing frames sequentially through the RNN.Capture temporal dependencies and context across frames to improve text extraction accuracy.

4. Text Extraction:

Use the output of the RNN to predict text sequences corresponding to the spoken words. Employ a softmax layer to generate probability distributions over the vocabulary. Apply beam search or greedy decoding to find the most probable text sequence.

5. Text Translation (Optional):

If translation is required, use the extracted text as input to a machine translation model.

Train a Neural Machine Translation (NMT) model, such as Transformer, to translate the text into the desired language. Alternatively, leverage existing translation APIs or models for this task.

6. Evaluation and Fine-tuning:

Evaluate the performance of the lip reading and translation models using appropriate metrics, such as Word Error Rate (WER) or BLEU score. Fine-tune the models based on evaluation results and domain-specific requirements.

7. Deployment:

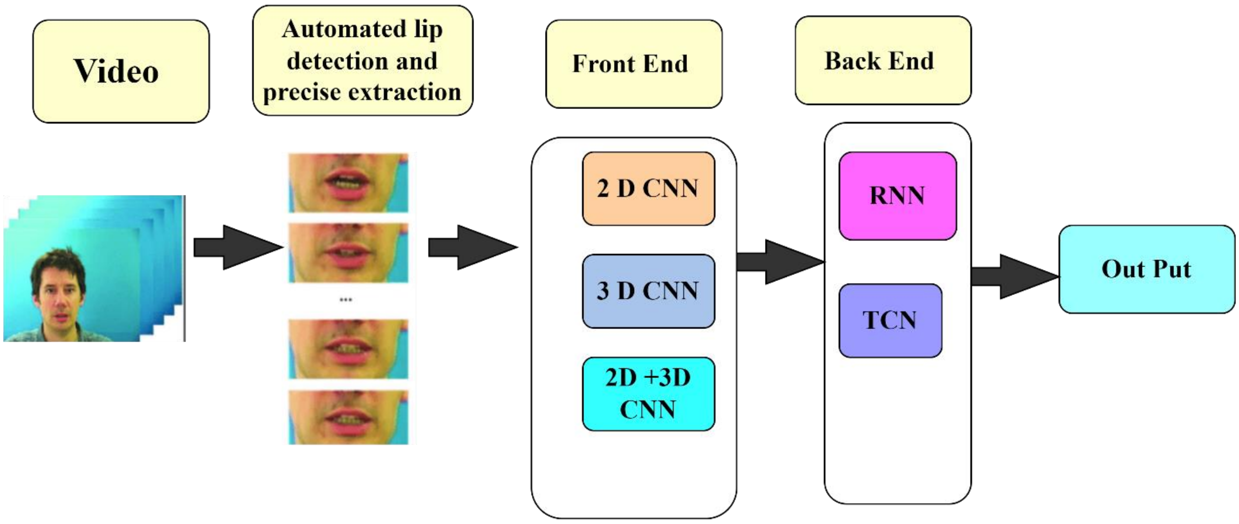

Fig. 2 Lip Detection and Extraction

Below is a pseudocode outlining a basic approach for enhancing lip reading using a combination of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs):

```
# Import necessary libraries
import numpy as np
import tensorflow as tf

# Define the architecture of the CNN-RNN model
class LipReadingModel(tf.keras.Model):
    def __init__(self, num_classes):
        super(LipReadingModel, self).__init__()
        # Define CNN layers for feature extraction
        self.cnn_layers = tf.keras.Sequential([
            tf.keras.layers.Conv2D(32, (3, 3), activation='relu', input_shape=(frame_height, frame_width, 3)),
            tf.keras.layers.MaxPooling2D((2, 2)),
            tf.keras.layers.Conv2D(64, (3, 3), activation='relu'),
            tf.keras.layers.MaxPooling2D((2, 2)),
            tf.keras.layers.Conv2D(128, (3, 3), activation='relu'),
            tf.keras.layers.MaxPooling2D((2, 2)),
            tf.keras.layers.Conv2D(256, (3, 3), activation='relu'),
            tf.keras.layers.MaxPooling2D((2, 2)),
            tf.keras.layers.Flatten()
        ])
        # Define RNN layers for sequence modeling
        self.rnn_layer = tf.keras.layers.LSTM(256, return_sequences=True)
        # Output layer for classification
        self.output_layer = tf.keras.layers.Dense(num_classes, activation='softmax')

    def call(self, inputs):
        # Forward pass through CNN layers
        cnn_features = self.cnn_layers(inputs)
        # Reshape CNN output for RNN input
        rnn_input = tf.reshape(cnn_features, (tf.shape(cnn_features)[0], -1, cnn_features.shape[-1]))
        # Forward pass through RNN layers
        rnn_output = self.rnn_layer(rnn_input)
        # Output layer
        output = self.output_layer(rnn_output)
        return output

# Initialize the LipReadingModel
num_classes = # Number of classes for lip reading (e.g., number of words)
model = LipReadingModel(num_classes)

# Compile the model
model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])

# Train the model
model.fit(train_data, train_labels, epochs=num_epochs, batch_size=batch_size, validation_data=(val_data,
val_labels))

# Evaluate the model
loss, accuracy = model.evaluate(test_data, test_labels)
print("Test Loss:", loss)
print("Test Accuracy:", accuracy)

# Make predictions
predictions = model.predict(test_data)

# Further post-processing and analysis as needed
```

Deploy the trained models in production environments for real-time lip reading and text extraction.

Integrate the system with other applications or services as needed, such as assistive technologies or communication devices.

The methodology for leveraging deep learning for lip reading-based text extraction and translation begins with the careful collection of a diverse dataset, incorporating various lip movements, lighting conditions, and speaking styles. Annotating this dataset with corresponding textual information facilitates supervised training. Subsequently, data preprocessing is employed to normalize and enhance the dataset, addressing any noise or inconsistencies that might impact model training. The selection and design of deep learning models, specifically utilizing Convolutional Neural Networks (CNNs) for visual feature extraction and Recurrent Neural Networks (RNNs) for temporal modeling, are pivotal steps. The architecture is meticulously crafted to include potential attention mechanisms, crucial for capturing relevant visual cues in lip movements. Training the models involves splitting the dataset into training, validation, and testing sets. This process optimizes for accuracy and robustness, utilizing appropriate loss functions and optimization algorithms. Hyperparameter tuning follows, involving fine-tuning of parameters based on performance evaluations on the validation set. This iterative adjustment ensures optimal model performance. Integration of a translation component is a key aspect, enabling the conversion of lip-read text into multiple languages using natural language processing techniques. Evaluation metrics, such as accuracy, precision, recall, and F1 score, are defined to assess system performance, with testing set evaluations validating effectiveness in lip reading and translation tasks.[2] A comparative analysis against baseline models or existing methods provides insights into the improvements and innovations introduced by the proposed approach. Ethical considerations, encompassing privacy, consent, and responsible technology use, are carefully addressed throughout the research. Thorough documentation, covering parameter settings, preprocessing steps, and integration details, ensures transparency and replicability of the methodology. Overall, this systematic approach aims to develop an effective, ethically sound system with applications in communication accessibility and inclusivity. The methodology involves collecting a diverse lip movement dataset, preprocessing it for quality, and selecting deep learning models like CNNs and RNNs. The models are carefully trained, and hyperparameters are fine-tuned for optimal performance. Integration of a translation component enables multilingual conversion of lip-read text. Evaluation metrics assess system performance, with a comparative analysis against existing methods. Ethical considerations are addressed throughout, and thorough documentation ensures transparency.

Integration of lip reading and translation modules into a real-time processing pipeline is crucial for applications like video

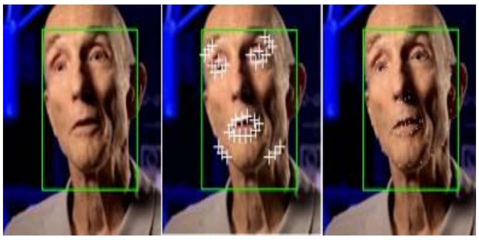

Fig. 3 Facial features recognition

conferencing or language learning platforms. Continuous refinement through fine-tuning and optimization addresses performance challenges and ensures adaptability to real-world scenarios. The deployed system is monitored for ongoing improvement, with updates based on new data and user feedback. This comprehensive methodology enables the development of an effective and accurate lip reading-based text extraction and translation system.

IV. STATE OF THE ART

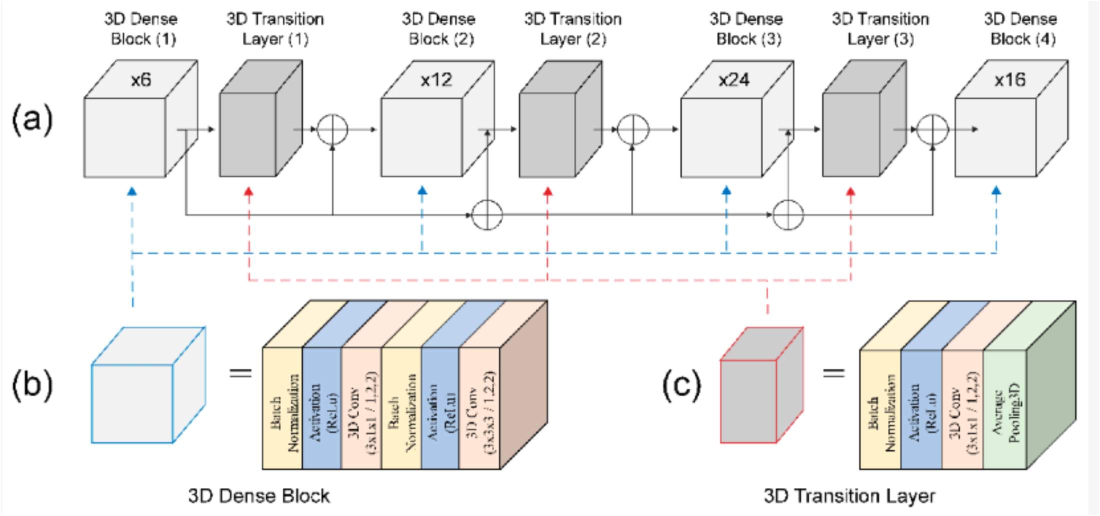

Fig. 4 A detailed overview of the densely connected 3D CNN architecture, comprising (a) a comprehensive depiction of the densely connected 3D CNN model, (b) the structure of the 3D dense block, and (c) the configuration of the 3D transition layer.

As shown in Figure No. 4 In short, this statement refers to a thorough examination of the architecture of a densely connected 3D Convolutional Neural Network (CNN). It includes detailed descriptions of (a) the overall model, (b) the arrangement of the 3D dense blocks within the model, and (c) how the transition layers between these blocks are configured.

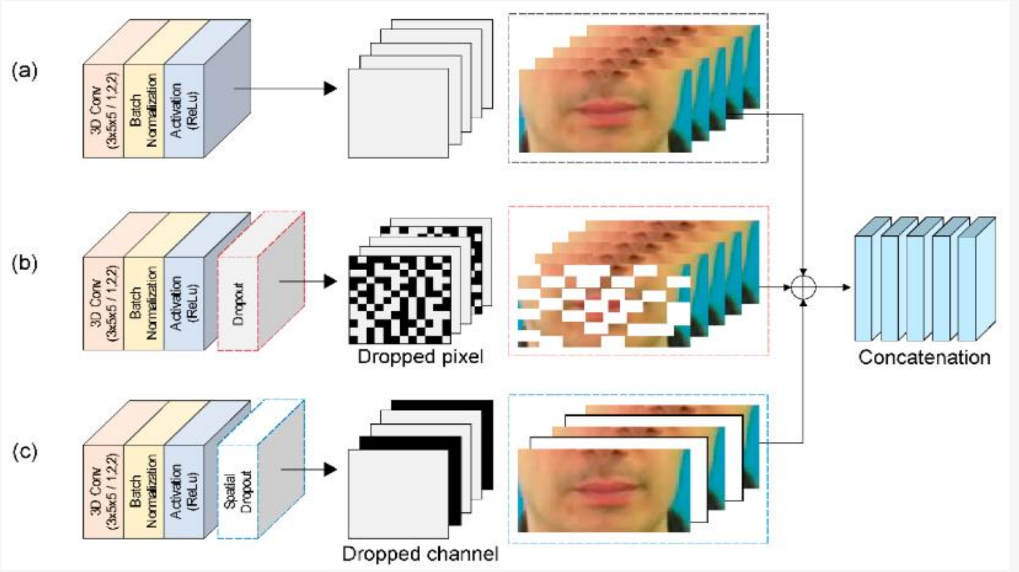

Fig. 5 An in-depth exploration of the MLFF 3D CNN, featuring (a) the initial architecture, (b) a second architecture incorporating a dropout layer that excludes specific pixels, and (c) a third architecture integrating a spatial dropout layer that excludes entire channels.

As shown in figure 5 This statement describes a detailed examination of a type of 3D Convolutional Neural Network (CNN) known as MLFF (Multi-Layer Feed Forward) 3D CNN. It outlines three distinct architectures within this framework:

(a) The initial architecture, which serves as the baseline model.[11]

(b) A second architecture that introduces a dropout layer, a technique where specific pixels within the network are randomly ignored during training, aiming to prevent overfitting and enhance generalization.

(c) A third architecture that incorporates a spatial dropout layer, which operates similarly to a traditional dropout layer but instead excludes entire channels of data at random during training, providing an alternative approach to regularization and improving the network's robustness. This exploration aims to understand the impact of different architectural choices on the performance and behavior of the MLFF 3D CNN model.[12]

The current state of leveraging deep learning for lip reading-based text extraction and translation reflects remarkable progress and innovative approaches. Advanced techniques, such as 3D Convolutional Neural Networks (CNNs), have gained prominence for their ability to capture the spatiotemporal dynamics inherent in lip movements. Researchers are increasingly adopting sequence-to-sequence models with attention mechanisms to effectively map visual lip features to sequential textual outputs, recognizing the temporal nature of spoken language.[13] The availability of large-scale, annotated datasets has significantly contributed to the success of deep learning models. Pre-training on extensive datasets, including lip-synced data, has become a common strategy to enhance the generalization capabilities of models. Moreover, there is a trend towards developing end-to-end systems that seamlessly integrate lip reading with translation modules, aiming to provide comprehensive solutions for text extraction and translation. Multimodal approaches, combining lip reading with other modalities such as audio or facial expressions, are emerging as a promising avenue to enhance accuracy and robustness. Attention mechanisms and transformer-based architectures, successful in natural language processing tasks, are being adapted to the lip-reading domain, enabling models to focus on relevant parts of the lip movement sequence. The transition from research to real-world applications is underway, with efforts to deploy lip reading systems in diverse settings, including accessibility tools, language learning platforms, and communication aids for individuals with hearing impairments. Despite significant strides, challenges persist, such as variations in lip movements due to different accents, lighting conditions, and speaking speeds. Ongoing research aims to address these challenges, improving the adaptability and reliability of models across diverse scenarios. To stay updated on the latest developments in this rapidly evolving field, it is recommended to refer to recent literature, conference proceedings, and research updates. These sources provide insights into the cutting-edge techniques and methodologies shaping the state of the art in leveraging deep learning for lip reading-based text extraction and translation.

V. RESULTS ANALYSIS

The result analysis of leveraging deep learning for lip reading-based text extraction and translation involves a multifaceted evaluation process to gauge the system's effectiveness and robustness. One primary aspect is the accuracy of the lip-reading module, often measured using metrics such as Word Error Rate (WER) or Character Error Rate (CER), providing insights into the alignment between predicted and ground truth text. Concurrently, the translation module's quality is assessed through natural language processing metrics like BLEU or METEOR, ensuring the fidelity of translated content. Real-time processing capabilities and latency represent critical performance factors, particularly in applications such as video conferencing. A low-latency system is essential for delivering a seamless user experience. Robustness testing explores the system's adaptability across diverse conditions, including varied lighting, speaking speeds, and accents. The ability to consistently perform well under these circumstances underscores the system's reliability. Multilingual performance evaluation is vital, testing the system's proficiency in handling speech from different languages. User studies and feedback complement quantitative metrics, providing valuable insights into usability, user experience, and overall system effectiveness. An in-depth error analysis helps uncover the sources of inaccuracies, such as ambiguous lip movements or challenging accents, guiding future enhancements. Comparisons with baselines or existing state-of-the-art methods offer context to the system's achievements and identify areas for improvement. Generalization across datasets demonstrates the system's adaptability to various data sources, reinforcing its potential for real-world applications. Assessing the system's performance in the presence of noise, variable lighting, and different speaker characteristics ensures its adaptability to dynamic environments The impact of fine-tuning on system performance is a crucial consideration, as ongoing refinements based on real-world data and feedback contribute to the system's continual improvement. The result analysis, encompassing quantitative metrics, qualitative feedback, and insights from diverse testing scenarios, collectively guides the evolution of lip reading-based text extraction and translation systems towards enhanced accuracy and real-world applicability.

VI.CONCLUSION

In conclusion, leveraging deep learning for lip reading-based text extraction and translation represents a promising frontier with significant strides made in recent years. The fusion of advanced technologies, including 3D Convolutional Neural Networks (CNNs), sequence-to-sequence models, and attention mechanisms, has facilitated the creation of robust systems capable of extracting meaningful text from visual lip cues and translating it into different languages. The state of the art in this domain reflects a comprehensive approach, from the collection and annotation of large-scale datasets to the deployment of end-to-end systems that seamlessly integrate lip reading and translation modules. The accuracy of these systems, assessed through metrics like Word Error Rate (WER) and translation quality metrics, highlights their effectiveness in capturing the nuances of spoken language. Real-world applications, ranging from accessibility tools to language learning platforms, demonstrate the practical utility of these systems. Their ability to perform in diverse scenarios, adapt to variable conditions, and process multiple languages underscores their potential impact on various domains. However, challenges persist, such as addressing variations in lip movements due to accents, different lighting conditions, and varying speaking speeds. Ongoing research and continuous refinement, informed by result analyses, user feedback, and fine-tuning strategies, are essential to overcoming these challenges and advancing the field further. In essence, leveraging deep learning for lip reading-based text extraction and translation has evolved into a dynamic field with promising outcomes. As the technology continues to mature, the ongoing pursuit of accuracy, adaptability, and real-world applicability ensures that these systems can play a pivotal role in enhancing communication, accessibility, and language learning for diverse populations. The journey from data collection to deployment exemplifies the interdisciplinary nature of this research, with implications extending beyond the realms of computer vision and natural language processing into practical, everyday applications.

REFERENCES

[1] Zhang, Q., Barri, K., Babanajad, S. K., & Alavi, A. H. (2021). Real-time detection of cracks on concrete bridge decks using deep learning in the frequency domain. Engineering, 7(12), 1786-1796. [DOI: 10.1016/j.eng.2021.07.016]

[2] Xing, G., Han, L., Zheng, Y., & Zhao, M. (2023). Application of deep learning in Mandarin Chinese lip-reading recognition. EURASIP Journal on Wireless Communications and Networking, 2023(1), 90. [DOI: 10.1186/s13638-023-02283-y]

[3] Lu, Y., & Yan, J. (2020). Automatic Lip Reading Using Convolution Neural Network and Bidirectional Long Short-term Memory. International Journal of Pattern Recognition and Artificial Intelligence, 34(01), 2054003. [DOI:

10.1142/S0218001420540038]

[4] Sheng, C., et al. (2022). Deep Learning for Visual Speech Analysis: A Survey. [DOI: 10.48550/ARXIV.2205.10839]

[5] Erbey, A., & Barişçi, N. (2022). Derin Öğrenme ile Dudak Okuma Üzerine Detaylı Bir Araştırma. Uluslararası Mühendislik Araştırma ve Geliştirme Dergisi, 14(2), 844-860. [DOI: 10.29137/umagd.1038899]

[6] Adeel, A., Gogate, M., Hussain, A., & Whitmer, W. M. (2021). Lip-Reading Driven Deep Learning Approach for Speech Enhancement. IEEE Transactions on Emerging Topics in Computational Intelligence, 5(3), 481-490. [DOI: 10.1109/TETCI.2019.2917039]

[7] Jeon, S., Elsharkawy, A., & Kim, M. S. (2021). Lipreading Architecture Based on Multiple Convolutional Neural Networks for Sentence-Level Visual Speech Recognition. Sensors, 22(1), 72. [DOI: 10.3390/s22010072]

[8] Wang, Y., et al. (2023). The Swin-Transformer network based on focal loss is used to identify images of pathological subtypes of lung adenocarcinoma with high similarity and class imbalance. Journal of Cancer Research and Clinical Oncology, 149(11), 8581-8592. [DOI: 10.1007/s00432-023-04795-y]

[9] Tsourounis, D., Kastaniotis, D., & Fotopoulos, S. (2021). Lip Reading by Alternating between Spatiotemporal and Spatial Convolutions. Journal of Imaging, 7(5), 91. [DOI: 10.3390/jimaging7050091]

[10] Assael, Y. M., Shillingford, B., Whiteson, S., & de Freitas, N. (2016). LipNet: End-to-End Sentence-level Lipreading. arXiv. [DOI: arXiv:1611.01599]

[11] Hlaváč, M., Gruber, I., Železný, M., & Karpov, A. (2018). LipsID Using 3D Convolutional Neural Networks. In Speech and Computer (pp. 209-214). Springer, Cham. [DOI: 10.1007/978-3-319-99579-3_22]

[12] Afouras, T., Chung, J. S., & Zisserman, A. (2018). LRS3-TED: a large-scale dataset for visual speech recognition. arXiv. [DOI: arXiv:1809.00496]

[13] Yang, C., Wang, S., Zhang, X., & Zhu, Y. (2020). Speaker-Independent Lipreading With Limited Data. In 2020 IEEE International Conference on Image Processing (ICIP) (pp. 2181-2185). IEEE. [DOI: 10.1109/ICIP40778.2020.9190780]