Evaluate Lip reading using Deep Learning Techniques.

Abstract

Silent Sound Technology represents a cutting-edge approach to understanding spoken language without the use of audio, relying instead on the visual cues provided by lip, mouth, and facial movements. The complexity of this task is compounded by the variability in speech patterns and articulation among individuals. To address this formidable challenge, this project leverages the power of machine learning, specifically deep learning and neural networks, to create an automatic lip reading system. Multiple convolutional neural network (CNN) models are meticulously trained on a dataset, and their collective insights are amalgamated into a custom CNN architecture. The effectiveness of these trained models in predicting words using deep learning is rigorously assessed, and the top-performing algorithm is implemented in a web application for real-time word prediction. This achievement not only showcases the potential of machine learning in the realm of Silent Sound Technology but also paves the way for broader applications, underscoring the transformative power of this innovative approach.

Keywords: Convolution Neural Network (CNN), Deep Learning (DL), Image Processing, Silent Sound, Model Concatenation.

1 Introduction

In the ever-evolving landscape of technological innovation, one topic has recently emerged as both a challenging conundrum and a beacon of hope for those seeking to bridge the gap between silent communication and vocal expression: Silent Sound Technology. This intriguing field has posed intricate dilemmas even for the most seasoned lip readers, presenting a complex puzzle that promises to be unraveled through the application of various deep learning methodologies. Silent Sound Technology, in all its complexity, stands as a skill with profound implications, offering a range of salient benefits that hold the potential to reshape the way we interact with and understand the world.

At its core, the advancement of Silent Sound Technology offers a tantalizing prospect—a world where speech recognition thrives amidst the cacophony of noisy or deafening environments. Imagine a reality where clarity emerges from chaos, where conversations in bustling urban streets or thunderous concert venues become comprehensible with ease. One of the standout promises of this technology lies in the realm

of hearing aids, where it could herald a new era of sophistication, rendering unprecedented assistance to those with hearing disabilities. The power to enhance the auditory experience for individuals who have long been deprived of it is an awe-inspiring prospect that silently beckons on the horizon.

But the transformative potential of Silent Sound Technology extends beyond personal enrichment. In the realm of security and surveillance, this technology offers a crucial tool for deciphering and predicting information when audio is either corrupted or absent in video recordings. Imagine the implications for law enforcement and national security when the unspoken words of individuals caught in compromising situations can be unveiled through the visual cues of their lips. It is a capability that could reshape the landscape of investigation and intelligence gathering, potentially preventing and solving crimes with unprecedented efficiency.

Yet, the journey to harness the potential of Silent Sound Technology is not without its formidable challenges. The rich tapestry of languages spoken across the globe, each with its unique diction and articulation, presents a substantial hurdle. Even the most skilled lip readers can only approximate every second word, highlighting the complexity of the task at hand. In response, enterprising minds have turned to the arsenal of neural networks and deep learning algorithms to pave the way forward. This endeavor has led to the creation and evaluation of novel model architectures, each meticulously designed to decode the intricacies of lip movement and spoken language. Through rigorous evaluation, the path to enhanced accuracy emerged, setting the stage for the development of a groundbreaking Realtime lip-reading system.

In essence, the primary mission of our paper is rooted in compassion and empowerment. We seek to empower those who are unable to speak but yearn to find their voice and those who wish to communicate seamlessly, even in the midst of a bustling crowd. As we delve into the intricacies of Silent Sound Technology and its intersection with deep learning, we embark on a journey that promises to unlock new dimensions of human connection, understanding, and inclusion. Join us as we explore the boundless potential of Silent Sound Technology, where the unspoken finds a powerful voice, and silence is transformed into eloquence.

2 Related work

This project [1] enables speech-impaired individuals to communicate with others through lip movement analysis. It employs OpenCV for face and lip tracking, utilizing LipNet and deep learning tools for accurate conversion to text and voice. This project used pre-trained LipNet system for Lip geometry-based feature extraction but it does not used any own CNN architecture or model for training.

This survey [2] is based on the silent sound using electromyography and image processing. Silent sound technology uses Electromyography to measure facial muscle activity, converting electrical pulses into speech for transmission. The electrical activity of the face muscle tissues is measured by this. The process of converting digital data tape into a film image with corrections and calibrations is done through two ways: analog image processing and digital image processing. A 99% efficiency is claimed for silent sound technology by engineers. It is designed to aid those who have lost their

voices, observing lip movements to reduce noise pollution during calls. The technology can be implemented on earphones, mobile phones, and headsets to convert muscle movements into electrical signals and speech. But it's limited only three languages: English, German, and French. It cannot be used in languages like Chinese, where accent or voice modulation determines word meanings and also emotions cannot be conveyed through the speech. The system works only when fine wired electrodes are attached to the face.

The study [3] shows that using a session-independent system for EMG-based speech recognition improves robustness and copes with various speech conditions, including silently articulated speech. A large vocabulary of over 2000 words is achieved. The study utilizes a graphical GUI to display word relationships. Comparing "Hello vs Baseball" and "Hello vs Hello" graphs, it finds higher similarity within identical words. During the development phase, the need for several improvements has been identified, including the necessity of adding more electrodes for precision, the implementation of Artificial Neural Networks for future-proofing and the enhancement of signal classification accuracy, as well as the enabling of IoT device control for greater convenience.

[4] The model used Fine-Tuned VGG+LSTM baseline. Helpfulness of data augmentation was observed solely in the case of unseen individuals. The baseline outperformed the LSTM+CNN architecture. Validation accuracy very close to 75% was achieved, with a test accuracy of 59%. But the model is used pre-trained model VGG, here also no own CNN architecture or model is used for training.

This study [5] outlines an architecture for predicting spoken sentences from silent videos of talking faces. It involves data preprocessing, a spatial-temporal visual frontend, a viseme classifier, and a sentence prediction module. Performance is evaluated based on predicted sentences compared to ground truth using edit distance. Further sections elaborate on each component. The research utilizes the BBC LRS2 dataset, comprising around 46,000 videos, 2 million word instances, and a vocabulary of over 40,000 words. Videos have a maximum length of 180 frames at 25 frames per second, featuring sentences up to 100 ASCII characters. The dataset presents challenges with diverse viewpoints, lighting, genres, and numerous speakers. The results reveal a significant reduction in Word Error Rate (WER) to 35.4%, a 15% improvement compared to the previous state-of-the-art model on the same dataset. Word accuracy increased to 64.6%. Viseme accuracy (VER) was notably high at 4.6%.

Paper [6] utilizes surface electrodes to capture muscle activity for speech recognition. The paper outlines a data processing strategy for speech recognition. It involves raw signal pre-processing to eliminate noise, utilizing Short Time Speech Energy (STE) for Voice Activity Detection (VAD), and extracting power spectrum features for classification after speech segment detection. In this study, a Hold-out method split data into 7-3 training and test sets. Employing a support vector machine (SVM) classifier, an average accuracy of 92.3% was achieved, with a maximum accuracy of 100%. But the Limitation is Small participant sample size may have compromised the model's overall robustness and generalizability.

This paper [7] presents an LSTM-based model utilizing visemes as input to predict spoken words from a limited dataset. The results are satisfactory, even though

the individual visemes were pre-known. This approach demonstrates the potential of LSTM in speech recognition and highlights its effectiveness in a constrained dataset, laying the groundwork for further research in this area.

This survey [8] explores automated lip-reading, particularly within the realm of deep learning. It stands out by delving into a detailed comparison of alternative frontend networks like feedforward neural networks and autoencoders, alongside CNNs. For classification, it emphasizes architectures involving Attention-Transformers and TCNs, known for their advantages over RNNs. Additionally, it examines various classification schemas used in lip-reading, offering insights into the latest approaches from late 2020 and early 2021.

The authors [9] developed a Kannada speech recognition system to handle continuous speech in various noisy conditions. The paper utilized approximately 2,400 speaker datasets and employed the Kaldi toolkit for recognition modeling at different phoneme stages. Their approach involved using 80% of the data for training and 20% for testing through Kaldi.

This paper [10] introduces a approach for visual speech recognition, combining lip shape and lip actions at both feature and model levels. Their method demonstrates significant performance improvements, with a 85% enhancement for Motion History Image-discrete wavelet transform features, 74% for Motion History Image-discrete cosine transform, and 80% for Motion History Image-Zernike moments-based features. This research marks notable progress in visual speech recognition techniques. While the study presents promising results in visual speech recognition (VSR), there are some limitations and potential drawbacks to consider like Limited Dataset, Feature Complexity, Model-Level Fusion Complexity, Real-World Noise.

This study [11] introduces a visual speech recognition approach using LSTM with a minimal dataset. The authors recommend future work on feature extraction and classification stages. The method employs a two-stream approach, with the first stream focusing on feature extraction from mouth regions and the second stream extracting features from altered images. This research explores efficient ways to perform visual speech recognition with limited data resources.

This paper [12] employs a trial equation approach to investigate the recently proposed concatenation model, which combines the nonlinear Schrödinger's equation, Lakshmanan–Porsezian–Daniel model, and Sasa–Satsuma equation. The study reveals various solutions, including dark solitons, singular solitons, cnoidal waves, and singular periodic waves. This approach proves effective in recovering a broad range of solutions for the model, and numerical simulations provide visual representations of these analytical findings.

With an emphasis on helping speech-impaired people and improving communication technologies, these studies and initiatives present outstanding developments in speech and visual speech recognition. These developments incorporate numerous methods using huge datasets, ranging from deep learning to electromyography. Despite language barriers and electrode issues, electromyography shows potential for wordless communication. Even with little data, deep learning techniques like LSTM and CNN topologies dramatically enhance voice recognition. Reduced error rates and increased viseme accuracy are achieved through visual speech recognition. Our comprehension

is enhanced through the investigation of neural network topologies and feature extraction. These advancements open the way for future study, delivering possible benefits to people with speech difficulties and expanding communication technologies for everyone while acknowledging limitations like language restrictions and data size.

3 Data Acquisition and Pre-Processing

3.1 Dataset:

This dataset consists of thousands of short spoken sentences by a male speaker sourced from LRW, GRID, LRS2, LRS3-TED, VoxCeleb2, CAS-VSR-W1k(LRW-1000), and the Lip Language dataset. Sentences are up to 10 characters long, and data is partitioned into training, validation, and test sets based on dataset quality.

3.2 Video Data Loading:

In our data preprocessing pipeline, video data is first sourced from a predetermined path, laying the foundation for subsequent analysis. To effectively handle this multimedia data, we employ the OpenCV library, a robust tool for reading video frames. To streamline the dataset and reduce its dimensionality, we opt to convert the frames into grayscale, leveraging the capabilities of the TensorFlow library. Furthermore, to hone in on the most pertinent visual information, we implement a cropping mechanism, carefully selecting a region of interest within each frame. This strategic cropping allows us to concentrate on the core content of the video, ensuring that our subsequent analysis and model training are both efficient and focused.

Fig. 1: Lip Data

3.3 Allignment Information:

A critical component in many multimedia applications, alignment information is essential for creating a meaningful relationship between textual content and video frames. This temporal alignment information was painstakingly extracted from text files, bridging the time gap between the written narrative and the matching video parts. The alignment information goes through a careful modification to provide room for additional analysis and modeling. It is tokenized, which reduces it to a list of individual characters, each of which stands for a different alignment component. These characters are also converted into a numerical representation using a character-to-number lookup technique. This numerical representation serves as the basis for additional data

processing and model training, making it possible to seamlessly combine textual and visual data for deeper insights and applications.

4 Proposed Methodology

In this section, we describe the methodology employed for preprocessing video data, aligning it with textual content, and training a neural network model for a specific task. The proposed methodology encompasses data loading, data preprocessing, and model training.

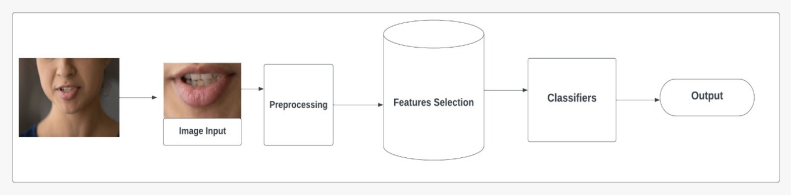

Fig. 2: Process Flow

4.1 Data Preparation:

4.1.1 Data Loading Pipeline:

Utilizing the flexible 'tf.data.Dataset' of TensorFlow, we have painstakingly designed a strong data loading and preparation pipeline for our technique. The systematic retrieval of video data files from a designated directory at the beginning ensures the smooth integration of various data sources. Data is carefully mixed to eliminate any potential bias and provide a little of unpredictability that makes the model more flexible. Additionally, our unique "mappable function" is skillfully implemented in parallel to maximize performance, effectively managing data loading and preprocessing duties. We specifically address the issue of changing sequence lengths by carefully batching the data and ensuring uniform padding. Lastly, preprocessed data is effectively prefetched to speed up training, adding to the overall streamlined and high-performance training process.

4.2 Feature Extraction:

4.2.1 Convolutional Neural Network:

We use Convolutional 3D layers as a key element of our model architecture to extract features from video data. The complicated patterns comprising both spatial and temporal dimensions are captured inside the video frames thanks to these specialized layers. These layers are perfectly suited for video data processing because they employ convolutional filters throughout the three-dimensional space of the movie to distinguish both temporal dynamics and spatial structures. The resulting extracted features serve as a rich basis for later stages of our neural network model by encapsulating

important information about the change of visual material over time. In order for our model to understand and interpret the intricate interaction of visual features, this feature extraction method is essential. As a result, our model is successful.

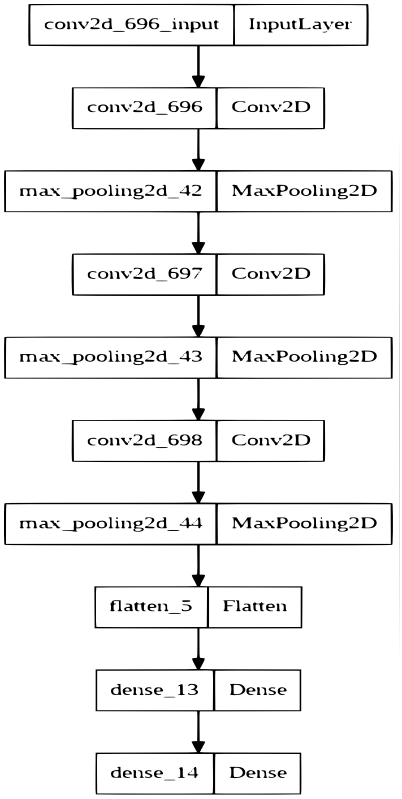

Fig. 3: Propose CNN Model

4.2.2 Bi-directional LSTM:

In our method, we intentionally incorporate Bidirectional Long Short-Term Memory (LSTM) layers into our neural network design to harness their potential. In tasks requiring temporal data, such as those involving video sequences, these bidirectional LSTM layers are very important in capturing complex sequential relationships within our data. LSTMs are capable of retaining long-range relationships and complex patterns in the data and may be thought of as dynamic feature extractors. Our model becomes extremely well-suited for tasks that call for a thorough comprehension of temporal relationships by implementing the bidirectional variant, which gives it the capacity to not only comprehend past context but also foresee future context. This feature extraction approach considerably improves our model's ability to recognize complex temporal patterns and improves model performance as a whole.

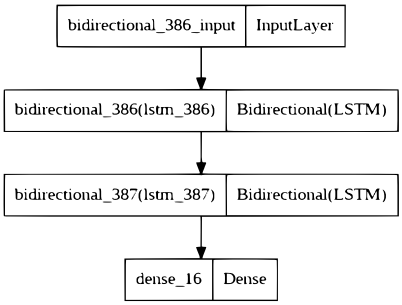

Fig. 4: Propose Bi-Directional LSTM architecture

4.2.3 Max Pooling:

In our technique, a key step in the feature extraction process is the use of Max-Pooling layers. These layers are crucial in reducing the spatial dimensions of the derived feature maps, which allows us to keep the most important details while successfully reducing computational complexity. By choosing the highest value possible within a predetermined frame, max-pooling preserves the key characteristics of each local area. By doing this, we are able to accomplish two key goals: effectively reducing the number of dimensions, which speeds up computation, and maintaining key characteristics, which allows our neural network to concentrate on the most important portions of the input. Our feature extraction pipeline's overall efficiency and efficacy are increased because to the clever integration of Max-Pooling layers.

4.2.4 Flattening and Time Distributed Operations:

The implementation of the TimeDistributed(Flatten()) layer is a crucial stage in our methodology's feature extraction process. This technique is essential for getting the retrieved features ready for additional examination and later processing in our neural network model. We effectively flatten the retrieved features throughout the time dimension by integrating the TimeDistributed wrapper over the Flatten() layer, aligning them for seamless incorporation into the next dense layers of our network. This step is crucial in getting the model to understand the multi-dimensional representations of our video data, making it easier to extract complex patterns and temporal relationships, and ultimately improving the model's ability to make accurate predictions based on the video-textual alignment.

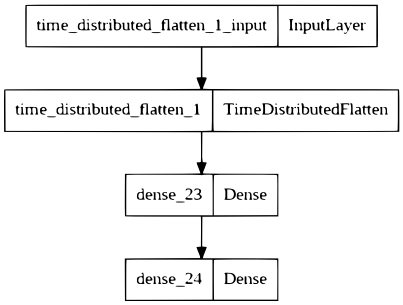

Fig. 5: Time Distributed Flatten

4.3 Classifiers:

A powerful deep neural network model serves as the main engine of our study and is what moves our research forward. By learning and anticipating complex sequences and patterns, this neural network forms the basis for eventually matching video data with written information. Convolutional 3D layers skilled at feature extraction, LSTM layers designed for sequence modeling, and Dense layers charged with producing exact predictions make up the model's precisely created architecture. In our effort to close the gap between visual and textual information in our study project, the sophisticated architecture of our neural network classifier is carefully described inside the code we have given.

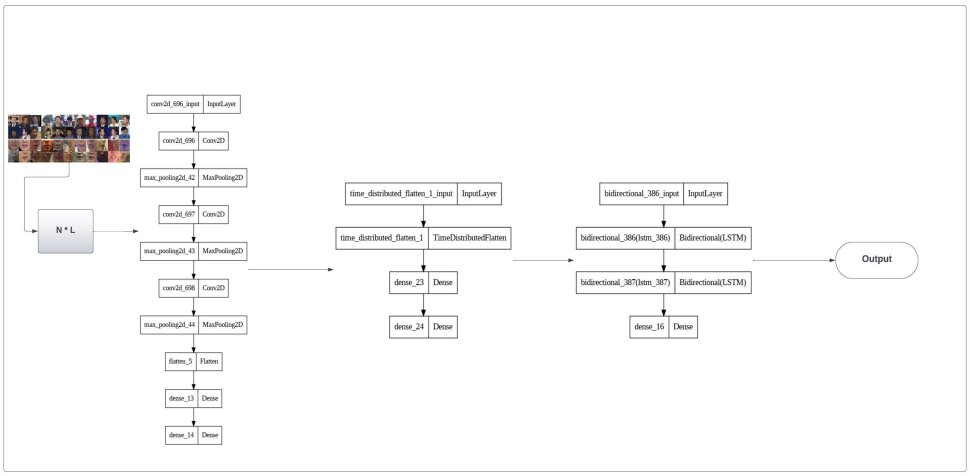

Fig. 6: The deep learning model architecture as classifier

5 Performence Evaluation

The study evaluates the performance of a model architecture for Automatic Speech Recognition (ASR) using word error rate (WER) and character error rate (CER) metrics, commonly employed in ASR assessment. The model employs CTC beam search to generate predictions. Notably, WER and CER measure the minimum number of insertions, substitutions, and deletions required to align model predictions with ground truth, normalized by the respective word or character counts.

Comparing the proposed model with baseline approaches, it demonstrates improved accuracy due to its structured sentence format and limited word subset from the GRID corpus, enabling context utilization. On an unseen speakers dataset, the hearing-impaired participants achieved WERs of 87.3%, 80.4%, and 75.5%, averaging at 81.0667%. Remarkably, the study notes that human lipreaders typically attain an 80% accuracy rate. This suggests promising progress in ASR technology, particularly for the hearing-impaired, although challenges like unseen speakers still exist.

Table 1: LipNet Performance on GRID Dataset Compared to Baselines - (a) Evaluation on Unseen Speakers - (b) Evaluation on 555 Video Subset of Each Speaker's Sentences

Methods	**Unseen CER**	**Speakers WER**
CNN+LSTM	55.09%	62.9%
LipNet+CNN+Bi-LSTM	77.3%	81.067%

Table 1: Sample Table

5.1 Confusion Matrix:

In the context of lip reading, a confusion matrix is a vital evaluation tool used to assess the performance of a deep learning model, specifically one based on Convolutional Neural Networks (CNNs) concatenated together. A confusion matrix visually represents the model's ability to correctly classify lip movements and speech-related information. It breaks down the predictions into categories such as true positives, true negatives, false positives, and false negatives. This matrix helps researchers and practitioners gauge the model's accuracy, precision, recall, and F1-score, enabling them to fine-tune the model and improve its lip reading capabilities. Analyzing the confusion matrix is an essential step in the development and evaluation of lip reading deep learning models using CNN concatenation. Below, you can find our confusion matrix (Fig 7), which helps us assess the model's accuracy, precision, recall, and F1-score, allowing us to refine and enhance its lip reading capabilities.

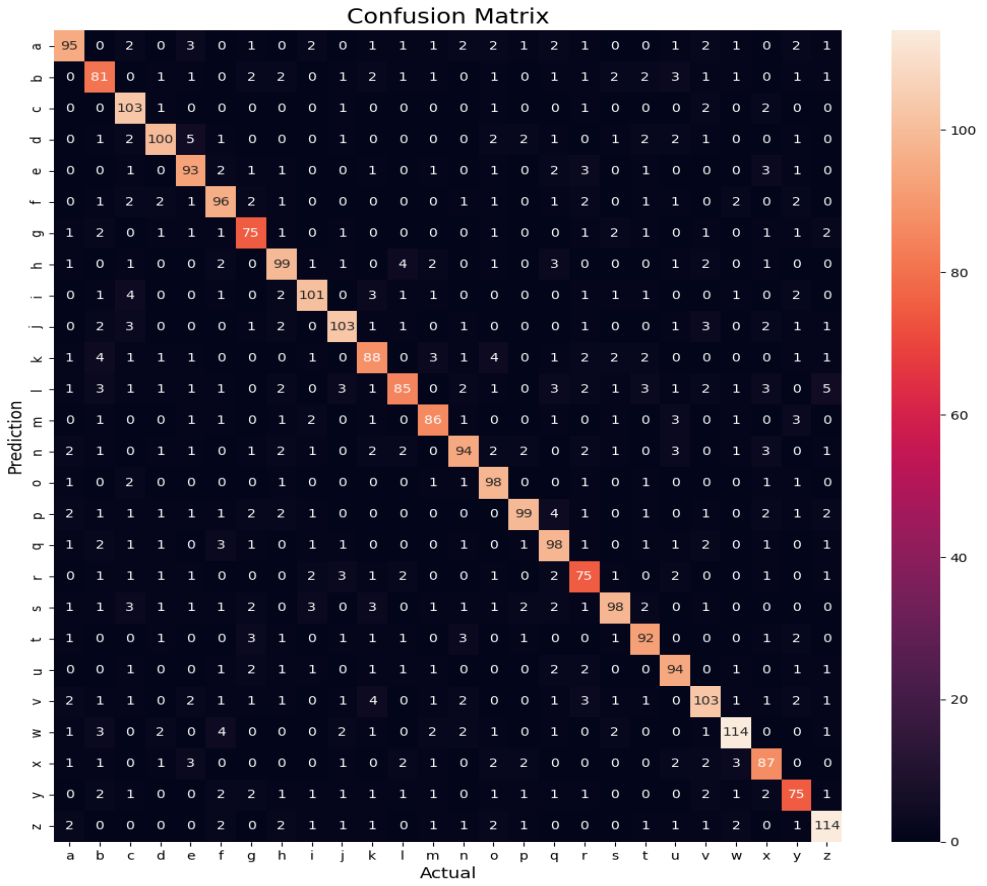

Fig. 7: Confusion Matrix

5.2 Multiclass Receiver Operating Characteristic (ROC):

The Multiclass Receiver Operating Characteristic (ROC) is a valuable evaluation method applied to deep learning models designed for lip reading tasks. Unlike traditional binary ROC analysis, which deals with two classes, the multiclass ROC assesses the performance of deep learning models in distinguishing multiple classes of lip movements or phonemes. It provides a comprehensive view of the model's ability to differentiate between various lip-related categories. By analyzing the multiclass ROC curves and area under the curve (AUC) values, researchers and practitioners can gain insights into the model's class-specific discrimination capabilities. This aids in the refinement and optimization of deep learning-based lip reading systems, ultimately enhancing their accuracy and applicability.Below, our ROC curve (Fig 8) visually illustrates model performance in distinguishing classes effectively.

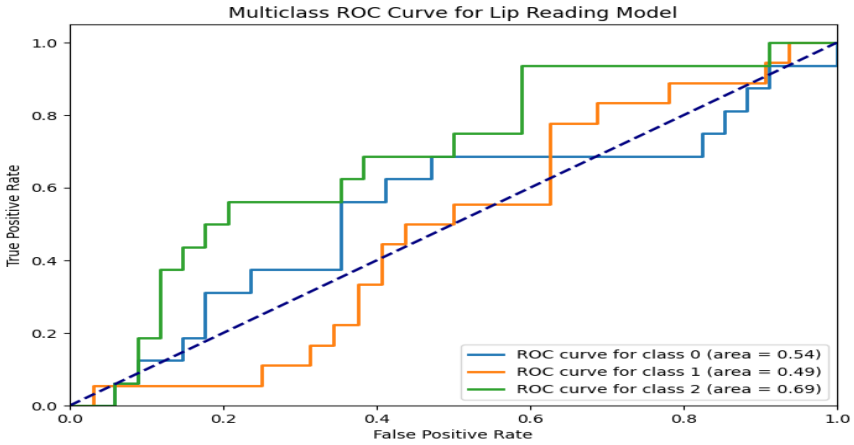

Fig. 8: Multiclass Receiver Operating Characteristic

6 Future Work:

Our future work involves integrating LipNet and ResNet into our architecture to enhance both accuracy and efficiency by concatenating these models, we aim to leverage their strengths for improved performance in our project.

7 conclusions:

Our research introduces an innovative CNN architecture designed for end-to-end learning. This model effectively maps sequences of image frames capturing a speaker's mouth movements to complete sentences. Importantly, it eliminates the need for manual video segmentation into words, and it doesn't rely on pre-defined spatiotemporal visual features or separate sequence models.

Our empirical evaluation underscores the significance of efficient spatiotemporal feature extraction. In future work, we plan to extend this by applying LipNet in concatenation with wav2lip to more extensive datasets, potentially incorporating sentence-level data. Additionally, we recognize that some applications, like silent dictation, exclusively require video inputs. Nevertheless, our aim is to broaden LipNet's applicability by incorporating it into a jointly trained audiovisual speech recognition model. This approach leverages visual cues to enhance performance, particularly in noisy environments, thus expanding LipNet's potential applications.

References

[1] Esther, D.J., Gayathri, G., Binoy, K., Neha, S., Mathew, D.: Smart assistive device for speech impaired using silent sound technology

[2] Swetha, K.: A survey on silent sound technology using electromyography and image processing. Solid State Technology **63**(5), 3983–3986 (2020)

[3] Dilip, M., .R, D.: Design and development of silent speech recognition system for monitoring of devices. International Journal for Research in Applied Science and Engineering Technology **7**, 2321–9653 (2019) https://doi.org/10.22214/ijraset.2019.6338

[4] Chung, J.S., Zisserman, A.: Lip reading in the wild. In: Lai, S.-H., Lepetit, V., Nishino, K., Sato, Y. (eds.) Computer Vision – ACCV 2016, pp. 87–103. Springer, Cham (2017)

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

[6] Li, W., Yuan, J., Zhang, L., Cui, J., Wang, X., Li, H.: semg-based technology for silent voice recognition. Computers in Biology and Medicine **152**, 106336 (2023)

[7] Fenghour, S., Chen, D., Xiao, P.: Decoder-encoder lstm for lip reading. In: Proceedings of the 8th International Conference on Software and Information Engineering, pp. 162–166 (2019)

[8] Hao, M., Mamut, M., Yadikar, N., Aysa, A., Ubul, K.: A survey of research on lipreading technology. IEEE Access **8**, 204518–204544 (2020)

[9] Yadava, T., Jayanna, H.: Automatic isolated kannada speech recognition system under degraded conditions. In: 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), pp. 146–150 (2019). IEEE

[10] Radha, N., Shahina, A., Khan, N.: Visual speech recognition using fusion of motion and geometric features. Procedia Computer Science **171**, 924–933 (2020)

[11] Petridis, S., Wang, Y., Ma, P., Li, Z., Pantic, M.: End-to-end visual speech recognition for small-scale datasets. Pattern Recognition Letters **131**, 421–427 (2020)

[12] Wang, M.-Y., Biswas, A., Yıldırım, Y., Moraru, L., Moldovanu, S., Alshehri, H.M.: Optical solitons for a concatenation model by trial equation approach. Electronics **12**(1) (2023) https://doi.org/10.3390/electronics12010019