Lip Reading AI

Prof. Manisha Vaidya[1], Shravani Bante[2], Arya Katre[3], Sejal Deoghare[4]

[1] *Prof. Manisha Vaidya, Artificial Intelligence, Priyadarshini J.L College of Engineering, Maharashtra,*

[2] *Shravani Bante, Artificial Intelligence, Priyadarshini J.L College of Engineering, Nagpur, Maharashtra, India*

[3] *Arya katre, Artificial Intelligence, Priyadarshini J.L College of Engineering, Maharashtra, India*

[4] *Sejal Deoghare , Artificial Intelligence, Priyadarshini J.L College of Engineering, Maharashtra, India*

ABSTRACT

Lip Reading AI is a functional model developed by our team to work with AI and it's advancement in the speech recognitions techniques. The work proposes a novel approach for lip reading and recognition of speech. The system uses modern deep learning techniques and CNNs. They are used to detect and count individuals on given scene while detecting their words and lip movements. The proposed model achieves high accuracy and leverages the capacities of CNN and classifications.

Keywords: *image recognition, neural networks, lip net, deep learning, stream-lit***.**

1. INTRODUCTION

Before technologies human interventions and dependencies were more and trained. They had enough man--power to train and accustom them well. But as the technology increased it's power they opened up new avenues for all fields. The advancement for proper facalities and accuracy has increased. Developing an AI system to bridge the communication gap for those to face challenges in verbal communication, for safety measures in miltary services, etc. the system relies on deep leaarning techniques and accurate lip mapping for real time data. The implications of a reliable lip reading system extend far beyond individual use cases; they encompass social inclusion and accessibility in various domains.

The core of the system leverages advanced algorithms, primarily Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, to analyze video inputs and predict spoken words based solely on lip movements.

1.1 Aim

The Main aim behind the project is to develop a more efficient system for acquiring the required accuruay for the real time data. Bridging the gap between people and maintaining an efficient balance in society is necessary but by relying on correct method.

1.2 Objective

The objective behind the model is to enhance the accessibility in reliable communication tool for people with speech impairment, enabling seamless interaction and ensuring the accuracy in various environment. We have ensured fast and efficient lip reading by training a robust model with different language and adaptability. Also we have explored various fields with the use of our model and have us allowed to be flexible.

1.3 Literature review:

[1] "LipNet: End-to-End Sentence-level Lipreading" - Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, Nando de Freitas.

Multimodal learning uses information from multiple sources, but the effectiveness of auxiliary modalities is unclear. In Multimodal Automatic Speech Recognition (MMASR) models, while showing gains over traditional architectures, they don't incorporate visual information when audio signals are corrupted, suggesting the need for better visually grounded adaptation techniques.

[2] "AI Lip Reader Detecting Speech Visual Data with Deep Learning"
Traditional automatic lip-reading systems generally consist of two stages: feature extraction and recognition, while the handcrafted features are empirical and cannot learn the relevance of lip movement sequence sufficiently. A proposed hybrid neural network architecture, which integrates CNN and bidirectional LSTM (BiLSTM) for lip reading. First, we extract key frames from each isolated video clip and use five key points to locate mouth region. Then, features are extracted from raw mouth images using an eight-layer CNN. The extracted features have the characteristics of stronger robustness and fault-tolerant capability. Finally, we use BiLSTM to capture the correlation of sequential information among frame features in two directions and the softmax function to predict final recognition result. The proposed method is capable of extracting local features through convolution operations and finding hidden correlation in temporal information from lip image sequences. The evaluation results of lip-reading recognition experiments demonstrate that our proposed method outperforms conventional approaches such as active contour model (ACM) and hidden Markov model (HMM)

[3]" Sequence Modelling with Connectionist Temporal Classification(CTC)", an algorithm used to train deep neural networks in speech recognition, handwriting recognition and other sequence problems.
Two experiments were conducted to assess observers' abilities to read speech information on a face under visual-only and visual-auditory discrepancy conditions. The first experiment involved identification and multiple choice testing, while the second experiment manipulated the eompellings of the visual-auditory discrepancy as a single speech event. Results showed that competing visual information had little effect on auditory speech recognition, but visual speech recognition was significantly interfered with when discrepant auditory information was present. Auditory bias during speech was found to be a moderately compelling conscious experience, not simply a case of fused responding or guessing. Recent research has shown that vision does not completely or inevitably dominate the processing of nonvisual information. By instructing or permitting subjects to attend to nonvisual information, visual dominance can be reduced or eliminated.

[4] "LipNet: End-to-End Sentence-level Lipreading" - GitHub Code implementation.
A proposed hybrid neural network model for Chinese lip-reading, integrating attention mechanisms into CNN and RNN. The model adds the convolutional block attention module (CBAM) to the ResNet50 neural network, improving feature extraction performance. The time attention mechanism is added to the GRU neural network, extracting features from consecutive lip motion images. Experiments show the model accurately recognizes Chinese numbers and frequently used words, outperforming other lip-reading systems

[5] "Keras Automatic Speech Recognition With CTC."
Lipreading is the task of decoding text from the movement of a speaker's mouth. Traditional approaches separated the problem into two stages: designing or learning visual features, and prediction. More recent deep lipreading approaches are end-to-end trainable (Wand et al., 2016; Chung & Zisserman, 2016a). All existing works, however, perform only word classification, not sentence-level sequence prediction. Studies have shown that human lipreading performance increases for longer words (Easton & Basala, 1982), indicating the importance of features capturing temporal context in an ambiguous communication channel. Motivated by the observation, LipNet, a model that maps a variable-length sequence of video frames to text, making use of spatiotemporal convolutions, an LSTM recurrent network, and the connectionist temporal classification loss, trained entirely end-to-end. To the best of the knowledge, LipNet is the first lipreading model to operate at sentence level, using a single end-to-end speaker-independent deep model to simultaneously learn spatiotemporal visual features and a sequence model. On the GRID corpus, LipNet achieves 93.4% accuracy, outperforming experienced human lipreaders and the previous 79.6% state-of-the-art accuracy.

[6] "LSTM Based Lip Reading Approach for Devanagiri Script"- M. S. Patil, S. Chickerur, Anand S. Meti, Priyanka M Nabapure, Sunaina Mahindrakar, Sonal Naik, Soumya Kanyal
Several training strategies and temporal models have been recently proposed for isolated word lip-reading in a series of independent works. However, the potential of combining the best strategies and investigating the impact of each of them has not been explored. It systematically investigate the performance of state-of-the-art data augmentation approaches, temporal models and other training strategies, like self-distillation and using word boundaries indicators.

[7] "Training Strategies for Improved Lip-Reading" IEEE International Conference on Acoustics, Speech, and Signal Processing Pingchuan Ma, Yujiang Wang, Stavros Petridis, Jie Shen, M. Pantic

A novel lip-reading solution, which extracts the geometrical shape of lip movement from the video and predicts the words/sentences spoken is proposed. An Indian-specific language data set is developed which consists of lip movement information captured from 50 persons. This includes students in the age group of 18 to 20 years and faculty in the age group of 25 to 40 years. All have spoken a paragraph of 58 words within 10 sentences in Hindi (Devanagari, spoken in India) language which was recorded under various conditions. The implementation consists of facial parts detection, along with Long short term memory's .

[8] "Automatic Lip Reading Using Convolution Neural Network and Bidirectional Long Short-term Memory" Int. J. Pattern Recognit. Artif. Intell. Yuanyao Lu, Jie Yan

In recent years, traditional lipreading methods have been gradually replaced by deep learning methods. The advantage of deep learning methods is that they can learn the best features from large databases. The article analyzes typical deep learning methods in detail according to their structural characteristics, and lists existing lipreading databases, including their detailed information and the methods applied to these databases. Finally, the problems and challenges of current lipreading methods are discussed, and the future research direction has prospected.

[9] "AI Lip Reader Detecting Speech Visual Data with Deep Learning 2023 4th International". Conference on Intelligent Technologies (CONIT) Geetha C, Rohan Jai D, Sandheep Krishna A, Seelam Sai Vara Prasad Reddy

Although automatic speech recognition (ASR) technology is mature, there are still some unsolved problems, such as how to accurately identify what the speaker is saying in a noisy environment. Lip-reading is a visual speech recognition technology that recognizes the speech content based on the motion characteristics of the speaker's lips without speech signals. Therefore, lip-reading can detect the speaker's content in a noisy environment, even without a voice signal.

2. PROBLEM STATEMENT AND BACKGROUND

2.1 Problem Statement

Real world problems are complex, dynamics and always require new advancements. Our model have a dynamic and robust approach to the modern problems. Lip reading AI aims to address the issue of challenges faced in interpreting the language and visual representation. The model rely on the real time videos and data to give a better solution and expressions. Environment factors can be issue but we have also determined all such drawbacks to give an advance model to the society.

2.2 Background

Lip reading AI is the process of analyzing the lip movements and expression of a person and detect the speech. This technology has implications in various domains like healthcare, education, and security. The emergence of deep learning has significantly improved the accuracy of lip reading model.

3. METHODOLOGY

The development of this lip reading AI system involves several key components:

The dataset we have used is a subset of the GRID corpus. The dataset includes videos of high quality performing various phrases.

- **Data Collection**: The system relies on high-quality video datasets of speakers performing various phrases. These datasets are essential for training the model to recognize and predict lip movements accurately.
- The dataset used for training the model is a subset of the Grid Corpus Dataset . Used gdown to download a subset (1 speaker) of the full dataset (34 speakers) from google drive.
- **Deep Learning Architecture**: The architecture consists of CNNs for spatial feature extraction from individual frames of the video and LSTMs for capturing the temporal dependencies between the lip movements over time. This combination allows the system to effectively learn both the static features of lip shapes and the dynamic changes during speech.

3.1 Model Design

- The core of the lip reading AI system involves deep learning models, typically leveraging the following architecture:

a. Convolutional Neural Networks (CNNs) Feature Extraction: CNNs are used to extract spatial features from individual video frames. They capture details of lip shapes and movements by applying convolution filters to input frames.

b. Recurrent Neural Networks (RNNs) Temporal Processing: To capture the sequential nature of speech, RNNs, specifically Long Short-Term Memory (LSTM) or Bidirectional LSTMs, are used. These networks can model the temporal dependencies between lip movements over time.

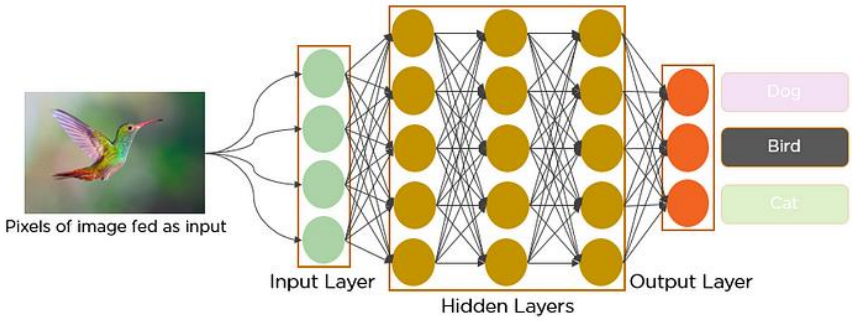

fig:3.1

c. Connectionist Temporal Classification (CTC)

- Loss Function: Since lip movements for words vary in length and timing, the CTC loss is employed to allow the model to output sequences of different lengths and align them with the transcription without requiring exact frame-to-word alignment.

3.2 Training the Model

- **Input Data**: The model is fed sequences of lip movement frames and their corresponding transcription.

fig:3.2

- **Batch Size and Epochs**: During training, the model is trained in batches, and the process is repeated over multiple epochs to learn optimal weights.

```
learner = cnn_learner(dls, resnet34, metrics=[error_rate, accuracy])

/usr/local/lib/python3.7/dist-packages/fastai/vision/learner.py:265: UserWarning: `cnn_learner` has been renamed to `vision_learner` -- please update your code
  warn("`cnn_learner` has been renamed to `vision_learner` -- please update your code")
Downloading: "https://download.pytorch.org/models/resnet34-b627a593.pth" to /root/.cache/torch/hub/checkpoints/resnet34-b627a593.pth
100% 83.3M/83.3M [00:01<00:00, 83.1MB/s]
```

fig:3.3

epoch	train_loss	valid_loss	error_rate	accuracy	time
0	2.469998	2.423700	0.666338	0.333662	17:30
1	2.321015	2.291972	0.646020	0.353980	17:31
2	2.188901	2.159503	0.617310	0.382690	17:31
3	2.017969	2.068784	0.595093	0.404907	17:32
4	1.881958	1.978835	0.571738	0.428262	17:31
5	1.798404	1.968829	0.570788	0.429212	17:29

fig:3.4

- **Optimizer**: **Adam** or **Stochastic Gradient Descent (SGD)** are commonly used optimizers for efficient learning.
- **Regularization**: Techniques like **dropout** and **batch normalization** are applied to avoid overfitting, ensuring the model generalizes well to new, unseen data.

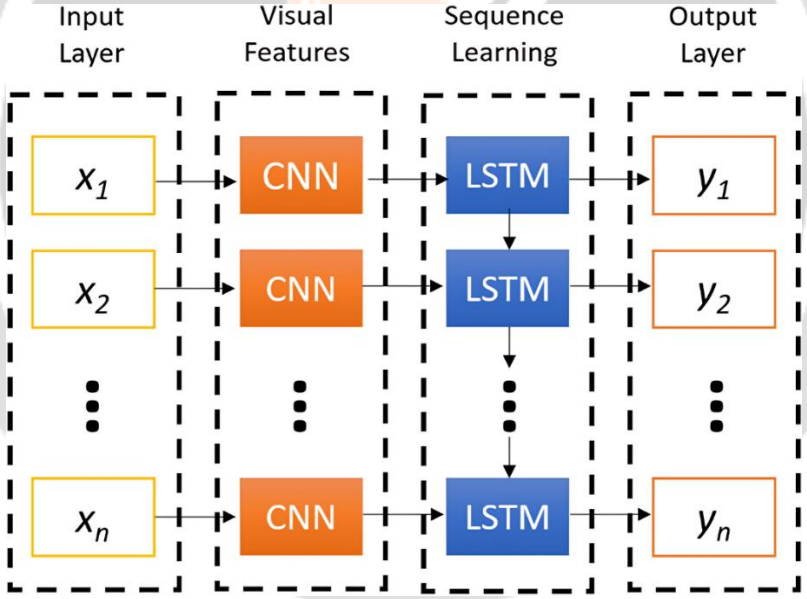

fig:3.5

4. CONCLUSION

Lip reading AI proceeds towards today's problems with a reassuring transformation through visual speech recognition, enabling a smooth interpretation through different techniques. CNN and RNN extracts lip movements inn real time data and provide high accuracy speech recognition without audio input. It's a promising model that may help in different fields and support multi-lingual speech and languages. It is a future interaction without any complications and making it a valuable experience for others in various domains.

References

[1] Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, Nando de Freitas," LipNet: End-to-End Sentence-level Lipreading", in International journal of image and graphics Technology, arXiv:1611.01599, Nov. 2022.

[2] Alex Graves et al., "Sequence Modelling with Connectionist Temporal Classification(CTC)", IMAC 2006, page 369-376.

[3] M. Nawaz, T. Nazir, M. Masood, F. Ali, M. A. Khan, "LipNet: End-to-End Sentence-level Lipreading", in IEEE vol 32, no 7, pp 2145-2148, 2022.

[4] Sharma, S. D., Sharma, S., Pathak , "Keras Automatic Speech Recognition With CTC", in 2023 2nd edition of IEEE delhi section flagship conference (pp 1 – 7). IEEE (2023 feb)

[5] M. S. Patil, S. Chickerur, Anand S. Meti, Priyanka M Nabapure, Sunaina Mahindrakar, Sonal Naik, Soumya Kanyal, "LSTM Based Lip Reading Approach for Devanagiri Script", Published in IEEE, 2018, IEEE International Conference on computational Intelligence and computating research vol 2018, page 1- 5.

[6] Pingchuan Ma, Yujiang Wang, Stavros Petridis, Jie Shen, M. Pantic, "Training Strategies for Improved Lip-Reading" IEEE International Conference on Acoustics, Speech, and Signal Processing 2021, ICASSP 2021, Pages 7608-761.

[7] Yuanyao Lu, Jie Yan, "Automatic Lip Reading Using Convolution Neural Network and Bidirectional Long Short-term Memory" Published in Int. J. Pattern Recognit. Artif. Intell, 2019, Volume 33, Issue 11, Pages 1-18.

[8] Geetha C, Rohan Jai D, Sandheep Krishna A, Seelam Sai Vara Prasad Reddy, "AI Lip Reader Detecting Speech Visual Data with Deep Learning" 2023 4th International Conference on Intelligent Technologies (CONIT), Pages 1-5.