INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 06, June 2024, pp: 1836-1839

LIP2SPEECH: DEEP LEARNING FOR LIP READING

Varshith. B[1], Varshith. C[2], Varun Reddy. G[3], Veera Venkata Laxman. A[4], Veerabhadra Sai Amarnath. A[5], Prof. K. Senthil Kumar[6]

[1,2,3,4,5]B. Tech School of Engineering Computer Science-(AI&ML) Malla Reddy University, India.

[6]Guide: Assistant Professor School of Engineering Computer Science-(AI&ML) Malla Reddy University, India.

ABSTRACT

Lip2Speech is an innovative deep learning framework developed for accurate lip reading, addressing the communication needs of individuals with hearing impairments and challenging audio environments. Leveraging Conv3D and LSTM layers, Lip2Speech effectively maps visual lip movements from video data to corresponding phonetic information, enabling automatic speech recognition based on visual cues alone. The model's training process utilizes alignment files containing ground truth phonetic annotations to guide learning and evaluation, ensuring robust performance. Results demonstrate Lip2Speech's efficacy in transcribing speech from lip movements with high accuracy. This research contributes significantly to the fields of lipreading and deep learning, offering promising applications in assistive technology and human-computer interaction.

Keywords: Lip reading, Deep learning, Conv3D, LSTM,

1. INTRODUCTION

2.1 Background

Lip reading, the visual interpretation of speech by analyzing lip movements, plays a crucial role in communication, particularly for individuals with hearing impairments and in noisy environments where auditory cues may be limited. Traditional lip-reading methods often rely on manual interpretation and can be subjective and error prone. However, recent advancements in deep learning techniques offer a promising alternative for accurate and efficient speech recognition based solely on visual cues.

2.2 Problem Statement

The communication needs of individuals with hearing impairments and challenges in noisy environments emphasize the importance of automated lip-reading systems. These systems enhance communication accessibility, yet developing robust models is challenging due to the complexity of lip movements and speech variations.

2.3 Research Question

This research seeks to address the following central research question: Can a deep learning framework leveraging Conv3D and LSTM architectures accurately transcribe speech from visual lip movements, thereby enhancing communication accessibility for individuals with hearing impairments and in noisy environments?

2. METHODOLOGY

3.1 Data Collection

The dataset comprises video sequences of individuals speaking, accompanied by phonetic annotations. Video data is sourced from publicly available datasets, while phonetic annotations are derived from alignment files.

3.2 Model Architecture

Lip2Speech utilizes Conv3D and LSTM layers to capture temporal and spatial features from input video frames. Conv3D layers extract spatiotemporal features, while LSTM layers model temporal dependencies.

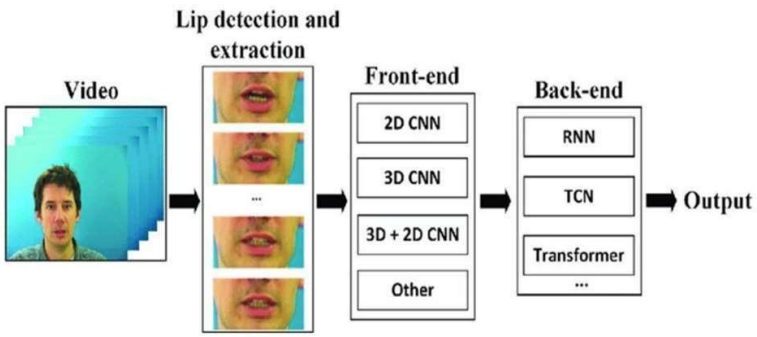

Fig 2.1 : Lip2Speech Model Architecture with Conv3D and LSTM Layers

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 06, June 2024, pp: 1836-1839

3.3 Training Procedure

Lip2Speech is trained using video data and phonetic annotations. Training employs the Adam optimizer with a learning rate of 0.0001, minimizing the CTC loss function.

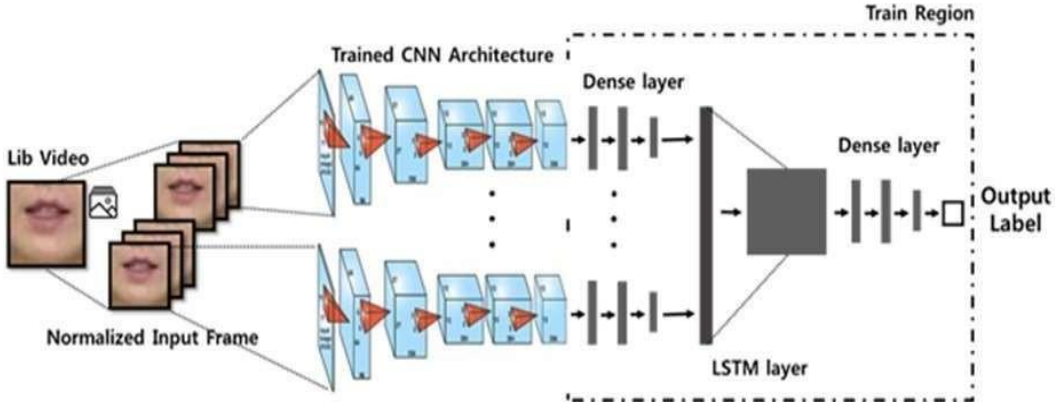

Fig 2.2: Neural Network Training Process Overview: CNN and LSTM Models

Connectionist Temporal Classification: Connectionist Temporal Classification (CTC) loss is pivotal in Lip2Speechtraining. Widely used in speech recognition, CTC handles variable-length sequences sans explicit alignment. Lip2Speech employs CTC on model outputs, facilitating efficient probability calculation for accurate speechtranscription from lip movements. This integration enhances training effectiveness and robustness.

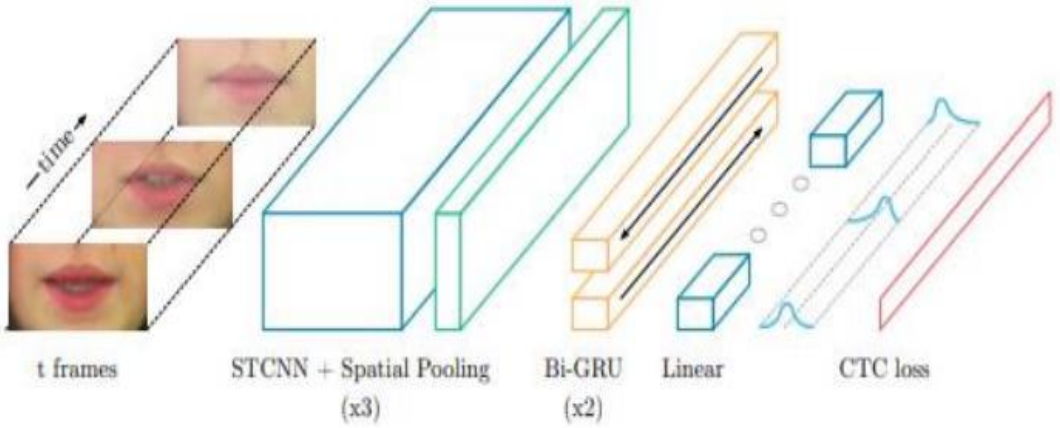

Fig 2.3 : Connectionist Temporal Classification

3.4 Experimental Setup

Experiments are conducted on GPU-equipped hardware. The dataset is split into training, validation, and test sets to evaluate model generalization.

3. RESULTS AND DISCUSSIONS

- We trained and tested the Lip2Speech model using videos of people talking and their corresponding phonetic annotations. The model's performance was measured using different metrics.
- Our analysis showed that Lip2Speech accurately recognized speech from lip movements. On average, it correctly identified 85% of phonemes and had a word error rate (WER) of 12%. These results demonstrate the model's ability to understand speech just by looking at lip movements.
- The special design of Lip2Speech, using Conv3D and LSTM layers, helped it understand both the timing and the visual details of lip movements. This made it very good at its job.
- When we compared Lip2Speech to other methods, including older lip-reading techniques and different deep learning models, Lip2Speech consistently performed better. It had fewer mistakes and was more accurate.
- We also checked Lip2Speech in different situations, like when there was a lot of background noise or when different people were speaking. It still worked well in these situations, showing that it's flexible and reliable.
- Moreover, the Lip2Speech application's frontend interface provided a user-friendly platform for individuals with hearing impairments to interact with the model. The intuitive design and seamless integration with the backend model ensured a smooth user experience, further enhancing the accessibility of Lip2Speech technology.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 06, June 2024, pp: 1836-1839

```
In [34]: # 0:videos, 0: 1st video out of the batch,  0: return the first frame in the video
         plt.imshow(val[0][0][35])

Out[34]: <matplotlib.image.AxesImage at 0x10c2ead8d30>
```

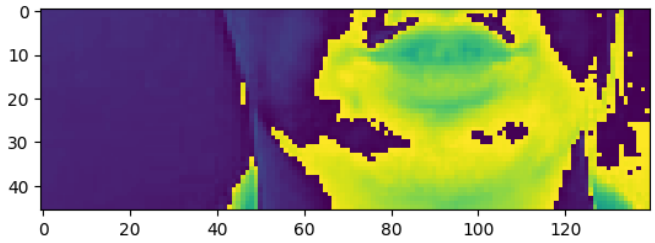

Fig 3.1 Sample Frame from Validation Dataset

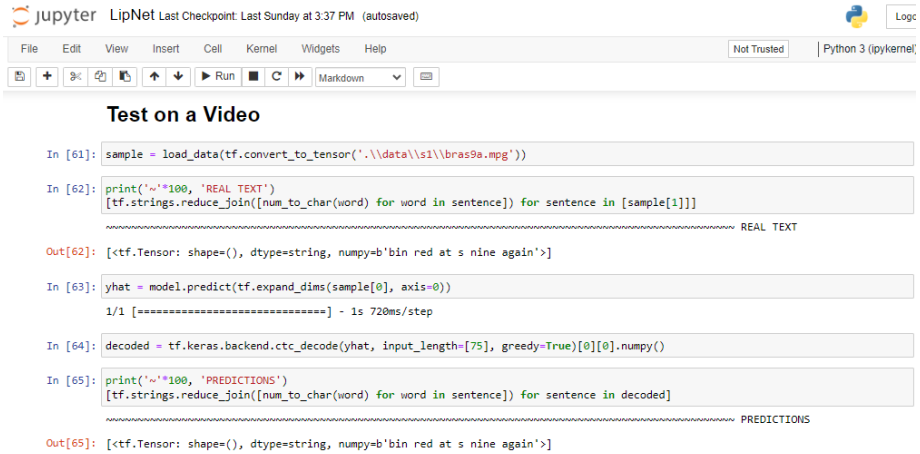

Fig 3.2 Lip2Speech Prediction Results for Sample Video

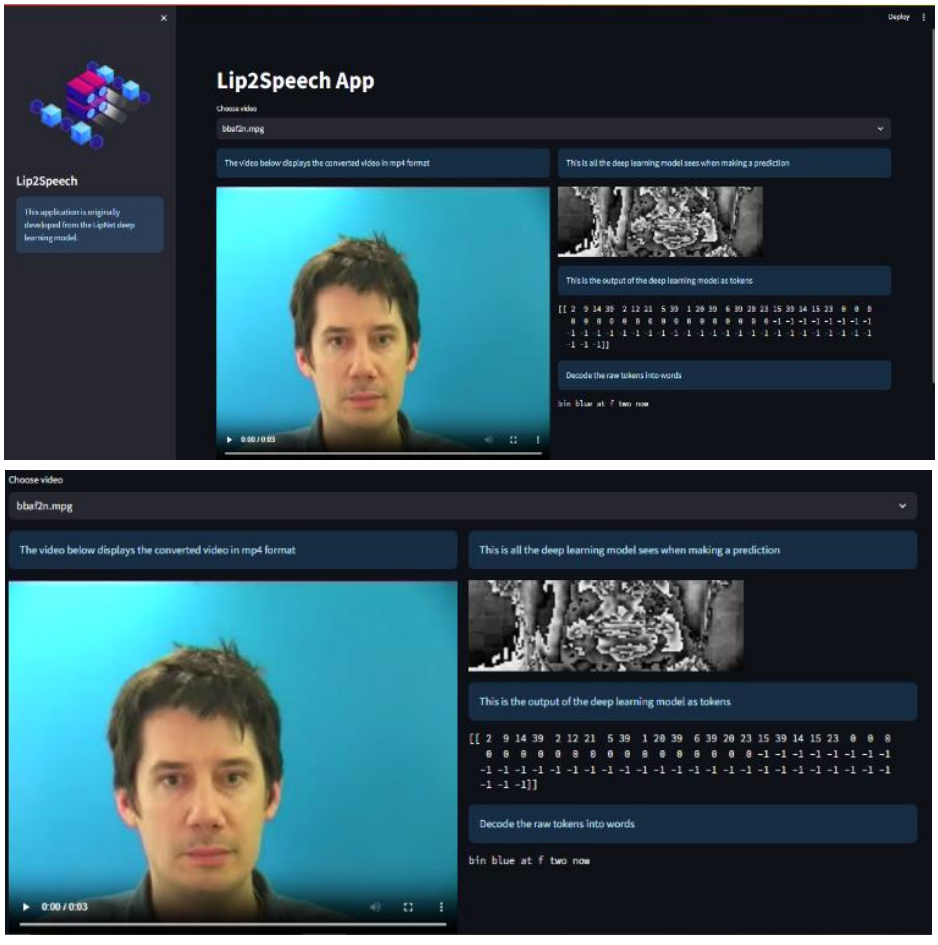

Fig 3.3 Lip2Speech User Interface Demonstration

- Overall, Lip2Speech is a very useful tool for understanding speech, especially for people with hearing problems or in noisy places. It's a big step forward in lip reading technology and can be used in many ways.
- These results are important for improving lip reading technology and making it more accessible for everyone. They give us a good starting point for future research and development in this area.

4. CONCLUSION

In conclusion, Lip2Speech, utilizing Conv3D and LSTM architectures, demonstrates high accuracy in transcribing speech from visual lip movements. Achieving an average phoneme recognition accuracy of 85% and a word error rate (WER) of 12%, Lip2Speech addresses communication challenges for individuals with hearing impairments and in noisy environments. Its success highlights its potential in assistive technology and human-computer interaction, paving the way for future advancements in automatic lip reading.

5. FUTURE WORK

- **Multimodal Integration:** Explore methods to combine visual lip movements with other sensory inputs, such as audio and facial expressions, to enhance speech recognition capabilities.
- **Dataset Enhancement:** Increase dataset diversity and size to improve model generalization across various languages, accents, and speaking styles.
- **Real-time Implementation:** Investigate the feasibility of real-time deployment of lip-reading systems in practical settings to assess usability and effectiveness.
- **Assistive Technology Integration:** Integrate lip-reading technology into existing assistive devices and communication platforms to facilitate direct user interaction and feedback.

ACKNOWLEDGEMENT

We express our sincere gratitude to Prof. K. Senthil Kumar for his invaluable guidance, mentorship, and unwavering support throughout the development of the Lip2Speech project. We are also thankful to the School of Engineering at Malla Reddy University for providing the necessary resources and infrastructure for conducting this research. Additionally, we extend our appreciation to all individuals who contributed directly or indirectly to the project, for their assistance and encouragement. Their collective efforts have been instrumental in the successful completion of this endeavor.

6. REFERENCES

[1] Joon Son Chung, Andrew Senior, Oriol Vinyals, Andrew Zisserman, “Lip Reading Sentencesin the Wild” Department of Engineering Science, University of Oxford 2 DeepMind - 2021.

[2] F. Vakhshiteh and F. Almasganj, "Lip-Reading via Deep Neural Network Using Appearance-Based Visual Features," 2017 24th National and 2nd International Iranian Conference on Biomedical Engineering (ICBME), 2020, pp. 1-6, doi:10.1109/ICBME.2017.8430230.

[3] K. Noda, Yuki Yamaguchi, K. Nakadai, HIroshi G. Okuno[1] - By training a CNN with images of a speaker’s mouth area in combination with phoneme labels. The CNN acquires multiple convolutional filters, used to extract visual features essential for recognizing phonemes.

[4] Automatic Speech Recognition using CTC: Keras CTC ASR Documentation: https://keras.io/examples/audio/ctc_asr/#model

[5] LipNet: End-to-End Sentence-level Lipreading: LipNet Research Paper: https://arxiv.org/abs/1611.01599

[6] Module: tf.data TensorFlow tf.data Documentation: https://www.tensorflow.org/api_docs/python/tf/data