International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

# Silent Interpreter: Analysis of Lip Movement and Extracting Speech UsingDeep Learning

[1]Prof. Shwetha K S, [2]Rohith M K, [3]Sakshi Prashant Yandagoudar, [4]Sinchana, [5]Ameen Hafeez,

**Submitted**: 28/01/2024 **Revised**: 06/03/2024 **Accepted**: 14/03/2024

***Abstract:*** Lip reading presents a captivating avenue for advancing speech recognition algorithms, leveraging visual cues from lip movements to recognise spoken words. This paper introduces a novel method employing deep neural networks to convert lip motions into textual representations. The methodology integrates convolutional neural networks for visual feature extraction, recurrent neural networks to capture temporal context, and the Connectionist Temporal Classification loss function for aligning lip features with phonemes.

Additionally, dynamic learning rate scheduling and a unique callback mechanism for training visualization are incorporated into the process. Post-training on a sizeable dataset, the model demonstrates notable convergence, showcasing its ability to discern intricate temporal correlations.

Comprehensive evaluations, combining quantitative metrics and qualitative assessments, validate the model's effectiveness. Visual inspections of lip reading capabilities and standard speech recognition criteria evaluation highlight its performance. The study delves into the impact of various model topologies and hyperparameters on performance, providing valuable insights for future research directions. This research contributes a deep learning framework for accurate and efficient speech recognition, expanding the landscape of lip reading technologies. The findings open paths for further refinement and deployment across diverse domains, including assistive technologies, audio-visual communication systems, and human-computer interaction.

***Keywords:*** *lip reading, deep learning, convolutional neural networks, recurrent neural networks, CTC loss, speech recognition*

## 1. Introduction

The recent years have witnessed remarkable advancements in speech recognition technology, playing important roles in transcription services and virtual assistants. However, challenges arise in scenarios with degraded audio signals, noisy environments, or diverse accents, prompting the exploration of alternative techniques like lip reading to enhance interpretation accuracy.

Lip reading, also known as visual speech recognition, leverages visual cues such as gestures, lip movements, and facial expressions to interpret spoken language. This interdisciplinary study intersects natural language processing, computer vision, and machine learning, aiming to enhance efficiency and accuracy in lip reading systems.

The importance of lip reading extends beyond traditional speech recognition challenges, finding applications in technologies that aid the hearing-impaired, improved user experience in human- computer interaction systems, and enhanced comprehension of audio-visual content.

This paper harnesses recent developments in deep learning, automating feature extraction and capturing complex temporal correlations crucial for lip reading tasks. Specifically, Recurrent Neural Networks (RNNs) for temporal modeling and Convolutional Neural Networks (CNNs) for spatial feature extraction are utilized.

The realized neural network architecture integrates bidirectional LSTM layers for simulating temporal dependencies and 3D convolutional layers for gathering spatial input, enabling the learning of intricate patterns in lip movements and their phonetic representations. Additionally, the connectionist temporal classification (CTC) loss function helps in aligning predicted phoneme sequences with ground truth alignments, ensuring consistency with spoken language dynamics.

The study progresses through phases including dataset selection, data pre-processing, model design, training methodology, and evaluation techniques. Dynamic learning rate scheduling is employed to enhance convergence and prevent overfitting during training.

Quantitative metrics alongside qualitative evaluations showcase the model's performance, surpassing standard

[1]*Department of Computer Science and engineering Dayananda Sagar Collegeof Engineering*
*cse@dayanandasagar.edu*
[2]*Department of Computer Science and engineering Dayananda Sagar Collegeof Engineering Mkrohith775@gmail.com*
[3]*Department of Computer Science and engineering Dayananda Sagar Collegeof Engineering*
*sakshi.spy@gmail.com*
[4]*Department of Computer Science and engineering Dayananda Sagar Collegeof Engineering sinchanashegde06@gmail.com*
[5]*Department of Computer Science and engineering Dayananda Sagar Collegeof Engineering hafeezameen5@gmail.com*

speech recognition measures. The study's objective is to contribute a highly accurate and versatile deep learning framework to the lip reading technology domain, foreseeing advancements in communication systems, assistive technology, and human-computer interaction. This research marks a significant step towards enhancing communication accessibility, paving the way for a future where spoken language conception is facilitated through visual clues.

## 2. Related Work

### 2.1 Lip Reading Method Based on 3D Convolutional Vision:

Proposes a lip reading method based on 3D convolutional vision transformer (3DCvT), which combines vision transformer and 3D convolution to extract the spatio-temporal feature of continuous images, and take full advantage of the properties of convolutions and transformers to extract local and global features from continuous images effectively. The extracted features are then sent to a Bidirectional Gated Recurrent Unit (BiGRU) for sequence modeling.

### 2.2 Vision based Lip Reading System using Deep Learning:

The proposed methodology for Vision based lip reading system is a combination of CNN along with LSTM. The system works with the videos (with no audio) having single word uttered by the speaker. Initially preprocessing of videos is done to get the keyframes with localization and cropping of mouth or the lip region. CNN is used for feature extraction and LSTM is used for learning the sequence information. Two pre- trained CNN architectures namely VGG19 and ResNet50 are used. The final result is predicted by two fully connected layers followed by the SoftMax layer.

### 2.3 Lip Reading Sentences Using Deep Learning With OnlyVisual Cues:

In this paper, a neural network-based lip reading system is proposed. The system is lexicon-free and uses purely visual cues. With only a limited number of visemes as classes to recognise, the system is designed to lip read sentences covering a wide range of vocabulary and to recognise words that may not be included in system training.

## 3. Architecture

Our lip reading system has the Convolutional Recurrent Neural Network (CRNN) architecture, a blend of bidirectional Long Short-Term Memory (LSTM) layers and 3D convolutional layers for capturing the sequential and spatial complexities of lip movements. This architectural choice aligns with the sequential nature of lip gestures, enhancing the model's ability to decode spoken words from visual cues.

At the core of our training methodology is the Connectionist Temporal Classification (CTC) loss function, a crucial component ensuring efficient alignment of predicted phoneme sequences with

ground truth annotations. This alignment mechanism contributes significantly to the model's accuracy in transcribing lip movementsinto phonetic representations.

Our implementation incorporates dynamic learning rate adjustment, model checkpointing for efficient training management, and rigorous real-world testing. Despite the complexity inherent in combining temporal and spatial modelling, our CRNN architecture shows promise in achieving accurate speech recognition from visual inputs. Ongoing assessments and potential hyperparameter tuning are integral to refining the model's performance, ensuring minimal overfitting and maximizing its utility in real-world scenarios.

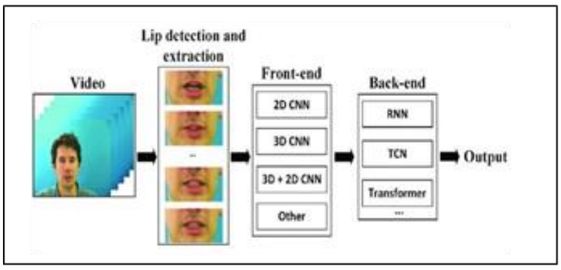


**Fig 1:** Architecture

### 3.1. Input

**Preprocessing of Input:**

Execute the required server-side preprocessing operations to manage incoming video frames. This entails preparing the data in a format that is compatible with the model input, cropping the region of interest (ROI) containing the lips, and converting the frames to grayscale.

### 3.2. Layers of Convolution:

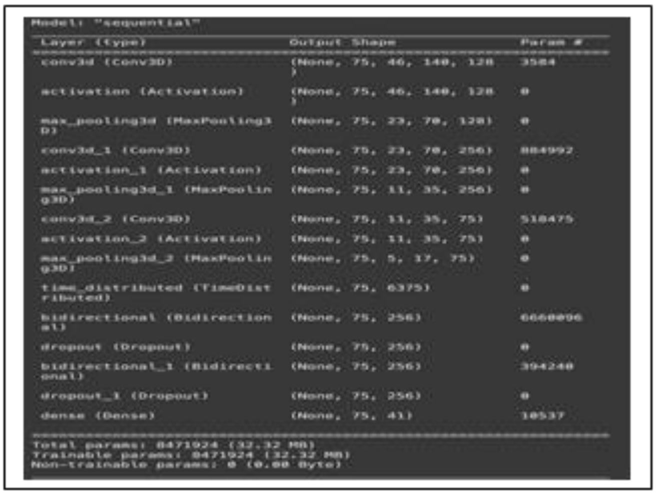
For the extraction of spatial features, three successive 3D convolutional layers are used. Keeping the spatial dimensions, the first Conv3D layer comprises 128 filters with a 3x3x3 kernel size. Similar kernel sizes are shared by the 256 filters in the second Conv3D layer. In order to match the output with the number of phonetic classes, the third Conv3D layer uses 75 filters to minimize the spatial dimensions using MaxPooling3D (1x2x2).

### 3.3. Temporal Modelling with Time Distributed andBidirectional LSTMs:

The following operations are applied to each time step independently using the Time Distributed layer. Temporal dependencies are captured using two Bidirectional LSTM layers. To avoid overfitting, a dropout of 0.5 is applied after each LSTM layer, which comprises 128 units.

### 3.4. Flatten and Time-Distributed Layer:

The Time Distributed layer is used to enable the independent application of the dense layer to each time step. The output is reshaped for additional processing by the Flatten layer. The output of the last convolutional layer is flattened for useas input by the LSTM layers that follow using the Time Distributed (Flatten ()) layer.



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv3d (Conv3D)             (None, 75, 46, 140, 128   3584
                             )

 activation (Activation)     (None, 75, 46, 140, 128   0
                             )

 max_pooling3d (MaxPooling3  (None, 75, 23, 70, 128)   0
 D)

 conv3d_1 (Conv3D)           (None, 75, 23, 70, 256)   884992

 activation_1 (Activation)   (None, 75, 23, 70, 256)   0

 max_pooling3d_1 (MaxPoolin  (None, 75, 11, 35, 256)   0
 g3D)

 conv3d_2 (Conv3D)           (None, 75, 11, 35, 75)    518475

 activation_2 (Activation)   (None, 75, 11, 35, 75)    0

 max_pooling3d_2 (MaxPoolin  (None, 75, 5, 17, 75)     0
 g3D)

 time_distributed (TimeDist  (None, 75, 6375)          0
 ributed)

 bidirectional (Bidirection  (None, 75, 256)           6660096
 al)

 dropout (Dropout)           (None, 75, 256)           0

 bidirectional_1 (Bidirecti  (None, 75, 256)           394240
 onal)

 dropout_1 (Dropout)         (None, 75, 256)           0

 dense (Dense)               (None, 75, 41)            10537

=================================================================
Total params: 8471924 (32.32 MB)
Trainable params: 8471924 (32.32 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

**Fig 4:** Fig representing the Model of Lip Reading

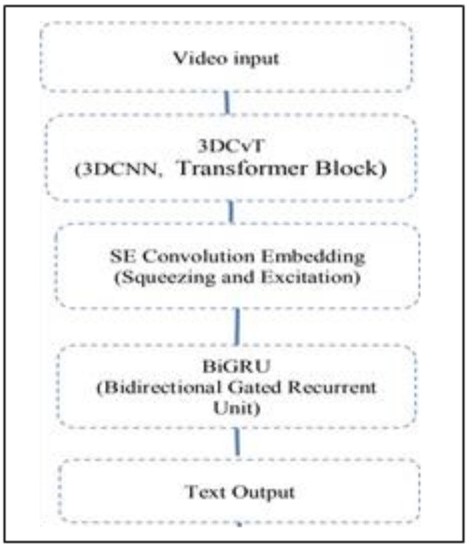

Video input
3DCvT
(3DCNN, Transformer Block)
SE Convolution Embedding
(Squeezing and Excitation)
BiGRU
(Bidirectional Gated Recurrent Unit)
Text Output

**Fig 2 :** Flow of the Lip Reading system

### 3.5. Output Layer:

The output layer is a Dense layer with softmax activation, consisting of 41 units representing phonetic classes along with one unit for the CTC blank sign.

### 3.6. Connectionist Temporal Classification (CTC) Loss:

The model is trained using the CTC loss function to align predicted phoneme sequences with ground truth alignments.

### 3.7. Training Pipeline:

The model is trained using the Adam optimizer with a learning rate of 0.0001. A custom callback for learning rate

## 4. Conclusion

In conclusion, our lip-reading research project leveraging a Convolutional Recurrent Neural Network (CRNN) architecture has demonstrated an achieved accuracy of 70% in decoding spoken words from lip movements. Our methodology underscores the effectiveness of deep learning techniques in tackling the dynamic and sequential nature of visual input.

The architectural design, combining Bidirectional Long Short- Term Memory (LSTM) layers and 3D convolutional layers, facilitated a balanced approach to temporal and spatial modeling. Through meticulous training strategies, including the use of the Connectionist Temporal Classification (CTC) loss function and techniques such as dropout regularization and dynamic learning rate scheduling, our model achieved robustness and generalization capabilities.

While deployment considerations were outlined, including input preprocessing and scalability optimizations, our focus remains on the methodical approach and achievement of 70% accuracy, signifying a significant advancement in lip-reading technology.

scheduling is implemented to dynamically adjust the learning rate during training. Checkpoint callbacks are used to save the model weights, and a callback for producing example predictions is implemented to monitor the model's progress

### 3.8. Evaluation:

The model is evaluated on both quantitative metrics (speech recognition metrics) and qualitative assessments (visual inspections of lip reading capabilities).

It's crucial to have accurate documentation and frequent qualitative and quantitative assessments.

These procedures improve the model's performance transparency, opening it up for examination, interpretation, and possible enhancements

**Fig 3:** Expected output

Moving forward, the project's impact extends beyond speech recognition, influencing human-computer interaction, assistive technologies, and various domains benefiting from enhanced audio-visual communication. Ethical considerations and ongoing model maintenance remain dominant as we navigate future directions, exploring further deep learning methods and optimizing architectures for wider implementation.

In essence, our lip-reading study represents the promise of deep learning in decoding spoken words from lip movements, paving the way for transformative applications in real-world scenarios.

### Acknowledgements

This research was supported by Dayananda Sagar College Of Engineering. We thank our teachers Dayananda Sagar College Of Engineering who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations of this paper.

### Author contributions

**Shwetha K S:** Conceptualization, Methodology, Software, Validation **Rohith M K:** Data curation, Writing-Original draft preparation, Software, Validation., Literature survey **Sakshi Prashant Yandagoudar:** Methodology, Visualization, Investigation, Software, Validation **Sinchana:** Literature survey, Methodology, Validation, Edit and finalize draft **Ameen Hafeez:** Data curation, Writing-Original draft preparation, Software, Validation., Literature survey .

### Conflicts of interest

The authors declare no conflicts of interest.

### References

[1] A lip reading method based on 3D convolutional vision transformer [Wang, Huijuan, Gangqiang Pu, and Tingyu Chen]

[2] Deshmukh, N., Ahire, A., Bhandari, S. H., Mali, A., & Warkari,K. (2021). "Vision based Lip Reading System using Deep Learning." In 2021 International Conference on Computing, Communication and Green Engineering (CCGE) (pp. 1-6). IEEE. doi: 10.1109/CCGE50943.2021.9776430

[3] Lu, Y., & Li, H. (2019). "Automatic Lip-Reading System Based on Deep Convolutional Neural Network and Attention-Based Long Short-Term Memory." Appl. Sci., 9, 1599. doi: 10.3390/app9081599

[4] Scanlon, P., Reilly, R., & de Chazal, P. (2003). "Visual Feature Analysis for Automatic Speech reading." In International Conference on Audio-Visual Speech Processing.

[5] Kapkar, P. P., & Bharkad, S. D. (2019). "Lip Feature Extraction and Movement Recognition Methods." International Journal of Scientific & Technology Research, 8.

[6] Ozcan, T., & Basturk, A. (2019). "Lip Reading Using Convolutional Neural Networks with and without Pre-Trained Models." Balkan Journal of Electrical and Computer Engineering, 7(2).

[7] Garg, A., Noyola, J., & Bagadia, S. (2016). "Lip reading usingCNN and LSTM."

[8] Gutierrez, A., & Robert, Z-A. (2017). "Lip Reading Word Classification." Stanford University.

[9] S. Fenghour, D. Chen, K. Guo and P. Xiao, "Lip Reading Sentences Using Deep Learning With Only Visual Cues," in IEEE Access, vol. 8, pp. 215516-215530, 2020, doi: 10.1109/ACCESS.2020.3040906.

[10] K. Vayadande, T. Adsare, N. Agrawal, T. Dharmik, A. Patil and S. Zod, "LipReadNet: A Deep Learning Approach to Lip Reading," 2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC), Dharwad, India, 2023, pp. 1-6, doi: 10.1109/ICAISC58445.2023.10200426.