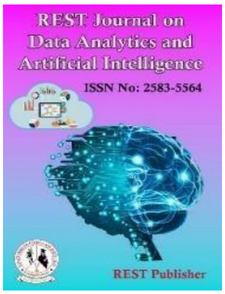
REST Journal on Data Analytics and Artificial Intelligence
Vol: 3(3), September 2024
REST Publisher; ISSN: 2583-5564
Website: http: //restpublisher.com/journals/jdaai/
DOI: https://doi.org/10.46632/jdaai/3/3/4



Visual Speech Recognition Using 3D CNN and LSTMS

[*]Sowjanya V, Harshavardhan Chary Vadla, Sai Ashrit Pinni, Shesha Sai Nishit Tirumala
Matrusri Engineering College, Saidabad, Hyderabad, Telangana, India.
**Corresponding author: sowjanya@matrusri.edu.in*

Abstract: *Lip reading, a challenging task within the domain of multimodal human-computer interaction, has garnered significant interest due to its potential applications in various fields, including assistive technologies, security systems, and human-computer interfaces. This research paper presents a deep learning-based approach to lip reading, leveraging Convolutional Neural Networks (CNNs) for feature extraction from lip images and employing a Bidirectional Long Short-Term Memory (Bi-LSTM) network for sequence modelling and transcription. The model integrates a 3D CNN architecture for spatiotemporal feature extraction, allowing it to capture spatial and temporal dependencies in lip movements. Training is facilitated using the Connectionist Temporal Classification (CTC) loss function, enabling end-to-end learning. Experimental results on the Grid Corpus dataset demonstrate the effectiveness of the proposed approach, achieving an impressive transcription accuracy rate of 85.65%. This high accuracy showcases the model's ability to accurately interpret subtle visual cues of lip movements and transcribe speech with high precision. Furthermore, evaluation metrics such as Word Error Rate (WER), Character Error Rate (CER), and overall accuracy provide insights into the model's performance. This research contributes to advancing automatic lip-reading systems, offering a deeper understanding of the challenges and opportunities in this domain. The proposed approach, combining 3D CNNs and recurrent networks, lays the foundation for future developments in improving the accuracy and robustness of lip-reading systems, ultimately enhancing their usability and applicability in diverse real-world scenarios.*

Keywords: *Lip Reading, Visual Speech Recognition (VSR), Deep Learning, Convolutional Neural Networks (CNNs), 3D CNNs, Bidirectional Long Short-Term Memory (Bi-LSTM), Connectionist Temporal Classification (CTC), Spatiotemporal Feature Extraction, Human-Computer Interaction, Speech Transcription, End-to-End Learning*

1. INTRODUCTION

Lipreading, the process of deciphering spoken language from visual cues of lip movements, holds immense potential in various applications, including human-computer interaction, assistive technologies, and surveillance systems [1]. In recent years, lipreading has witnessed significant advancements, driven primarily by integrating deep learning techniques into traditional approaches [2].

Lipreading is indispensable in communication, particularly for individuals with hearing impairments, offering them a means to comprehend spoken language through visual cues alone, enhancing their communication capabilities [3]. Moreover, lipreading proves valuable in environments where auditory cues are compromised, such as noisy settings or when speakers are distant, aiding comprehension and facilitating interaction [4]. Its significance extends to law enforcement, where surveillance footage analysis relies on lipreading for interpreting conversations and gathering intelligence [5]. In education, lipreading complements auditory instruction, providing visual reinforcement of speech sounds and aiding language learning and teaching [6]. Furthermore, advancements in lipreading technology hold promise for applications in human-computer interaction, automatic speech recognition, and accessibility features in digital communication platforms, contributing to inclusivity and improved communication access [7].

Lipreading poses numerous challenges, exacerbated by the inherent variability and ambiguity in visual speech cues [8]. One significant challenge arises from the lack of distinct visual features corresponding to individual phonemes, as many phonetic sounds share similar lip movements, making accurate differentiation difficult [9].

Moreover, environmental factors such as lighting conditions, speaker variability, occlusions, and facial expressions introduce noise and variability into the visual signals, further complicating lipreading. Additionally, the rapid pace of speech and the co-articulation of phonemes within words add complexity, necessitating sophisticated temporal modelling techniques for accurate interpretation. Despite these obstacles, recent advancements in computer vision and deep learning have shown promise in mitigating these challenges by enabling the extraction of meaningful spatial and temporal features from visual speech data, thereby improving lipreading performance [7].

Our study revolves around proposing and evaluating a deep-learning architecture tailored for lipreading tasks on the GRID corpus [10]. The GRID corpus, a well-established benchmark dataset for lipreading tasks, comprises videos of subjects uttering phrases from a predefined grammar set. The proposed model is designed to capture both spatial and temporal features from input video sequences, enabling it to transcribe spoken words from visual lip movements effectively.

Through rigorous experimentation and evaluation, we assess the performance of our proposed model against established benchmarks on the GRID corpus dataset. Our methodology involves training the model on a subset of the dataset and evaluating its performance on unseen data to ensure robustness and generalization.

Furthermore, we delve into the interpretability of the proposed model, shedding light on the underlying principles of lipreading and the role of deep learning in understanding visual cues of speech. By elucidating the intricacies of lipreading using deep learning techniques, we aim to contribute to the advancement of multimodal processing and facilitate the development of innovative applications in human-computer interaction and communication technologies.

2. RELATED WORK

Visual Speech Recognition (VSR) has witnessed substantial progress, fueled by advancements in deep learning techniques. This section categorizes and discusses relevant studies based on their methodologies, encompassing temporal modelling, transfer learning, architectural innovations, end-to-end systems, integrated approaches, and specialized applications.

Temporal modelling, crucial for capturing the sequential dynamics inherent in speech, has been a focal point of research. Wand et al. [11] pioneered using Long Short-Term Memory Recurrent Neural Networks (LSTM-RNNs) in VSR tasks, achieving a notable word accuracy of 79.6% on the OuluVS database (The OuluVS dataset contains 52 speakers, each uttering 10 different phrases five times, captured in various lighting conditions. It is used for developing and evaluating visual speech recognition systems). Their work primarily delved into word and phrase classification dynamics, albeit without addressing sentence-level sequence prediction, illuminating the challenges in modelling temporal dependencies effectively.

Transfer learning has emerged as a potent strategy to enhance VSR performance by leveraging pre-trained models and domain adaptation techniques. Garg et al. [12] delved into this domain, employing Convolutional Neural Networks (CNNs) and LSTM architectures on the MIRACL-VC1 dataset. Despite facing challenges, their methodology demonstrated potential, yielding accuracies of 56.0% and 44.5% for word and phrase classification, respectively. This underscores the utility of leveraging existing knowledge for improved performance in VSR tasks.

Architectural innovations constitute another cornerstone in advancing VSR capabilities. Chung and Zisserman [13] introduced spatial and spatiotemporal CNN architectures tailored specifically for word classification tasks. Their models achieved word accuracy of 76.20% on the BBC TV dataset, addressing challenges in handling variable sequence lengths. However, the absence of robust sentence-level prediction remains a notable challenge.

End-to-end systems signify a departure from conventional methodologies by concurrently optimizing feature extraction and classification stages. A seminal work by [14] proposed such an approach, attaining state-of-the-art performance on the OuluVS2 and CUAVE databases, boasting accuracies of 84.5% and 85.0%, respectively. This paradigm shift mitigates the limitations inherent in two-stage methods, offering promising avenues for further exploration.

Integrated approaches, amalgamating multiple modalities or harnessing extensive datasets, present another avenue for enhancing VSR performance. [15] introduced an integrated lipreading system, achieving a word error rate of

40.9% on a held-out set, leveraging the LSVSR dataset. Despite showcasing robustness, challenges persist in adapting to diverse conditions and ensuring dataset quality, emphasizing the need for continued refinement.

Dynamic features and CNNs have been explored in novel lip-reading methods, as demonstrated by [16]. Despite achieving promising accuracies of 71.76% on the ATR Japanese speech database, the applicability of these methods in real-world scenarios remains a challenge, necessitating further refinement.

Dimensionality reduction strategies, such as the PCA-based approach proposed by [17], have shown promise in enhancing visual-only lip-reading systems' efficiency. While achieving accuracies around 77% on various datasets, further optimization and exploration of dataset characteristics are essential for scalability and accuracy improvement.

Some studies target specialized applications or address domain-specific challenges in VSR. For instance, [18] focused on audio-visual speech recognition (AVSR) system robustness to noise, achieving approximately 65%-word accuracy on a Japanese dataset. Similarly, [19] designed an AVSR system for individuals with articulation disorders, achieving 50.9% accuracy in lip reading but facing scalability and generalization limitations.

Specialized applications of VSR have also been explored, such as Ulkümen and Ozturk's work [20], which developed an automatic image-based lip-reading method to detect emergency words in noisy and large-scale events. They constructed an original dataset capturing silent video frames of emergency words spoken by different speakers. Through experimentation with the SSD deep learning method, they achieved up to 76% accuracy, underscoring the potential of image-based lip-reading techniques for enhancing security and public safety measures in diverse environments. Our goal is to combine the strengths of various approaches to create a strong and effective model for lipreading entire sentences. We aim to harness the capabilities of 3DCNNs, Bi-GRUs, and CTC loss functions to develop an advanced system for visual speech recognition. By integrating these components, we aspire to improve the accuracy and reliability of lipreading technology, making it suitable for practical use in real-world scenarios.

3. PROPOSED ARCHITECTURE

A. ***Data Collection****:* The dataset utilized in this study is the GRID corpus [10], a widely used resource in lip reading research. It consists of recordings from 34 speakers, each narrating 1000 sentences. While the corpus initially encompasses a vast amount of data, specific videos were found to be missing or corrupted, resulting in a total of 32746 usable videos for analysis. The sentences in the GRID corpus are structured according to a specific grammar pattern, which includes a combination of command, color, preposition, letter, digit, and adverb categories. For example, the command category offers choices such as 'bin,' 'place,' 'lay,' and 'set,' while the colour category includes options like ‘green,' 'blue,' 'red,' and 'white.' Each category provides a predetermined number of word choices, resulting in 64000 possible sentences. Moreover, the GRID corpus incorporates a unique split approach for evaluation, withholding data from select speakers for testing while utilizing the remainder for training. A sentence-level split variant is employed, where a subset of sentences from each speaker is reserved for evaluation purposes. The GRID corpus provides a rich and varied dataset containing numerous linguistic structures ideal for training and evaluating lip reading models.

FIGURE 1. Random Frame from Grid Dataset

B. Model Architecture

Our lipreading architecture, emphasizing its nature as an end-to-end model for sentence-level lipreading. Figure 2 illustrates the architecture's implementation diagram, while the Table 1 provides detailed specifications regarding its implementation.

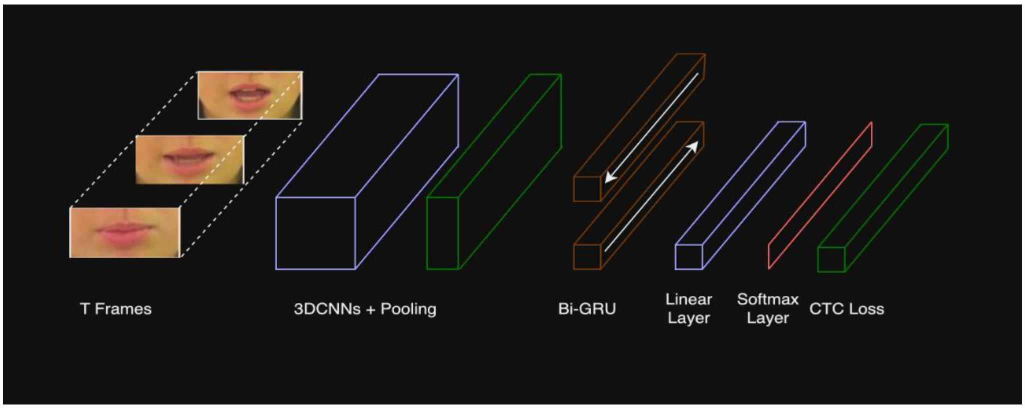


FIGURE 2. Model Architecture. A series of frames goes through 3 layers of 3D CNNs and spatial max-pooling. Features are then processed by 2 Bi-GRUs. Each Bi-GRU output step is fed into a linear layer and softmax. The model is trained end-to-end with CTC.

TABLE 1. Model architecture hyperparameters.

Layer	**Size/ stride/pad**	**Input size (dimension order)**
Conv3D	3x5x5 / 1,2,2 / 1,2,2	75x46x140x1 (TxCxHxW)
MaxPool3D	1x2x2 / 1,2,2	75x23x70x128 (TxCxHxW)
Conv3D	3x5x5 / 1,2,2 / 1,2,2	75x23x35x64 (TxCxHxW)
MaxPool3D	1x2x2 / 1,2,2	75x12x17x128 (TxCxHxW)
Conv3D	3x3x3 / 1,2,2 / 1,1,1	75x12x8x64 (TxCxHxW)
MaxPool3D	1x2x2 / 1,2,2	75x6x4x96 (TxCxHxW)
Bi-GRU	256	75x(6x4x96) (Tx(CxHxW))
Bi-GRU	256	75x512 (TxF)
Linear	vocabulary_size()+1	75x512 (TxF)
Softmax	-	75x28 (TxV)
CTC loss	1	

Here,

Size: Refers to the dimensions of the kernel/filter used in convolution operations.

Stride: The number of pixels the filter moves across the input image.

Pad: Padding refers to the process of adding zeros around the border of the input image to ensure the output size remains the same after convolution.

T (Time steps): The length of the sequence of frames or inputs over time used for processing.

C (Number of channels): The number of distinct data streams or feature maps in each frame, often representing color channels in an image.

H (Height of the input): The vertical dimension of each input frame, typically measured in pixels.

W (Width of the input): The horizontal dimension of each input frame, typically measured in pixels.

F (Number of features): The dimensionality of the feature vector extracted from each input frame.

V (Number of vertices or points): The number of key points or vertices used to represent the shape or structure in each input frame, often relevant in graph-based or skeletal representations.

C. 3D Convolutional Neural Network:

A 3D Convolutional Network (3D CNN) is a type of neural network architecture commonly used for processing spatio-temporal data, such as videos. Unlike traditional 2D CNNs [21], which operate on two-dimensional input data (e.g., images), 3D CNNs extend this concept to three-dimensional volumes, allowing them to capture spatial and temporal features simultaneously.

In a 3D CNN, convolutional filters are applied across the width and height of an input volume and along the temporal dimension, capturing information over time [22]. This enables the network to learn representations that encode spatial patterns (e.g., object shapes and textures) and temporal dynamics (e.g., motion and changes over time) within the input data.

Our architecture uses 3D Convolutional Neural Networks (3DCNNs) to interpret spoken language from video data. Unlike traditional 2D CNNs that process static images, 3DCNNs extend the convolution operation to three dimensions, allowing them to capture spatial and temporal features simultaneously. This capability is crucial for lip reading, as it enables the network to analyze the appearance of lip movements and their evolution over time.

The operation of a 3DCNN can be expressed using the following formula:

$$[3DCNN(x,w)]_{c_0tij} = \sum_{c=1}^{C}\sum_{t_0=1}^{kt}\sum_{i_0=1}^{kw}\sum_{j_0=1}^{kh} w_{c_0ct_0i_0j_0}x_{c,t+t_0,i+i_0,j+j_0}$$

This formula represents the output of the 3DCNN operation applied to the input x with the filter w at a specific output location (c0, t, i, j).

- $[3DCNN(x,w)]_{c_0tij}$ denotes the output of the 3DCNN operation at the specified location in the output volume.
- $\sum_{c=1}^{C}$ iterates over all input channels (C) of the input volume x.
- $\sum_{t_0=1}^{kt}, \sum_{i_0=1}^{kw}, and \sum_{j_0=1}^{kh}$ iterate over the temporal extent and spatial dimensions of the filter w, respectively.
- $w_{c_0ct_0i_0j_0}$ represents the weight of the filter at position (c0,c,t0,i0,j0).
- $x_{c,t+t_0,i+i_0,j+j_0}$ denotes the input value at position (c,t+t0,i+i0,j+j0).

In lip reading context, the 3DCNN processes video frames as input x and learns to extract relevant spatiotemporal features using the filter w. By convolving across both spatial and temporal dimensions, the network can capture intricate patterns of lip movements and phonetic articulation over time.

This approach offers several advantages for lip reading applications. First, 3DCNNs can effectively interpret spoken language from video data by modelling spatial details and temporal dynamics. Second, the end-to-end learning framework enables the network to directly learn features from raw video input directly, eliminating the need for handcrafted features or intermediate processing steps.

D. ***Bidirectional Gated Recurrent Unit:*** The bidirectional gated recurrent unit (Bi-GRU) is a variant of recurrent neural networks (RNNs) designed to capture past and future dependencies in sequential data. In lip reading, the Bi-GRU plays a crucial role in modelling the temporal dynamics of speech articulation over time.

The operation of a Bi-GRU can be mathematically represented as follows

$$T = \sigma(W_{zz}z_t + W_{hh}h_{t-1} + b_g)$$

$$\tilde{h}_t = tanh(U_{zz}z_t + U_{hh}(r_t \odot h_{t-1}) + b_h)$$

$$h_t = (1-T) \odot h_{t-1} + T \odot \tilde{h}_t$$

Where:

- z represents the input sequence to the RNN.

- $\sigma(r) = \frac{1}{1+exp(-r)}$ is the sigmoid activation function.
- $\tilde{h_t}$ is the candidate activation.
- T is the update gate, which controls how much of the previous hidden state h_{t-1} to retain and how much of the candidate activation $\tilde{h_t}$ to incorporate into the current hidden state h_t.
- r_t is the reset gate, determining how much previous state information should be forgotten or reset.

In the Bi-GRU architecture, two separate GRU units process the input sequence in forward and backward directions. The forward GRU computes the hidden state $\overrightarrow{h_t}$ based on the current input z_t and the previous hidden state $\overrightarrow{h_{t-1}}$. Similarly, the backward GRU computes the hidden state $\overleftarrow{h_t}$ based on the current input z_t and the previous hidden state $\overleftarrow{h_{t-1}}$.

The final hidden state h_t of the Bi-GRU at time step t is obtained by concatenating the forward and backward hidden states $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$. This combined representation captures dependencies from past and future contexts, providing a more comprehensive representation of the input sequence. To parameterize a distribution over sequences, a softmax layer is applied to the output of the Bi-GRU at each time step t, yielding the probability distribution $p(u_t \mid z)$ over possible outputs. The distribution over the entire length of the T sequence is then defined as the product of the individual probabilities

$$p(u_1, \dots, u_T \mid z) = \prod_{t=1}^{T} p(u_t \mid z).$$

In the model architecture, the 3D Convolutional Neural Network (3DCNN) output serves as the input z to the Bi-GRU. By incorporating both spatial and temporal information extracted from the video data, the Bi-GRU effectively models the sequential nature of speech articulation, contributing to the accurate interpretation of lip movements and phonetic content.

E. ***Connectionist Temporal Classification (CTC) loss function:*** The connectionist Temporal Classification (CTC) loss function [23] is a powerful tool used in sequence labelling tasks, particularly in scenarios where the alignment between input and output sequences is unknown [24]. In the context of lip reading, CTC is employed to compute the probability of a sequence of tokens given the output of a neural network model without the need for explicit alignments between the input video frames and the corresponding labels.

In CTC, the model outputs a sequence of discrete distributions over token classes, augmented with a special "blank" token. The set of tokens the model classifies at a single time-step of its output is denoted by V (vocabulary), and the vocabulary augmented with the blank symbol is denoted by $\tilde{V} = V \cup \{blank\}$.

The CTC loss function defines a mapping function $B: \tilde{V}^* \rightarrow V^*$ that removes adjacent duplicate characters and blank tokens from a given string over the augmented vocabulary. For a label sequence $y \in V^*$, the probability p(y|x) is computed as the sum of the probabilities of all possible alignments of the label sequence to the input sequence x, where the length of the aligned sequences is equal to the number of time steps T in the model's output. Mathematically, the probability p(y|x) is computed as:

$$p(y \mid x) = \sum_{u \in B^{-1}(y) s.t. |u|=T} p(u_1, \dots, u_T \mid x)$$

Here, $B^{-1}(y)$ represents the set of all sequences that map to y after removing adjacent duplicate characters and blank tokens. The sum is computed efficiently using dynamic programming techniques, allowing maximum likelihood estimation of the label sequence given the input sequence.

In our architecture, the CTC loss function is applied after the model's output has been processed by a series of layers, including spatial pooling, Bi-GRU layers, and linear layers, followed by SoftMax activation. By leveraging the CTC loss function, model can effectively train and optimize its parameters to accurately predict label sequences from input video data without requiring explicit alignments.

4. EXPERIMENTAL SETUP

A. Preprocessing: In the preprocessing phase of our research, we meticulously prepared the GRID corpus dataset to ensure its compatibility with our lipreading model. The preprocessing pipeline involved several essential steps tailored to the datasetʹs characteristics and our model architectureʹs requirements.

Initially, the video data was loaded from the GRID corpus dataset. Each video file was parsed to extract individual frames and subsequently processed to focus on the mouth region, a critical aspect for lipreading tasks. This involved converting each frame to grayscale and applying cropping operations to isolate the mouth area.

In parallel, the textual data provided in the dataset, including phonetic transcriptions, underwent preprocessing. Characters from the transcriptions were mapped to numerical indices, facilitating their representation in a format suitable for model training and evaluation. This mapping process ensured that textual data could be effectively encoded and decoded during the training and inference phases.

Phonetic transcriptions were further processed to extract relevant phonemes while excluding irrelevant information such as silence. This step involved parsing alignment files to identify and map phonemes to numerical indices. By focusing solely on pertinent phonetic information, we ensured that the model could effectively learn from the textual cues provided in the dataset.

Finally, methods for creating Tensor Flow Datasets from lists of file paths were implemented to streamline data loading and integration within the Tensor Flow framework. These methods facilitated the seamless integration of the preprocessing pipeline into our training workflow, ensuring the dataset was appropriately formatted and ready for model training.

B. Training: During training, we employed the Adam optimizer with a learning rate 0.0001 to optimize the model parameters. Additionally, we implemented a custom learning rate scheduler function, which exponentially reduced the learning rate after a certain number of epochs, aiding in the convergence of the training process.

We monitored the modelʹs performance on a validation dataset throughout the training to prevent overfitting and ensure generalization to unseen data. We employed various callbacks, including Model Checkpoint for saving model weights, Learning Rate Scheduler for dynamically adjusting the learning rate, and a custom callback named Produce Example, which generated examples of original and predicted transcriptions at the end of each epoch for qualitative evaluation.

The training process was conducted over 100 epochs, with the model iteratively updating its parameters to minimize the CTC loss and improve its performance in transcribing lip movements accurately. Furthermore, the model was trained in batches of two videos, optimizing computational resources and expediting the training process while maintaining data diversity for robust learning. This comprehensive training regimen enabled us to optimize the model's parameters and evaluate its effectiveness in capturing temporal dynamics and extracting meaningful features from the input video data.

C. *Evaluation Protocol*: In our research, we rigorously evaluate the performance of our lipreading model using a comprehensive set of evaluation metrics encompassing accuracy, transcription quality, and error rates.

Accuracy is a fundamental metric that represents the overall correctness of the model's predictions. It measures the percentage of correctly predicted characters in the transcriptions compared to the ground truth, indicating the model's proficiency in accurately transcribing lip movements into textual representations.

Word Error Rate (WER) and Character Error Rate (CER) offer more granular assessments of transcription quality. WER quantifies the modelʹs accuracy at the word level by measuring the percentage of incorrectly predicted words in the transcriptions, accounting for substitutions, deletions, and insertions. Similarly, CER assesses accuracy at the character level, computing the percentage of incorrectly predicted characters in the transcriptions, providing a finer-grained evaluation of transcription precision.

Word-level accuracy complements WER and CER by measuring the percentage of correctly predicted words in the transcriptions relative to the total number of words. It offers a straightforward assessment of the model's ability to transcribe complete utterances accurately, disregarding individual character errors.

In our evaluation protocol, we conducted experiments on the GRID corpus dataset, focusing on videos from a single speaker. We employed a standard train-test split, allocating 90% of the videos for training and 10% for testing. This systematic approach ensured robustness and reliability in assessing the performance of our lipreading model, allowing us to make informed decisions and optimize model parameters.

5. RESULTS

Our lipreading model exhibited remarkable performance during evaluation, particularly highlighted by its impressive accuracy of 85.65%. This signifies the model's ability to effectively transcribe spoken words by interpreting visual cues from video data. The validation of this accuracy was further emphasized through the computation of additional metrics: an average Word Error Rate (WER) of 20.33% and an average Character Error Rate (CER) of 15.17%. These metrics underscore the model's robustness and precision in deciphering speech, showcasing its capability to convert visual lip movements into corresponding textual representations accurately.

4. CONCLUSION

This research paper presents a deep learning-based approach to lip reading, employing Convolutional Neural Networks (CNNs) for feature extraction and Bidirectional Long Short-Term Memory (Bi-LSTM) networks for sequence modelling. Integrating a 3D CNN architecture enables the model to capture both spatial and temporal dependencies in lip movements, facilitating accurate transcription of spoken language from visual cues. By leveraging the Connectionist Temporal Classification (CTC) loss function for end-to-end learning, our proposed architecture achieves an impressive transcription accuracy rate of 85.65% on the GRID corpus dataset. Additionally, the model demonstrates an average Word Error Rate (WER) of 20.33% and an average Character Error Rate (CER) of 15.17%. These metrics and high accuracy showcase the model's proficiency in accurately transcribing lip movements into textual representations. Through rigorous evaluation and experimentation, our study contributes to advancing automatic lip-reading systems, offering insights into the challenges and opportunities in this domain. The proposed architecture, combining 3D CNNs, Bi-LSTMs, and CTC loss functions, lays a solid foundation for future developments aimed at improving the accuracy and robustness of lip-reading systems, thus enhancing their usability and applicability in diverse real-world scenarios.

REFERENCES

[1]. Zhao, X., Yang, S., Shan, S., & Chen, X. (2020, November). Mutual information maximization for effective lip reading. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (pp. 420-427). IEEE.

[2]. Deng, L., & Yu, D. (2014). Deep learning: methods and applications. Foundations and trends® in signal processing, 7(3–4), 197-387.

[3]. Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. The journal of the acoustical society of America, 26(2), 212-215.

[4]. AuerJr, E. T., & Bernstein, L. E. (2007). Enhanced visual speech perception in individuals with early-onset hearing impairment.

[5]. Theobald, B. J., Harvey, R., Cox, S. J., Lewis, C., & Owen, G. P. (2006, September). Lip-reading enhancement for law enforcement. In Optics and Photonics for Counterterrorism and Crime Fighting II (Vol. 6402, pp. 24-32). SPIE.

[6]. Summerfield, Q. (1992). Lipreading and audio-visual speech perception. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 335(1273), 71-78.

[7]. Qu, L., Weber, C., & Wermter, S. (2022). Lipsound2: Self-supervised pre-training for lip-to-speech reconstruction and lip reading. IEEE transactions on neural networks and learning systems.

[8]. Potamianos, G., Neti, C., Luettin, J., & Matthews, I. (2004). Audio-visual automatic speech recognition: An overview. Issues in visual and audio-visual speech processing, 22, 23.

[9]. Chandrasekaran, C., & Ghazanfar, A. A. (2009). Different neural frequency bands integrate faces and voices differently in the superior temporal sulcus. Journal of Neurophysiology, 101(2), 773-788.

[10]. Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America, 120(5), 2421-2424.

[11]. Wand, M., Koutník, J., & Schmidhuber, J. (2016, March). Lipreading with long short-term memory. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6115-6119). IEEE.

[12]. Garg, A., Noyola, J., & Bagadia, S. (2016). Lip reading using CNN and LSTM. Technical report, Stanford University, CS231 n project report.

[13]. Son Chung, J., Senior, A., Vinyals, O., & Zisserman, A. (2017). Lip reading sentences in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6447-6456).

[14]. Petridis, S., Li, Z., & Pantic, M. (2017, March). End-to-end visual speech recognition with LSTMs. In 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 2592-2596). IEEE.

[15]. Shillingford, B., Assael, Y., Hoffman, M. W., Paine, T., Hughes, C., Prabhu, U., ... & de Freitas, N. (2018). Large-scale visual speech recognition. arXiv preprint arXiv:1807.05162.

[16]. Hong, X., Yao, H., Wan, Y., & Chen, R. (2006, December). A PCA-based visual DCT feature extraction method for lip-reading. In 2006 International Conference on Intelligent Information Hiding and Multimedia (pp. 321-326). IEEE.

[17]. Li, Y., Takashima, Y., Takiguchi, T., & Ariki, Y. (2016, June). Lip reading using a dynamic feature of lip images and convolutional neural networks. In 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS) (pp. 1-6). IEEE.

[18]. Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., & Ogata, T. (2015). Audio-visual speech recognition using deep learning. Applied Intelligence, 42, 722-737.

[19]. Takashima, Y., Kakihara, Y., Aihara, R., Takiguchi, T., Ariki, Y., Mitani, N., ... & Nakazono, K. (2015). Audio-visual speech recognition using convolutive bottleneck networks for a person with severe hearing loss. IPSJ Transactions on Computer Vision and Applications, 7, 64-68.

[20]. Ülkümen, B., & Öztürk, A. (2024). Detection of Emergency Words with Automatic Image Based Lip Reading Method. Intelligent Methods In Engineering Sciences, 3(1), 1-6.

[21]. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pp. 1097–1105, 2012.

[22]. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1725–1732, 2014

[23]. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling ´ unsegmented sequence data with recurrent neural networks. In ICML, pp. 369–376, 2006.

[24]. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In International Conference on Machine Learning, pp. 1764–1772, 2014.