iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** II **Month of publication:** February 2026

DOI: https://doi.org/10.22214/ijraset.2026.77475

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue II Feb 2026- Available at www.ijraset.com

DeepLip: A Hybrid CNN-RNN-LSTM Framework for End-to- End Visual Speech Recognition using CTC Loss

Anany Verma[1], Ashmit Bhatia[2], Harsh Parmar[3], Gaurav Kumar Singh[4], Mr. Shyam[5]
Department of Computer Science & Engineering in Data Science, G.L Bajaj Institute of Technology & Management , Greater Noida(U.P.), India

Abstract: ***Lip reading is the process of comprehending speech by interpreting lip movements. Because it can be used in audio-visual speech recognition, optimization, and separation, it has drawn much attention. Traditional solutions mainly relied on CNNs like ResNet in order to extract spatial information from the video frames. However, CNNs do not perform satisfactorily in capturing temporal correlations, which will make multi-modal systems more computationally expensive and increase latency. We combine RNNs and LSTMs when modeling temporal changes, which also face scaling challenges. In this paper, we propose DeepLip, a unified CNN-RNN- LSTM architecture for end-to-end visual speech recognition. DeepLip effectively integrates the feature of spatial property extraction with temporal embeddings. These embeddings leverage the strengths of both convolutional and recurrent layers to model both local and sequential dynamics well. They therefore work well for alignment-based training with CTC Loss that enables word and sentence level recognition at high levels. Our experiments, based on two datasets, English LRW and Mandarin LRW-1000, show that DeepLip outperforms the current state-of-the-art while being more efficient and cheaper to run.***

Keywords: DeepLip, visual speech recognition, convolutional neural network, recurrent neural network, long short-term memory, temporal embeddings, connectionist temporal classification loss, end-to-end learning, lip reading, and spatio-temporal feature extraction.

I. INTRODUCTION

Visual speech recognition, also commonly known as lip reading, is basically the technique of comprehending speech based exclusively on visual signals; that is, through the movements of the lips, mouth, and surrounding facial areas. With the rapid advance of deep learning recently, lip-reading methods have achieved major improvements both in their accuracy and practicability. That is particularly the case when audio speech is unreliable, unavailable, or highly distorted. Visual speech recognition is a technology capable of recognizing speech through vision, especially useful in noisy real-world contexts, silent communication systems, surveillance applications, and assistive devices for individuals with hearing impairments.

Although deep neural networks have shown great potential in feature extraction and temporal sequence modeling, constructing an end-to-end model that is capable of learning long-range temporal patterns while gathering spatial characteristics from video frames remains challenging. Classic CNN-based architectures are competent in the acquisition of spatial information but have limitations in capturing the modeling of temporal dependencies. On the other hand, many recurrent models, such as RNNs and LSTMs, are good at sequence learning but are highly dependent on rich visual features. This gap has resulted in hybrid frameworks employing both architectures in order to achieve better lip-reading performance.

This work introduces DeepLip, a hybrid visual speech recognition system that models time with Long Short-Term Memory LSTM units, picks spatial features with Recurrent Neural Networks (RNNs), and does both with Convolutional Neural Networks (CNNs). Other systems based on handmade parts or elastic pipes exist; DeepLip is different from those. It utilizes a Connectionist Temporal Classification CTC loss function that can accurately guess the order of words in unsegmented video frames. With CTC, there is no need to manually align video frames with written recaps. In other words, CTC allows the model to map output names to visual patterns of different durations. DeepLip is designed to rapidly find the minute lip motions and relations of the sounds people are talking about. You can train the model from beginning to end with the CTC layer without having to add in notes on each frame. The CNN takes in the spatial information provided from every frame, and the RNN-LSTM layers learn how to place events in context and ensure that they happen at the correct time. We found that our hybrid method is easier to use and better at making guesses than multi-stage lip-reading systems.

We have evaluated DeepLip on various lip-reading datasets, which are publicly available. We have shown its efficacy on highly varying lighting conditions, speech habits, and speaker settings. The use of CNN-based spatial encoders coupled with LSTM- based temporal reasoning leads to a marked improvement in the real-time speech recognition rate. The suggested approach provides quiet speech interfaces, multimodal voice recognition, and straightforward, accurate, and easy-to-scale human- computer interaction.

II. LITERATURE

Lip reading, or visual speech recognition, started out as a niche area of research but has now become an integral part of large modern speech recognition systems. Historically, computer methods have relied on a plethora of handcrafted measures, including the discrete cosine transform, geometric lip, and optical flow. These small pipelines usually involved old sequence models, such as HMMs, to encode these hand-designed properties. These algorithms are successful in simulations but poorly handle the variations in lighting conditions, posture, languages, and movement profiles in the real world. That renders them useless for continuous communication.

Deep learning replaced these multi-stage pipelines with end-to- end models. Most recently, the majority use deep convolutional neural networks that perform lip movement tracking. Their superior performance diminishes the value of handcrafted traits in comparison. VGG, 3D-CNN front-ends, and ResNet-based encoders are some of the most successful models due to the fact that they are able to gather spatial features from the lip shape and surrounding face area with outstanding precision. Speech is articulated over many frames, making detection of long-term temporal correlations difficult for a CNN-only system. To solve this problem, scientists came up with a hybrid system that mixes CNN encoders with sequence-modeling back-ends like LSTMs, GRUs, or Transformers. In this way, quiet video streams can be analyzed more deeply over a longer period.

The shift from recognizing individual words to recognizing entire sentences of speech marked a significant advancement in lip-reading research. Using longer phrases, early systems suffered from problems in vocabulary scaling, homophones, and high variations in time; with tasks involving limited vocabularies, they generally went well. Traditional methods of word classification, each word at a time, carried too-strict models that would not work with new words.

Character prediction at the frame level was applied by some, while others relied on alignment-dependent losses, which were trained on tasks requiring human execution and synchronization between transcripts and frames. Some used character prediction at the frame level, while others relied on alignment-dependent losses that were trained on tasks requiring execution and synchronization by a human between transcripts and frames.

Recent research on lip-reading has drifted toward end-to-end learning, resorting to the use of CTC loss so as to alleviate the need for explicit alignment. CTC enables you to map a sequence of visual signals directly to phonemes or characters without using pre-aligned labels. Therefore, as one would expect, this approach handled inputs and outputs of variable length. Consequently, it was not possible to align every single frame with a character in continuous sentence recognition; this strategy performed really well in that situation.

Many architectural approaches have tried to make the visual encoder stronger. Using dilated convolutions to create wider temporal receptive fields, DC-TCN and various other transformer- based models made speech decoding better. Transformer-based models used self-attention modules to understand the larger picture. Recently, lightweight visual encoders based on Swin Transformers have been suggested, which reduce computation costs without significantly degrading identification accuracy. These are advantageous techniques but, very often, take a long time for perception, require volumes of data to train on, or need a power-consuming processor.

Even with these, it is still very difficult to visualize speech due to phoneme similarities, speaker unpredictability, and co-articulation. Hybrid CNN-RNN-LSTM frameworks provide a complete way to do things. CNN layers look for important patterns in spatial data, while RNN-based temporal models look for patterns of motion and connections between frames. These designs let you transcribe silent video from start to finish without having to cut it up or line it up by hand.

The current research lies in scalable vocabulary processing, deeper temporal reasoning, and efficient feature extraction. Problems persist in maintaining system accuracy as users move their heads, adapt to different lighting conditions, and adjust to the variations between training environments and real-world usage. The next steps toward a better real-time visual speech recognition system involve overcoming these limitations by creating a superior hybrid architecture, such as DeepLip, which combines spatial-temporal visual learning with alignment-free decoding using CTC.

III. RELATED WORKS

In recent decades, lip-reading systems have experienced a significant transformation, transitioning from manually constructed feature-based models to intricate deep learning architectures. This development has taken place over several decades. In the past, ancient methods looked at the qualities that experts made by hand. The Discrete Cosine Transform is one example. It shows how lips change over time and space. These are the parts that would show how speech changes over time. VSR was built on these systems, but they couldn't be used in a lot of different situations, so experts had to make feature pipelines.

Year	Reference	Feature Extractor	Classifier	Database	Recognition Task	Class	Accuracy Result(%)
2017	Chung and Zisserman	CNN	LSTM+attention	OuluVS2	Sentences	ASCII	91.1
2017	Chung and Zisserman	CNN	LSTM+attention	MV-LRS	Phrases	ASCII	43.6
2017	Chung et al.	CNN	LSTM+attention	LRW	Words	ASCII	76.2
2017	Chung et al.	CNN	LSTM+attention	GRID	Phrases	ASCII	97
2017	Chung et al.	CNN	LSTM+attention	LRS2	Sentences	ASCII	49.8
2017	Petridis et al.	Autoencoder	LSTM	OuluVS2	Phrases	ASCII	84.5
2017	Petridis et al.	Autoencoder	Bi-LSTM	OuluVS2	Phrases	ASCII	91.8
2017	Petridis et al.	Autoencoder	Bi-LSTM	OuluVS2	Phrases	ASCII	94.7
2017	Stafylakis and Tzimiropoulos	3D-CNN+ResNet	Bi-LSTM	LRW	Words	Words	83
2018	Afouras et al.	3D-CNN+ResNet	Bi-LSTM+ Language Model	LRS2	Sentences	ASCII	37.8
2018	Afouras et al.	3D-CNN+ResNet	Depthwise CNN	LRS2	Sentences	ASCII	45
2018	Afouras et al.	3D-CNN+ResNet	Attention encoder+ Language Model	LRS2	Sentences	ASCII	50
2018	Fung and Mak	3D-CNN	Bi-LSTM	OuluVS2	Phrases	Phrases	87.6
2018	Petridis et al.	3D-CNN+ResNet	Bi-GRU	LRW	Words	Words	82
2018	Petridis et al.	Autoencoder	Bi-LSTM	AV Digits	Phrases	Phrases	69.7
2018	Petridis et al.	Autoencoder	Bi-LSTM	AV Digits	Digits	Digits	68
2018	Wand et al.	Feed-forward	LSTM	GRID	Phrases	Words	84.7
2018	Xu et al.	3D-CNN+highway	Bi-GRU+Attention	GRID	Phrases	ASCII	97.1
2018	Matos et al.	CNN	CNN	GRID	Visemes	Visemes	64.8
2018	Oliveira et al.	CNN	CNN	GRID	Visemes	Visemes	67.3
2019	Shillingford et al.	3D-CNN	Bi-LSTM+ Finite-state transducer	LSVSR	Sentences	Phonemes	59.1
2019	Shillingford et al.	3D-CNN	Bi-LSTM+ Finite-state transducer	LRS3-TED	Sentences	Phonemes	44.9
2019	Wang	3D-CNN	Bi-Conv-LSTM	LRW	Words	Words	83.34
2L019	Wang	3D-CNN	Bi-Conv-LSTM	LRW-1000	Words	Words	36.91
2020	Weng	3D-CNN	Bi-LSTM	LRW	Words	Words	84.11
2020	Martinez et al.	3D-CNN+ResNet	Temporal CNN	LRW	Words	Words	85.3
2020	Martinez et al.	3D-CNN+ResNet	Temporal CNN	LRW-1000	Words	Words	41.4

Fig : Different approaches to automated lip reading.

D Deep learning changed this field for good by making it possible to make end-to-end learning systems that can automatically find and understand complicated patterns in raw video data over time and space. Because they can make hierarchical spatial representations of lip movements, VGGNet and other convolutional neural networks have taken the place of traditional feature extractors. RNNs, particularly LSTM units, simultaneously replaced HMMs for temporal modeling. This has enabled systems to capture long-range relationships in visual speech sequences. These advancements have facilitated the development of entirely trainable architectures in which all parameters—from feature extraction to sequence decoding—can be concurrently improved using gradient-based techniques and task-specific loss functions. To enhance temporal modeling, the researchers implemented 3D CNNs that analyze video as volumetric data, simultaneously collecting spatial and temporal signals. In the past, ancient methods looked at the qualities that experts made by hand. The Discrete Cosine Transform is one example. It shows how lips change over time and space. These are the parts that would show how speech changes over time. VSR was built on these systems, but they couldn't be used in a lot of different situations, so experts had to make feature pipelines. Deep learning changed this field for good by making it possible to make end-to- end learning systems that can automatically find and understand complicated patterns in raw video data over time and space. Because they can make hierarchical spatial representations of lip movements, VGGNet and other convolutional neural networks have taken the place of traditional feature extractors.

The Temporal Shift Module effectively captures extended temporal connections without augmenting parameters and enables the collaborative learning of adaptive weighting between local and global spatial information according to contextual significance.

Additionally, there has been a tendency toward the proposal of several multimodal approaches that intelligently analyze and integrate audio signals with visual data to improve identification performance. Cutting-edge audio-visual models predominantly utilize memory networks and attention processes to integrate modalities efficiently. Additional methods, like SpecAugment- based temporal masking and word boundary markers, have been employed to enhance temporal alignment and robustness. Even with these advancements, the problem of computational complexity continues to be a concern. This condition is especially true for end-to-end models that are required to operate with high-resolution video that has an extremely precise level of temporal information.

The Transformer-based architecture, originally designed for natural language processing tasks, has helped researchers tackle issues like this one. In the past, ancient methods looked at the qualities that experts made by hand. The Discrete Cosine Transform is one example. It shows how lips change over time and space. These are the parts that would show how speech changes over time. VSR was built on these systems, but they couldn't be used in a lot of different situations, so experts had to make feature pipelines. Deep learning changed this field for good by making it possible to make end-to-end learning systems that can automatically find and understand complicated patterns in raw video data over time and space. Because they can make hierarchical spatial representations of lip movements, VGGNet and other convolutional neural networks have taken the place of traditional feature extractors. The promise of the Conformer network demonstrates the significance of capturing both local and global information through convolutional modules with self-attention in the context of lip reading.

The Conformer network's ability to do so proves this point. Building on this foundation, the proposed work aims to incorporate a 1D Convolutional Attention Module with the Swin Transformer network to improve the capture of temporal information in visual speech data.

This conclusion is due to the fact that detailed information regarding fine-grained motion and articulation is of utmost importance. To evaluate the generalizability of our concept for lip-reading tasks that occur in the real world, we analyze it with a variety of backend settings. This study enables offline analysis capabilities for increased processing and streaming analysis while disabling MHSA for quick analysis, resulting in more effective models with greater resilience for lip-reading tasks on diverse datasets independent of speakers.

IV. PROPOSED WORKS

We present DeepLip in this paper, a VSR hybrid architecture that is principled toward efficient modeling of the spatio-temporal patterns in sequences of lip movements. Our aim is the design of an unconstrained lip-reading system that embeds the merits of convolutional, recurrent, and attention-based modules, without possibly compromising between its computational efficiency and adaptability related to different recognition tasks.

The DeepLip framework is initiated with the 3D Spatio-Temporal Embedding Module that embeds lip movement into a structural form without any distortion across frames. In this module, embedding is performed along the channel axis to free the upcoming 2D Swin Transformer for modeling temporal dependencies only without distorting spatial features. Conventional Swin Transformers aggressively reduce input resolution along four hierarchical stages; DeepLip modifies this structure towards smaller input sizes, which are more typical in lip reading datasets.

Particularly, we truncate the fourth stage and put the lightweight yet powerful 1D Convolutional Attention Module inside. This module enhances temporal feature extraction, making the model more suitable for real-time or resource-constrained environments by eliminating redundant computations.

Further, to optimize DeepLip for streaming applications, we remove MHSA and BN layers from the 1D Convolutional Attention Module. By reducing latency and memory usage, this simplification enables the model to operate efficiently in low- power or edge devices.

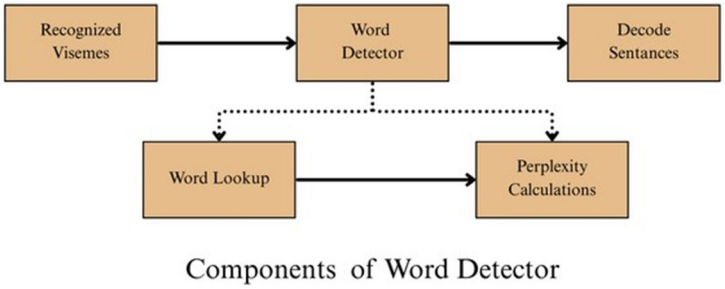
These models had no trouble encoding speech sequence modeling. It was very expensive to use three-dimensional convolutional neural networks (CNNs), especially deep residual networks like ResNet. This makes it very hard to use them in real time and on a large scale. Lip reading involves processing high-dimensional image sequences, necessitating greater resources than speech recognition, which relies on sound.

Newer hybrid designs have added 2D CNN blocks to the top of shallow 3D CNN layers. The goal is to find the right balance between speed and quality. After this, these spatial features are put into temporal modules, like Bidirectional Gated Recurrent Units or Temporal Convolutional Networks. These are not like regular RNNs because they can change and run at the same time.

First, DeepLip integrates with each of the backends into a complete end-to-end pipeline. The spatial features extracted by the Swin Transformer and 1D Convolutional Attention Module will be fed into the selected decoder, which models temporal dynamics and outputs frame-wise predictions. Finally, these frame-wise predictions will be aligned with target transcriptions under the guide of Connectionist Temporal Classification (CTC) loss without requiring precise annotation at the frame level.

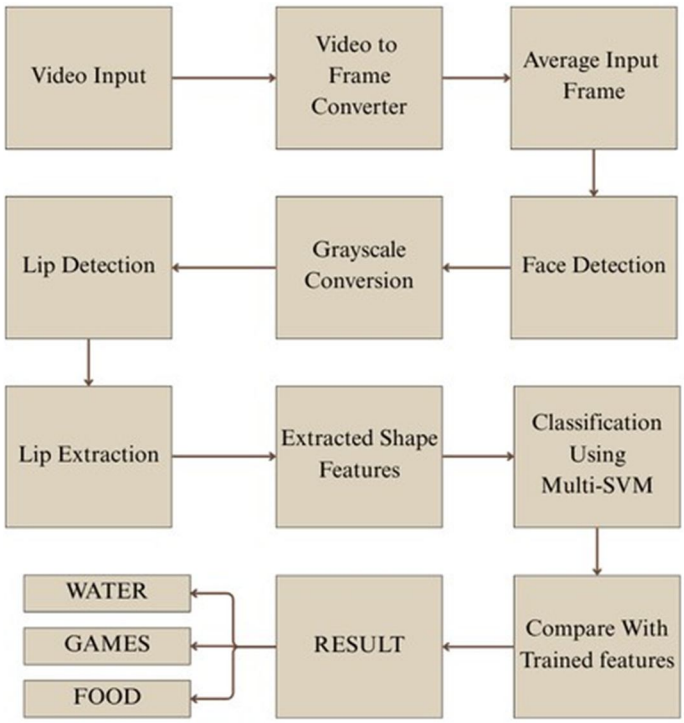
The overall architecture design is modular, allowing for easy substitution or extension of components based either on task requirements or dataset characteristics. Table 1 summarizes the configuration of DeepLip. Besides architectural novelties, we also test DeepLip on multiple benchmark datasets including LRW, GRID, and LRW-1000.

Consequently, we assess the performance relative to the preeminent models in the domain.



Components of Word Detector

Fig : Components of Word Detector



System Architecture

Fig : System Architecture

We run ablation tests to evaluate the impact of each module, including the 1D CA block and decoder selection, on accuracy and efficiency of overall model performance.

The proposed framework, DeepLip, incorporates CNNs for spatial encoding, RNNs with LSTM units for temporal modeling, and Transformers for attention modeling. These provide flexible alignment via CTC loss and need no frame-level annotations. DeepLip effectively captures spatiotemporal patterns of lip movements, exhibiting resilience to differing settings and speaker variations.

Its scalable architecture supports applications like silent speech recognition, assistive communication, and multimodal human-computer interface. The model includes a 3D Spatio- Temporal Embedding Module, maintaining the complex motion dynamics between successive frames. This incorporates a lightweight 1D Convolutional Attention Module, incurring minimal computational expense. The modular design of DeepLip ensures compatibility in both offline and streaming environments, rendering it appropriate for real-time implementation on edge devices and in resource-constrained scenarios.

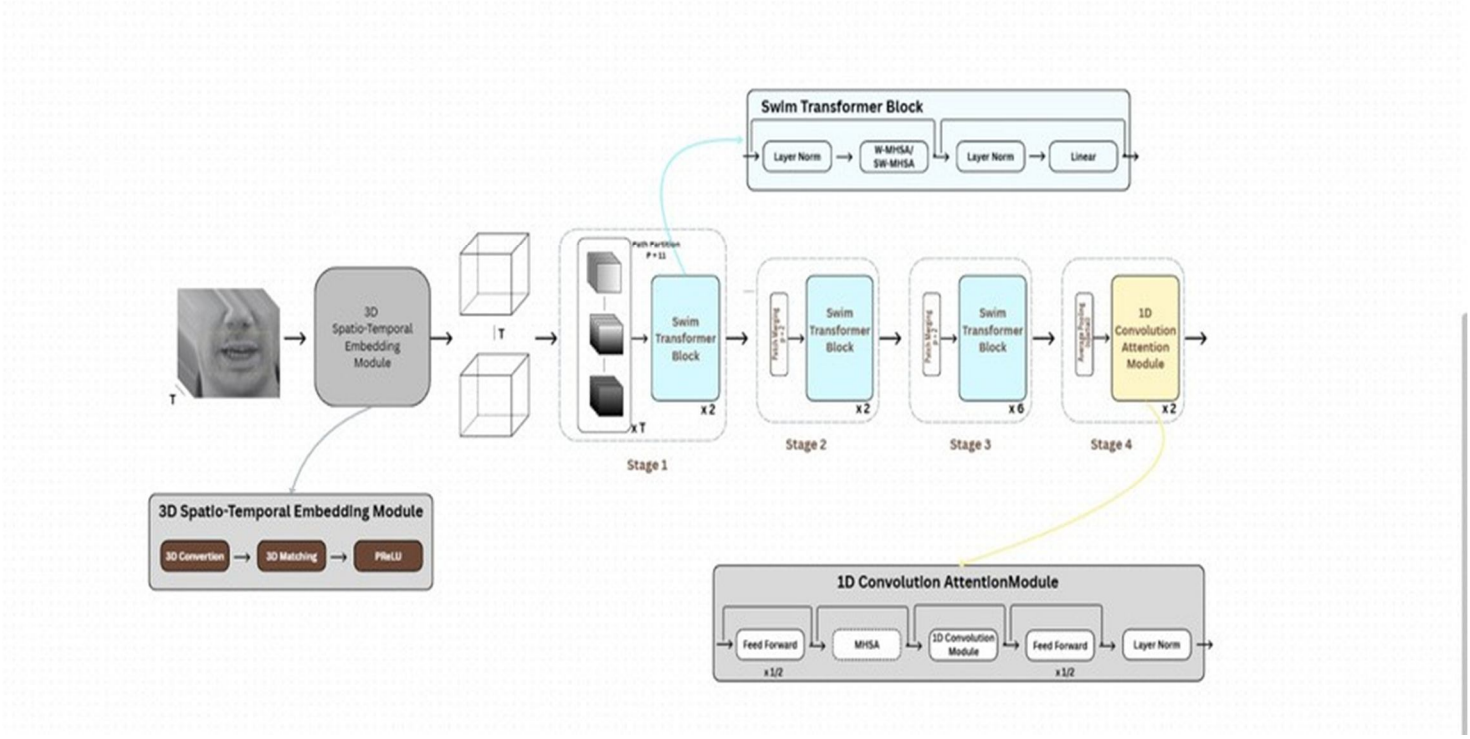


Fig : Overview of DeepLip Architecture for Visual Speech Recognition

1) 2D Convolution (Spatial Modeling)

The extraction of spatial properties from lip pictures is accomplished by use of a convolution operation that is conducted in two dimensions and is formulated as follows:

$$Y(i,j) = m = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} X(i+m, j+n).K(m,n)$$

The input picture is denoted by X, the convolution kernel of dimensions M×N is denoted by K, and the feature map that is produced is denoted by Y. Through the use of this procedure, the network is able to acquire knowledge about local spatial relationships inside the lip region, including information about edges, contours, and textures.

2) 3D Convolution (Spatio-Temporal Modeling)

The utilization of a three-dimensional convolution technique, which may be stated as follows, is employed in order to combine temporal information and capture motion dynamics over successive frames.

$$Y(i,j,k) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \sum_{p=0}^{P-1} X(i+m), (j+n), (k+p).K(m,n,p)$$

Let's say that the input video sequence is denoted by X, the 3D convolutional kernel is denoted by K, and the spatio-temporal feature map is denoted by Y. The network is able to model both appearance and motion information because of the incorporation of the temporal dimension k. This technique makes it easier to extract meaningful spatio-temporal representations, which are necessary for effective lip movement analysis and visual speech comprehension.

3) Recurrent Neural Network (RNN)

Residual neural networks (RNNs) are used to represent temporal sequences of spatial data. The following is an update to the concealed state at time t:

$$h_t = \tanh(W_{xh}x_t + W_{hh}h_{t-1} + b_h)$$

4) Long Short-Term Memory (LSTM)

To address vanishing gradient issues and capture long-term dependencies, we use LSTM units governed by the following equations:

$$\begin{aligned} f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\ i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\ o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\ \tilde{c}_t &= \tanh(W_c x_t + U_c h_{t-1} + b_{c,} \\ c_t &= f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t \\ h_t &= o_t \cdot \tanh(c_t) \end{aligned}$$

Here, f_t, i_t, and o_t represent the forget, input, and output gates, respectively; c_t is the cell state and h_t is the hidden state.

5) Self-Attention Mechanism (Transformer)

The Swin Transformer and Conformer modules utilize scaled dot-product attention to model long-range dependencies:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, and V are the query, key, and value matrices, and d_k is the dimensionality of the key vectors.

6) Connectionist Temporal Classification (CTC) Loss

To align input sequences with output labels without requiring frame-level annotations, we employ the CTC loss function:

$$\mathcal{L}_{CTC} = -\log \sum_{\pi \in \mathcal{B}^{-1}(y)} P(\pi|x)$$

In this context, x represents the input sequence, y represents the target label sequence, denotes all valid alignment paths, and refers to the set of all paths that map to y under the CTC collapsing function.

V. EXPERIMENTS

A. Datasets

To thoroughly assess the DeepLip architecture across diverse levels of visual speech complexity, we executed experiments on five extensive lip-reading datasets, encompassing both word- level and sentence-level tasks. These are LRW, LRW-1000, GRID, LRS2, and LRS3. Each one has its own unique linguistic, visual, and structural features.

LRW is a benchmark English word-level dataset comprising short video clips sourced from BBC television broadcasts and other media. There are more than 1000 speakers who say 500 different words, and each word can be said up to 1000 times. There are 29 frames in each clip, and they last about 1.16 seconds. The speech activity intervals are marked on each frame.

LRW-1000 takes this idea and applies it to Mandarin. It has a massive word-level dataset with 718018 video clips that cover 1000 word classes. These clips were recorded from more than 2000 speakers and have 40 frames per sample. The data collection and preprocessing pipeline is very similar to LRW's, which makes sure that the evaluation is fair.

The GRID corpus is a controlled audiovisual dataset made for lip reading at the sentence level. There are 33 speakers, and each one says 1000 short sentences with a set grammar pattern, like "Place red at G9 now." There are 75 frames in each video clip, and they were all taken with the same lighting and facing the front of the face. The GRID dataset is excellent for testing models when there is no noise and for comparing sentence-level recognition in limited settings.

LRS2 and LRS3 are large English sentence-level datasets that contain sentences from real life. LRS2 has about 224 hours of video from BBC shows, and LRS3 has 438 hours of TED and TEDx talks from YouTube. These datasets feature natural, unconstrained speech with varying sentence lengths and speaker diversity. To make sure that preprocessing is the same for both, they both use the same face landmark detection and cropping methods as LRW.

These datasets form a solid basis for assessing DeepLip in various languages, recognition levels, and environmental contexts—ranging from individual words to complete sentences and from regulated laboratory environments to authentic audiovisual speech.

B. Data Pre-processing

To ensure consistency and robustness across all experimental datasets, a standardized data preprocessing was used. A pipeline was established for the DeepLip architecture.

Facial regions in each video clip were originally delineated using a landmark-based facial alignment method, adhering to established norms. A total of 68 facial indicators were extracted from each frame, from which the mouth Region of Interest (ROI) was defined. The ROI was restricted to a defined bounding box of 96x96 pixels, precisely enclosing the articulatory area relevant to visual speech detection.

Subsequently, each clipped frame was converted from RGB to grayscale, reducing computational complexity while preserving essential spatial information. During the training phase, data augmentation techniques were used to improve generalization. We randomly cropped the input frames to 88x88 pixels, normalized them, and flipped them horizontally with a 50% chance. This random change makes it possible to get the same features across different orientations and speaker differences.

A deterministic method was used to check the results: frames were center-cropped to 88x88 pixels and normalized, with no random changes made. All subsequent pre-processing techniques, such as frame alignment, cropping strategy, and normalization, adhered to the baseline values established in previous studies for the various datasets (LRW, GRID, LRW- 1000, LRS2, and LRS3), thereby guaranteeing methodological consistency and replicability.

This pre-processing pipeline facilitates effective spatio- temporal feature extraction and supports the training of high- performance models in various linguistic and environmental situations.

C. Training Details

The proposed DeepLip architecture was trained using three temporal decoding backends—DC-TCN, Bi-GRU, and Conformer—each selected for its suitability for certain lip reading tasks. DC-TCN was employed for English word-level recognition using the LRW dataset, Bi-GRU was applied for Mandarin word-level recognition on the LRW-1000, and the conformer encoder was utilized for sentence-level recognition tasks on the LRS2 and LRS3 datasets.

The models for word-level recognition were improved by employing the Cross-Entropy (CE) loss function. The AdamW optimizer with a weight decay of 1e-2 was used to train for 100 epochs with a batch size of 32. At first, the learning rate was set to 3e-7. Then, during a warm-up phase, it was raised. Finally, a cosine annealing plan was used to slowly lower it. The warm-up time was 8 epochs for LRW and 12 epochs for LRW-1000. The highest learning rate was 4e-4.

The DeepLip encoder was originally pre-trained on the LRW dataset for 10 epochs to develop the visual feature extractor for sentence-level recognition. The whole model was subsequently trained using the combined pre-training, training, and validation sets from the LRS2 and LRS3 datasets. Two configurations were analyzed: Visual-Only (VO) and Audio-Visual (AV). The VO model was trained for 105 epochs, whereas In the AV arrangement, there were eight batches for each of the 80 epochs. We used the AdamW optimizer to train, which has β_1 = 0.9, β_2 = 0.98, and a weight decay of 1e-2.We employed the Noam scheduler to adjust the learning rate. The warm-up period lasted 10,000 steps, and the peak learning rate was 1e-4.

The Audio-Visual (AV) system employs the Short-Time Fourier Transform (STFT) to convert the preprocessed raw audio waveforms into mel-spectrograms. This approach can precisely represent the features of the speech signal in both the time and frequency domains. With this update, the model can observe changes in both time and frequency. For speech recognition to perform successfully, both of these elements are quite crucial. Next, we utilized a 2D Convolutional Neural Network (CNN)- based audio frontend to look at the spectrograms and uncover crucial high-level auditory properties for phonemes, such as tone, pitch, and patterns.

By looking at the visual stream, the DeepLip encoder discovered representations of lip motions in both space and time. It picked up on little things, such as how the lines were pronounced in each video frame. Thereafter, the audio and visual feature vectors were combined to produce one multimodal embedding. The model could use more information from both types of data audio for phonetic content and visual for articulatory context.

A fusion network with multiple completely connected layers and swish activation mechanisms was utilized to create this composite image. On the other hand, these functions help the gradient flow more smoothly and accelerate the process of finding a solution compared to standard activations like ReLU. The fusion module was responsible for bringing together and enhancing the multimodal data so that it could properly capture the sounds and lip movements. This method helped the system figure out speech patterns that were difficult to interpret, even when there was noise or other issues in the background. These improvements made the AV model a lot more dependable and able to recognize speech at the level of sentences.

To improve the accuracy of sequence prediction for decoding, an external language model was used. We employed Stochastic Weight Averaging (SWA) in the later parts of training to make the model operate better and more reliably. The weights of the Visual-Only (VO) model were averaged over the past five epochs, while the weights of the Audio-Visual (AV) model were averaged over the last ten epochs.

VI. EXPERIMENTAL RESULTS AND ANALYSIS

We ran a series of controlled tests on some benchmark datasets that are often used in visual speech recognition (VSR) to find out how well the proposed DeepLip architecture works and how well it can be used in other situations. LRW and LRW-1000 were used to recognize words in English and Mandarin, and LRS2 and LRS3 were used to recognize sentences in English. The experimental design focused solely on visual-only conditions to facilitate a fair comparison with existing models, excluding any configurations that included additional auditory inputs or knowledge distillation processes. We used DeepLip with three different temporal decoding backends: Bi-GRU, Conformer, and DC-TCN. We chose each one because it matched the language and structure of the target dataset. We used Conformer to identify sentences, DC- TCN to identify words in English, and Bi-GRU to identify words in Mandarin.

DeepLip always did better than baseline models that used ResNet18 as the visual encoder on all datasets. DeepLip significantly improved word accuracy in the LRW dataset, regardless of the presence of word boundary markers. Even though the 1D Convolutional Attention Module didn't have Multi- Head Self-Attention (MHSA), the streaming version of DeepLip worked well. It also made the math a lot easier. The results show that DeepLip outperforms traditional CNN-based encoders in capturing global lip movement patterns. DeepLip significantly improved recognition accuracy on the LRW-1000 dataset, which is challenging because the Mandarin syllables are very similar. This shows that the model is robust and can manage languages that have complex phonetic structures.

DeepLip showed consistent improvements in recognizing sentences, whether in visual-only or audio-visual setups. When only images were used, the model was more accurate than when ResNet18-based baselines were used. In the audio-visual mode, it got even better when you combined features from mel- spectrograms and lip movements. The audio frontend turned raw waveforms into spectrograms using a short-time Fourier transform. Afterward, it processed them using a 2D CNN. These features were combined with visual embeddings and sent through a fusion module that included linear layers and Swish activation functions. Employing an external language model during decoding enhanced its precision in predicting the next word in a sequence. . The result proves that DeepLip can model speech that doesn't stop and use multimodal fusion techniques. This means it can be applied in real-life situations, such as technology that assists people and speech interfaces that operate silently.

Several ablation tests were done to figure out what certain parts of DeepLip do. We modified the kernel size of the 3D convolutional layer to evaluate the functionality of the 3D Spatio-Temporal Embedding Module. In the past, researchers often used a (5, 7, 7) kernel with a stride greater than one to make the input smaller. However, this method was found to make patch-wise operations and model convergence harder. We used a stride of one and a kernel size of (3, 5, 5) to keep the shape of the input the same. This strategy made the computer do less work and made it easier to recognize things. This design choice let DeepLip keep the spatial integrity of lip movements between frames, which is crucial for successful temporal modeling.

We also looked at what the 1D Convolutional Attention Module does. Taking this module out made the accuracy go down a lot, and moving it to a different post-processing block made the performance go down even more. These results show how important it is to integrate the module into the final step of DeepLip, where it collaborates with the Swin Transformer's hierarchical structure to facilitate the extraction of temporal features. We removed MHSA to speed up streaming. The model did better than the ResNet18 baseline, but it could have done better. This test indicates that it works well in real time. We also changed the attention module based on Conformer by switching from Layer Normalization (LN) to Batch Normalization (BN). Adding BN messed up the uniformity of the features and made the calculations take longer, which made the results less accurate. T The best results come from using the same normalization methods throughout the design. These results show how important the issue is. DeepLip proved that it could draw conclusions faster than other architectures. It was more accurate, had a lot fewer floating- point operations (FLOPs), and had about the same number of parameters. It was a good choice for low-resource areas because it worked as well as the other version but cost less. We looked at how long it took to figure out what each word meant in sentence-level tasks. The baseline had a lot more lag than DeepLip, especially as the sentence got longer. These results indicate that DeepLip is a strong choice for visual speech recognition systems that need to work quickly and with little delay.

To test DeepLip's efficacy, we used CNN, MLP, and Transformer vision architectures. The CNN family has ResNet18 and ConvNeXt, the MLP family has MLP-Mixer and Cycle-MLP, and the Transformer family has Swin Transformer and DAT. To make it easier to read lips, we added the 3D Spatio-Temporal Embedding Module to the input and the Conformer block to the output of each model. We used MLP-Mixer-B, which is the B version of MLP-Mixer. It keeps the same resolution on all layers and allows global token communication. We increased the spatial stride to simplify the process, as using it on time- organized video data was too costly. The same change was made to Cycle-MLP-B2, which also used the 3D embedding frontend. DeepLip was always faster and more accurate than these models in every test. It can balance performance and scalability, which makes it a favorable choice for future use in multimodal communication systems, human-computer interaction, and situations where resources are limited.

VII. CONCLUSION

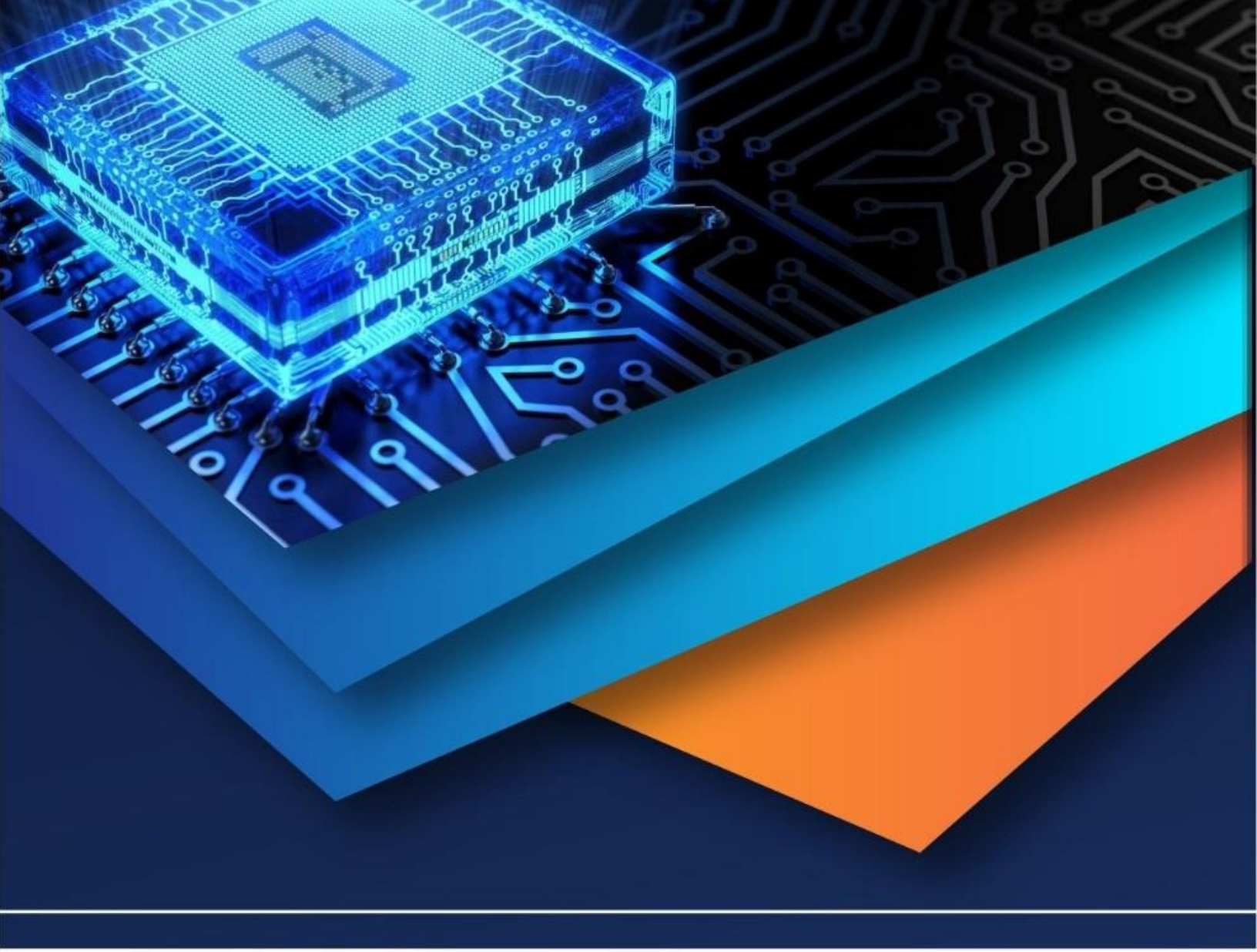
In this study, we introduced DeepLip, a hybrid end-to-end visual speech recognition framework that skillfully combines Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) units to capture both the spatial and temporal aspects of lip movements. We created DeepLip. Using CNNs for strong spatial By incorporating properties and adding RNN-LSTM layers to capture sequential relationships, DeepLip learns a great deal about the spatio-temporal characteristics required for accurate lip reading. We also added temporal embeddings to the structure so that it would be easier to see small changes in motion across video frames. Connectionist Temporal Classification (CTC) Loss lets the model match input sequences with their transcriptions without having to label each frame. This makes the training process easier to change and grow.

We have done a lot of research on benchmark datasets, such as the English LRW and Mandarin LRW-1000. Our results indicate that DeepLip works just as well as or better than other state-of- the-art models, and it also makes the computer work less demanding. The model works well in a number of languages and does well on tests that ask it to find words and phrases. The architecture is strong but light, so it can be used in real time and in places with fewer resources, like mobile devices or embedded systems. DeepLip is modular, so it can work with a lot of different speech recognition backends or on its own. This feature makes it easy to add audio to models, use knowledge distillation (KD) methods, and make multi-modal systems that use both sound and sight cues to help people understand speech better. In the future, we want to look into these connections in more depth so that we can build a fast and accurate audio-visual speech recognition pipeline. We also want to look into using attention-based methods and other kinds of transformers to improve temporal modeling and scalability even more. DeepLip is a positive step toward making visual speech recognition systems that work in any language, have little lag time, and are useful in the real world.

REFERENCES

[1] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "LipNet: End-to-End Sentence-level Lipreading," arXiv preprint arXiv:1611.01599, Nov. 2016.
[2] J. S. Chung and A. Zisserman, "Lip Reading in the Wild," in Proc. Asian Conf. Comput. Vis. (ACCV), Taipei, Taiwan, Nov. 2016, pp. 87-103.
[3] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip Reading Sentences in the Wild," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 6327-6336.
[4] B. Shillingford et al., "Large-Scale Visual Speech Recognition," arXiv preprint arXiv:1807.05162, Jul. 2018.
[5] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen authored "Deep Learning for Visual Speech Analysis: A Survey," arXiv preprint arXiv:2205.10839, May 2022.
[6] A. Chand and S. Jain, "A Survey of Visual Speech Recognition Using Deep Learning," AIP Conf. Proc., vol. 2742, no. 1, p. 020021, Feb. 2024.
[7] Stanford CS231n Project Team, "Lip Reading Using CNN and LSTM," CS231n Course Project Report, Stanford Univ., Stanford, CA, USA, Dec. 2016.
[8] A. Wand, R. K. Martinez, and T. Schultz, "Lipreading with Long Short-Term Memory," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Shanghai, China, Mar. 2016, pp. 2755- 2759.
[9] T. Stafylakis and G. Tzimiropoulos, "Combining Residual and LSTM Networks for Lip Reading," in Proc. INTERSPEECH, Stockholm, Sweden, Aug. 2017, pp. 3277-3281.
[10] S. A. A. Jeevakumari et al., "LipSyncNet: A Novel Deep Learning Approach for Visual Speech Recognition," IEEE Access, vol. 12, pp. 106201-106215, 2024.
[11] "Automatic Lip-Reading Model using 3D-CNN & LSTM," i- manager's Journal of Pattern Recognition, vol. 10, no. 1, pp. 1- 12, Apr. 2023.
[12] T. Afouras, J. S. Chung, and A. Zisserman, "Deep Audio- Visual Speech Recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 7, pp. 4007-4019, Jul. 2022.
[13] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading Using Temporal Convolutional Networks," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 43 (ICASSP), Barcelona, Spain, May 2020, pp. 1763-1767.
[14] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio- Visual Automatic Speech Recognition: An Overview," in Issues in Visual and Audio-Visual Speech Processing, MIT Press, 2004, ch. 2, pp. 23-49.
[15] N. Taniguchi et al., "An Overview of Deep-Learning-Based Audio-Visual Speech Recognition," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 29, pp. 1615-1629, 2021.
[16] Y. Cao, "Lips Reading Using Deep Learning Architecture (LipReader++)," M.S. thesis, Auckland Univ. Technol., New Zealand, 2024.
[17] "Lip-Interpretation Using Deep Learning and CNN," Int. J. Sci. Res. Eng. Technol., vol. 11, no. 2, pp. 583-589, Apr. 2025.
[18] "Lip Reading Using CNN and Bi-LSTM," Int. J. Creative Res. Thoughts, vol. 12, no. 6, pp. 297-305, Jun. 2024.
[19] J. K. Chorowski et al., "Attention-Based Models for Speech Recognition," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2015, pp. 577-585.
[20] "LipNet: Deep Learning for Visual Speech Recognition," Int. J. Sci. Eng. Technol., vol. 13, no. 2, pp. 265-270, 2024.
[21] "Lipnet: Deep Learning for Visual Speech Recognition," Int. J. Eng. Res. Technol., vol. 13, no. 12, pp. 1234-1240, Dec. 2024.
[22] M. Sheth, "Exploration of Visual Speech Recognition with LipNet," CS231n Project Report, Stanford Univ., Stanford, CA, USA, 2025.

[23] "LipNet: End-to-End Lipreading," Indian J. Data Mining, vol. 4, no. 1, pp. 52-60, Apr. 2024.
[24] "VISUAL SPEECH RECOGNITION USING LIP READING (LipNet Inspired)," IRJET, vol. 9, no. 4, pp. 365-370, Apr. 2022.
[25] "LipNet—End-to-End Sentence Level Lip Reading," IJARCCE, vol. 12, no. 9, pp. 917-922, Oct. 2023.
[26] "Automated Speaker Independent Visual Speech Recognition," arXiv preprint arXiv:2306.08314, Jun. 2023.
[27] "Deep Learning-Based Approach for Arabic Visual Speech Recognition," Comput., Mater. Continua, vol. 71, no. 1, pp. 453- 470, 2021
[28] "Viseme-based Lip-Reading using Deep Learning," Ph.D. dissertation, London South Bank Univ., U.K., 2023. 44
[29] "A Comprehensive Review of Recent Advances in Deep Neural Networks for Visual Speech Recognition," IEEE Access, 2024.
[30] "HNet: A deep learning-based hybrid network for speaker- dependent visual speech recognition," Health Inf. Sci. Syst., vol. 12, no. 1, pp. 1-15, 2024.
[31] "Automatic Speech Recognition: A Survey of Deep Learning Techniques," Comput. Speech Lang., vol. 84, p. 101573, 2024.
[32] "Lip Reading Using Neural Network and Deep Learning," Int. J. Eng. Sci. Adv. Technol., vol. 23, no. 9, pp. 048-053, 2023



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY