Results in Control and Optimization 20 (2025) 100608

Contents lists available at ScienceDirect

Results in Control and Optimization

journal homepage: www.elsevier.com/locate/rico

Deep learning based 3D residual convolutional and Multi-Head Attention (3D-RMA) for lip-reading

Archana Chaudhari [a,1], Masuk Abdullah [b,*,1], Vivek Deshpande [c,1], Tushar Zanke [a], Samrudhi Wath [a,1], Snehashish Mulgir [a,1], Stuti Jagtap [a,1]

[a] Instrumentation Engineering, Vishwakarma Institue of Technology, Upper Indira Nagar, Bibwewadi, Pune, 411037, Maharashtra, India
[b] Department of Vehicles Engineering, Faculty of Engineering, University of Debreceen, Otemeto strt. 2-4 4028, Debreceen, Hungary
[c] Vishwakarma Institute of Information Technology, Kondhwa, Pune, India

ARTICLE INFO

Keywords:
Lip reading
Deep learning
3D convolutional networks
Multi-Head Attention
Connectionist temporal classification (CTC)

ABSTRACT

Lip reading, an essential yet intricate facet of communication, has seen notable progress through the application of advanced deep learning techniques. This research introduces a deep learning-based lip-reading model that integrates Conv3D layers, Multi-Head Attention mechanisms, Bidirectional LSTMs, and a Dense output layer, combined with a custom Connectionist Temporal Classification (CTC) loss function. Our comprehensive data preprocessing pipeline extracts video frames, normalizes pixel values, and converts textual alignments into numerical tokens, enabling effective model integration. The model architecture is carefully structured to capture spatiotemporal features, with Conv3D layers addressing spatial information, while Multi-Head Attention mechanisms and Bidirectional LSTMs effectively manage temporal dependencies. Residual connections and Max-Pooling layers are incorporated to enhance feature extraction and abstraction, supporting improved performance. The use of Layer Normalization and Dropout layers contributes to stable learning and mitigates overfitting. Through extensive training and evaluation, our model demonstrates a 96% accuracy rate in decoding lip movements and predicting corresponding words. The implementation of the CTC loss function allows for effective handling of variable-length sequences, further contributing to the model's performance. This research provides a technically sound approach to lip reading, contributing to the advancement of visual speech recognition and offering potential benefits for communication accessibility among individuals with hearing impairments.

1. Introduction

Lip reading is the ability to receiving speech by observing the movements of lips and that method useful for those individuals who have some problem with ears [1]. This ability is very important for enhancing the interaction in everyday life, whether it is small talk or business discussions etc. But the process is highly challenging in lip reading. The movements of the lips are small and speaking as such is highly relative. That is why lip reading is quite difficult. Factors such as accent, speaking rate, and individual pronunciation variations significantly hinder accurate lip reading. Over the recent past however, improvements in deep learning have gone a long way in teaching computers how to interpret lip movements [2]. This is because with the aid of machine learning

* Corresponding author.
E-mail addresses: archana.chaudhari@vit.edu (A. Chaudhari), masuk@eng.unideb.hu (M. Abdullah), vivek.deshpande@viit.ac.in (V. Deshpande), tushar.zanke21@vit.edu (T. Zanke), samrudhi.wath21@vit.edu (S. Wath), snehashish.mulgir21@vit.edu (S. Mulgir), stuti.jagtap21@vit.edu (S. Jagtap).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.rico.2025.100608
Received 16 January 2025; Received in revised form 2 June 2025; Accepted 22 August 2025
Available online 29 August 2025

especially neural networks, researchers have been able to construct systems that analyze not only the spatial details of the lip image in individual frame but also the temporal details of the images. CNN and RNN-based approaches are among the most widely adopted methods with CNNs being more effective in handling dynamic features of speech. This research proposes a new model to help lip reading for better recognition by predicting sequences of characters which are related to movements of lips in captured videos. Therefore, the proposed model adopts a carefully designed architecture based on 3D Convolution layers alongside Bidirectional LSTM networks supported by the Multi-Head Attention mechanisms. A third component of the proposed model is the Dense output layer These components enable the model to learn both, the spatial features of the model and the temporal or time-related features of the speech. To ensure separated input data brings forth accurate predictions, it undergoes rigorous preparation—the videos are preprocessed through framing, normalization, and tokenization. The final combined model assists in solving every difficulty in lip reading and augments the general performance proficiencies.

2. Related work

This paper focuses on lip reading using facial feature extraction and deep learning. It employs deep learning techniques such as (CNNs) for feature extraction from facial images. The accuracy achieved is 85.5%. Also they have shown the graphical representation of CNN and LSTM . The novelty lies in the method of integrating facial feature extraction with deep learning algorithms to improve lip reading accuracy. However, the study does not address the challenges posed by variations in head poses, lighting conditions, and individual differences, which can impact the robustness of the model [3]. This paper proposes an automatic lip-reading system using a combination of deep CNNs and attention-based LSTM networks. The novelty lies in the integration of CNNs for feature extraction and attention-based LSTM for sequence modeling. The results show that the accuracy of the proposed model is 88.2% in the test dataset and 84.9% for the contrastive model. Despite the improved performance, the model's reliance on large datasets and computational resources may limit its applicability in real-time systems [4]. This paper introduces an automated lip-reading system using deep learning and neural networks, with a focus on real-time word prediction. Object detection, CNNs, and Keras are utilized. A lightweight model outperforms EF-3 architecture, with Model D achieving the highest accuracy. The system is implemented in a web application using object detection and the Keras library. While the system outperforms the EF-3 architecture, the study does not provide quantitative performance metrics or address potential issues related to scalability and generalization across diverse datasets [5]. An overview of the major studies on lipreading, from traditional methods to deep learning methods, is given in this work. It does a thorough structural analysis of popular deep learning algorithms and summaries the state-of-the-art lipreading databases, together with all of their detailed information and the methods applied to them. The paper highlights the advancements in deep learning but lacks a critical analysis of the limitations and challenges inherent in current lip reading technologies [6]. By employing a small number of coordinate points that indicate the basic form of the mouth instead of the massive volume of data, the author aimed to overcome the problems associated with training the neural network and test its classification capabilities. This work focused on a de-scoped and foundational version of the CV lip reading problem, using basic face-forward video frames of individuals speaking single English words as the data source. While this approach reduces data complexity, it may oversimplify the nuances of lip movements, potentially affecting accuracy [7]. This research presents a lip-reading framework to improve the identification rate in low-quality videos. In this study, a several Pose (MP) collection of low-quality movies featuring multiple extreme stances is built. The proposed framework improves the input video by dividing it into frames and applying the CLAHE approach on each frame. The proposed approach surpassed the state-of-the-art methods, obtaining 90% accuracy in phrase prediction on a testing dataset consisting of 100 silent, low-quality movies with varying locations. The method's reliance on Contrast Limited Adaptive Histogram Equalization (CLAHE) may not generalize well to videos with varying noise levels and artifacts [8]. In this paper, the latest advances in deep learning based VSR research is discussed in detail, with an emphasis on task-specific complexities, data issues, and their related solutions. Developments along these lines will hasten the transition of silent speech interface theory to reality. The paper emphasizes the need for large, diverse datasets but does not delve into the ethical considerations and potential biases associated with dataset collection and usage [9]. The author of this research uses a VGG Net that has been pre-trained on celebrity faces from IMDB and Google Images. The VGG Net is used in conjunction with LSTMs to extract temporal information, and it is trained on images that have been concatenated from numerous frames in each sequence. For various reasons, the concatenated image model utilizing nearest-neighbor interpolation performed better than the LSTM models, yielding a validation accuracy of 76%. While the approach shows promise, the model's performance may be influenced by the quality and diversity of the training data [10]. The paper talks about the lip reading algorithm LIPNET. It focuses on the use of deep neural networks to transcribe the sequences of spoken words. Despite its innovative approach, the model's complexity may hinder real-time applications and scalability [11]. The paper talks about lip reading algorithm based of Ghost Net. Ghost net is a lightweight convolutional NN architecture designed for efficient object detection. Ghost Net is particularly known for its emphasis on computational efficiency. While the architecture aims for efficiency, its performance may be compromised when dealing with complex lip movements and diverse speakers [12]. The paper gives a review on LSTM. It was introduced to address the vanishing gradient problem, a challenge faced by traditional RNNs when learning patterns that are separated by many time steps. It consists of Memory cells, gates and cell state. The review highlights the advantages of LSTMs but does not address their limitations, such as the vanishing gradient problem and computational demands [13]. This paper proposed an Indonesian lip-reading system utilizing a Long-Term Recurrent Convolutional Network (LRCN) and MediaPipe Face Mesh for precise lip landmark detection. Their model achieved high accuracy of 95.42% on word-level and 95.63% on phrase-level tests, surpassing Conv-LSTM baselines. However, evaluations were conducted on controlled datasets, which may limit generalizability to real-world conditions with varied lighting and occlusions. Additionally, reliance on MediaPipe Face Mesh could affect performance if facial landmark detection is compromised [14]. This research proposed

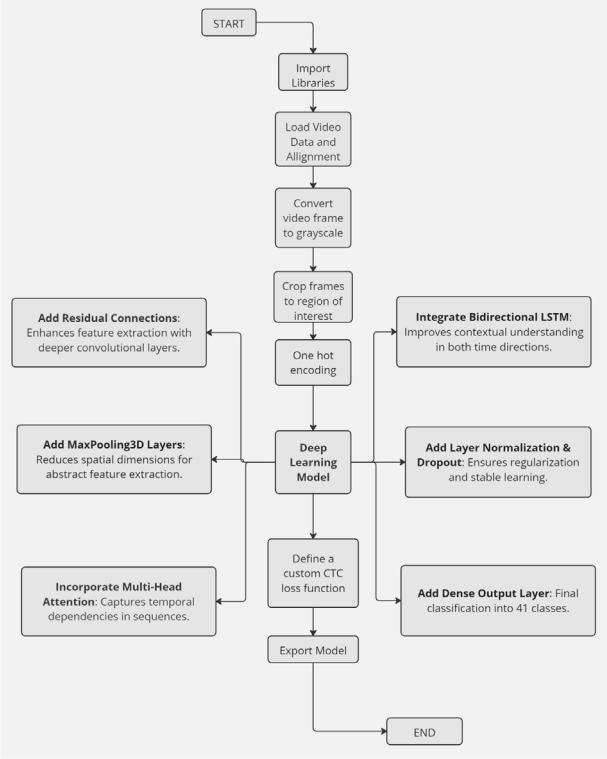

Fig. 1. Flowchart of the proposed 3D residual convolutional and multi-head attention approach.

a lip-reading model utilizing a twelve-layer convolutional neural network with batch normalization, processing unified images of isolated lip regions from video sequences to predict complete words. While the model addresses challenges like speech pace and lighting variations, its reliance on static image amalgamation may limit temporal dynamics capture, and the absence of detailed performance metrics hinders comprehensive evaluation [15]. This paper introduced an innovative lip-reading classification approach that integrates Optimized Quaternion Meixner Moments (OQMMs) with Convolutional Neural Networks (CNNs). By employing the Grey Wolf Optimization (GWO) algorithm to fine-tune the local parameters of Meixner polynomials, the method effectively extracts discriminative features from lip images. This technique addresses challenges in feature extraction and model capability for visual speech recognition. Evaluated on datasets such as AVLetters, GRID, AVDigits, and LRW, the OQMMs-CNN approach demonstrated superior classification accuracy and reduced computational complexity compared to existing deep learning models [16].

3. Proposed methodology

3.1. Dataset description

The work employed, the GRID Audio-Visual Sentence Corpus [17], a comprehensive dataset widely utilized in visual speech recognition research. The corpus comprises high-quality audio and video recordings of 34 speakers (18 male and 16 female), each articulating 1,000 sentences, culminating in a total of 34,000 utterances. Each sentence adheres to a fixed six-word structure: command, color, preposition, letter, digit, and adverb (e.g., "place red at C zero again"), facilitating controlled linguistic variability. The video recordings are provided in MPEG format with a resolution of 720 × 576 pixels at 25 frames per second, while the audio is sampled at 25 kHz in WAV format. Accompanying annotations include word-level time alignments for each utterance. For preprocessing, we extracted the mouth region using a 68-point facial landmark detector, followed by grayscale conversion and pixel value normalization to standardize the input data.

3.2. Approach overview

Fig. 1 shows a Flow chart of the proposed 3D Residual Convolutional and Multi-Head Attention approach. The methodology includes several stages, such as data preprocessing, design of the neural network architecture, training procedure, and model evaluation. Each one of these is carefully designed so that a good and accurate lip-reading system can be produced. The initial step in proposed methodology is very comprehensive data preprocessing, which needs to be done on the raw video data so it can

be put into effective training and evaluation. The implementation begins with importing essential libraries and frameworks such as TensorFlow and Keras. These libraries offer the necessary tools for constructing and training neural networks, providing a robust foundation for the lip-reading model. Loading a dataset comprising video sequences where each sequence corresponds to spoken words. It is important that the data aligns correctly to ensure that the video frame properly maps to the corresponding temporal sequence of the words spoken. For this reason, proper alignment becomes essential in the data to sustain temporal consistency. To simplify and reduce the computation complexity of the data, it is converted from RGB to grayscale video frames. This conversion helps in retaining essential features while reducing the dimensionality of the data, thus easing the processing load and focusing on the critical aspects of the lip movements. Each frame is cropped to concentrate on the region around the lips. This step is essential to eliminate irrelevant background information and center the model's attention on the lip movements, which are pivotal for accurate speech recognition. The spoken words in the dataset are converted into one-hot encoded vectors. One-hot encoding converts categorical labels into a binary matrix, making it easier for the neural network to classify the data by transforming the labels into a suitable format for model training. The neural network architecture designed for this lip-reading model integrates Convolutional 3D (Conv3D) layers, Bi-LSTMlayers, Multi Head Attention Mechanism and Dense layers. This architecture is designed to properly capture both the spatial and temporal aspects of lip movement. Conv3D layers extract the spatial feature of video frames. These types of layers will conduct 3D convolution on data, suitable for video information as they analyze not only two spatial dimensions in each frame of height and width but also in the depth that involves time steps. The initial Conv3D layer applies convolutional filters to the input video frames, thereafter max pooling and a ReLU activation function. This layer reduces spatial dimensions and extracts fundamental features. Additional Conv3D layers further refine the extracted features, deepening the feature maps while progressively reducing spatial dimensions through further pooling layers. Residual connections are incorporated to enhance feature learning and mitigate vanishing gradient issues. After the final Conv3D layer, a TimeDistributed layer flattens the output across the temporal dimension, allowing each time step to be processed independently. This step retains the temporal structure of the data and prepares it for the subsequent LSTM layers. The first Bi-LSTM Layer processes the time-distributed features, capturing long-range dependencies within the lip movement sequences. A third Bi-LSTM layer further refines the temporal features, allowing the model to better detect complex temporal patterns. Multi-Head Attention Mechanism has been employed to attend to different parts of the input sequence to allow the model to better handle variable speech patterns and contextual information in the input. The last Dense layer contains a SoftMax activation function to output the probability for each possible spoken word by the model. The model uses a custom connectionist temporal classification loss function. This type of a loss function is specifically effective where the alignment between the input and output sequences is unknown in sequence-to-sequence problems. It mathematically presents the loss function as is demonstrated in Eq. (1).

$$L_{\text{CTC}} = -\log\left(p(y \mid x)\right) \tag{1}$$

In this formulation, $p(y \mid x)$ denotes the probability of the correct label sequence y given the input sequence x. The forward–backward algorithm is used to calculate the probability, as it is very efficient in handling sequence alignment, which is crucial for accurate training of the model. The training process uses the Adam optimizer with an adaptive learning rate and excellent performance. Upon completing the training phase, the model is evaluated on a separate test dataset to assess its performance. Key metrics such as accuracy and loss are computed to gauge the model's effectiveness. Once evaluated, the trained model is exported for deployment.

3.3. Training configuration and reproducibility

The training of the proposed 3D-RMA lip-reading model was conducted using the TensorFlow 2.x and Keras deep learning frameworks. The model was optimized using the Adam optimizer with an initial learning rate set to 1e−4, which was adaptively adjusted during training to ensure stable convergence. A batch size of 8 was employed to balance memory efficiency and convergence speed. The model was trained for up to 60 epochs, with early stopping implemented based on validation loss to prevent overfitting. To enhance generalization, dropout layers with a dropout rate of 0.3 were integrated within the LSTM layers. The Connectionist Temporal Classification (CTC) loss function was used, as it is well-suited for sequence prediction tasks where the alignment between input frames and output sequences is unknown. Evaluation was performed using character-level accuracy, computed by comparing decoded predictions to the ground truth labels. For reproducibility, all experiments were conducted in a controlled software environment, and training scripts, model checkpoints, and configuration files were archived.

4. Model architecture

Table 1 provides a detailed summary of the Proposed model architecture, including the various layers, their input shapes, parameter counts, and output shapes.

4.1. Input layer

The input layer is designed to accommodate video data, specifically grayscale sequences of frames. It accepts tensors, as defined in Eq. (2), the input to the model is represented by a 5D tensor $\mathbf{X}$. with the shape $(\text{None}, 75, 46, 140, 1)$, where None denotes the batch size, 75 represents the number of video frames, and each frame is of size 46×140 pixels with a single grayscale channel. This structure is crucial for capturing both temporal and spatial information necessary for effective lip reading.

$$\mathbf{X} \in \mathbb{R}^{N \times T \times H \times W \times C} \tag{2}$$

Table 1
Model summary of proposed 3d residual convolutional and multi-head attention approach.

Layer (type)	Input shape	Param #	Output shape
input_layer_1 (InputLayer)	(None, 75, 46, 140, 1)	0	–
conv3d_9 (Conv3D)	(None, 75, 46, 140, 128)	3584	input_layer_1[0][0]
conv3d_10 (Conv3D)	(None, 75, 46, 140, 128)	442,496	conv3d_9[0][0]
conv3d_11 (Conv3D)	(None, 75, 46, 140, 128)	442,496	conv3d_10[0][0]
add_4 (Add)	(None, 75, 46, 140, 128)	0	conv3d_9[0][0], conv3d_11[0][0]
activation_3 (Activation)	(None, 75, 46, 140, 128)	0	add_4[0][0]
max_pooling3d_3 (MaxPooling3D)	(None, 75, 23, 70, 128)	0	activation_3[0][0]
conv3d_12 (Conv3D)	(None, 75, 23, 70, 256)	884,992	max_pooling3d_3[0][0]
conv3d_13 (Conv3D)	(None, 75, 23, 70, 256)	1,769,728	conv3d_12[0][0]
conv3d_14 (Conv3D)	(None, 75, 23, 70, 256)	1,769,728	conv3d_13[0][0]
add_5 (Add)	(None, 75, 23, 70, 256)	0	conv3d_12[0][0], conv3d_14[0][0]
activation_4 (Activation)	(None, 75, 23, 70, 256)	0	add_5[0][0]
max_pooling3d_4 (MaxPooling3D)	(None, 75, 11, 35, 256)	0	activation_4[0][0]
conv3d_15 (Conv3D)	(None, 75, 11, 35, 75)	518,475	max_pooling3d_4[0][0]
conv3d_16 (Conv3D)	(None, 75, 11, 35, 75)	151,950	conv3d_15[0][0]
conv3d_17 (Conv3D)	(None, 75, 11, 35, 75)	151,950	conv3d_16[0][0]
add_6 (Add)	(None, 75, 11, 35, 75)	0	conv3d_15[0][0], conv3d_17[0][0]
activation_5 (Activation)	(None, 75, 11, 35, 75)	0	add_6[0][0]
max_pooling3d_5 (MaxPooling3D)	(None, 75, 5, 17, 75)	0	activation_5[0][0]
time_distributed_1 (TimeDistributed)	(None, 75, 6375)	0	max_pooling3d_5[0][0]
multi_head_attention_1 (MultiHeadAttention)	(None, 75, 6375)	7,657,275	time_distributed_1[0][0]
add_7 (Add)	(None, 75, 6375)	0	time_distributed_1[0][0], multi_head_attention_1[0][0]
bidirectional_2 (Bidirectional)	(None, 75, 256)	6,660,096	add_7[0][0]
layer_normalization_2 (LayerNormalization)	(None, 75, 256)	512	bidirectional_2[0][0]
dropout_4 (Dropout)	(None, 75, 256)	0	layer_normalization_2[0][0]
bidirectional_3 (Bidirectional)	(None, 75, 256)	394,240	dropout_4[0][0]
layer_normalization_3 (LayerNormalization)	(None, 75, 256)	512	bidirectional_3[0][0]
dropout_5 (Dropout)	(None, 75, 256)	0	layer_normalization_3[0][0]
dense_1 (Dense)	(None, 75, 41)	10,537	dropout_5[0][0]

4.2. Conv3D layer

The Conv3D Layer plays a crucial role in feature extraction from video frames. It utilizes 128 filters with a kernel size of $3 \times 3 \times 3$ to perform 3D convolution, extracting both spatial and temporal features from the input frames. The ReLU activation function introduces non-linearity to the model, allowing it to pick up complex patterns. This layer processes the input tensor while maintaining the temporal and spatial dimensions but increasing the depth of feature maps, resulting in an output tensor with the shape $(\text{None}, 75, 46, 140, 128)$. The convolution operation is mathematically described by the following Eq. (3):

$$Z_{i,j,k,f} = \sum_{l=0}^{K_T-1} \sum_{m=0}^{K_H-1} \sum_{n=0}^{K_W-1} \sum_{c=0}^{C-1} X_{i+l,j+m,k+n,c} \cdot W_{l,m,n,c,f} + b_f \tag{3}$$

where $Z_{i,j,k,f}$ denotes the output feature map at position (i, j, k) for filter f, X is the input tensor, $W_{l,m,n,c,f}$ represents the learnable weights of the filter, b_f is the bias term for filter f, and K_T, K_H, and K_W are the kernel sizes expressed in terms of time, height, and breadth, respectively. The activation function of ReLU is utilized to the convolution output, which is given by Eq. (4):

$$A_{i,j,k,f} = \text{ReLU}(Z_{i,j,k,f}) = \max(0, Z_{i,j,k,f}) \tag{4}$$

where $A_{i,j,k,f}$ is the activated feature map at position (i, j, k) for filter f, and ReLU represents the function of the Rectified Linear Unit, which sets all negative values in $Z_{i,j,k,f}$ to zero.

4.3. Residual Block 1

The Residual Block 1 is structured with two Conv3D layers designed to enhance feature learning and mitigate the vanishing gradient problem. The first Conv3D layer applies 128 filters with a $3 \times 3 \times 3$ kernel and uses the ReLU activation function to process the input, as described by Eq. (5):

$$Z^{(1)}_{i,j,k,f} = \sum_{l,m,n,c} X_{i+l,j+m,k+n,c} \cdot W^{(1)}_{l,m,n,c,f} + b^{(1)}_f \tag{5}$$

where $Z^{(1)}_{i,j,k,f}$ represents the output feature map at position (i, j, k) for filter f, X is the input tensor, $W^{(1)}_{l,m,n,c,f}$ denotes the learnable weights, and $b^{(1)}_f$ is the bias term. The ReLU activation applied to this layer is given by (6):

$$A^{(1)}_{i,j,k,f} = \text{ReLU}(Z^{(1)}_{i,j,k,f}) \tag{6}$$

The second Conv3D layer also employs 128 filters with a $3 \times 3 \times 3$ kernel but without an activation function, aiming to learn residual features, computed as (7):

$$Z^{(2)}_{i,j,k,f} = \sum_{l,m,n,c} A^{(1)}_{i+l,j+m,k+n,c} \cdot W^{(2)}_{l,m,n,c,f} + b^{(2)}_f \tag{7}$$

where $Z^{(2)}_{i,j,k,f}$ is the output feature map from the second Conv3D layer, and $W^{(2)}_{l,m,n,c,f}$ represents the learnable weights for this layer with bias $b^{(2)}_f$. As illustrated in Eq. (8) a skip connection then adds the original input of the block to the output of the second Conv3D layer, producing:

$$Z^{(\text{res})}_{i,j,k,f} = X_{i,j,k,f} + Z^{(2)}_{i,j,k,f} \tag{8}$$

The final output is passed through a ReLU activation function as represented in (9);

$$A^{(\text{res})}_{i,j,k,f} = \text{ReLU}(Z^{(\text{res})}_{i,j,k,f}) \tag{9}$$

This architecture allows for easier gradient flow through the network, enabling effective training of deeper networks by learning residual features and mitigating the vanishing gradient problem.

4.4. MaxPooling3D layer

The feature maps are downsampled using the MaxPooling3D layer, which keeps the most significant characteristics while decreasing their spatial dimensions. With a pool size of $1 \times 2 \times 2$, this layer performs pooling operations across the height and width dimensions, resulting in an output shape of $(None, 75, 23, 70, 128)$. The purpose of this downsampling is to lower computational complexity and mitigate overfitting by abstracting the features. The maximum value within each pooling window is chosen by the max-pooling procedure, which can be mathematically described as shown in Eq. (10).

$$P_{i,j,k,f} = \max_{m,n} \left(A^{(\text{res})}_{i,2j+m,2k+n,f} \right) \tag{10}$$

where $P_{i,j,k,f}$ represents the pooled output value at position (i, j, k) for filter f, and m and n are the indices within the pooling window. This operation ensures that only the most prominent features are retained, aiding in efficient feature extraction and model training.

4.5. Conv3D layer

The following Conv3D layer will raise the depth of the feature maps using 256 filters of size $3 \times 3 \times 3$ and ReLU activation. The job of this particular layer is to extract deep and complex features from the video frames that are downsampled, and thus, the resulting output tensor would be of size

$$(None, 75, 23, 70, 256)$$

Increasing the filters allow the model to perceive more complex interrelations and patterns in the data.

4.6. Residual Block 2

Residual Block 2 is the same structure as that of Residual Block 1 but uses 256 filters. The block comprises two Conv3D layers; the first applies convolution with ReLU activation followed by the second convolution without using activation. A skip connection is made to add the original input into the output of the second convolution layer, followed by ReLU activation. This design improves the capacity of the model to learn complex features while keeping the gradient flow effective during training. The resulting output tensor is of shape $(None, 75, 23, 70, 256)$.

4.7. MaxPooling3D layer

An additional MaxPooling3D layer is included with a size of $1 \times 2 \times 2$, that reduces the spatial dimensions of feature maps to $(None, 75, 11, 35, 256)$ The latter pooling continues the process of data downsampling in the manner which is capable of regulating the learned feature complexity. It increases model robustness as well.

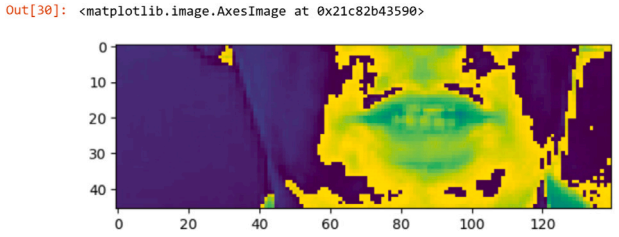

Fig. 2. Heatmap output form a model.

4.8. Conv3D layer

The next Conv3D layer uses 75 filters, a $3 \times 3 \times 3$ kernel, and ReLU activation. This reduces the number of feature maps to 75, thus shaping the output to the dimensions required by the attention mechanism. The output tensor is of shape $(None, 75, 11, 35, 75)$. This stage prepares the data for the next attention mechanism by focusing on a reduced number of feature maps, just like the previous Conv3D layer but with 75 filters.

4.9. Residual Block 3

Residual Block 3 is similar to the residual blocks above it, except that it uses 75 filters. It consists of two Conv3D layers, with a skip connection, reflecting the preceding Residual blocks. This structure continues to learn complex features. This results in effective training through skip connections that prevent the vanishing gradient problem. The output shape of this block is $(None, 75, 11, 35, 75)$.

4.10. MaxPooling3D layer

The Final layer in the MaxPooling3D, with a pool size of $1 \times 2 \times 2$, samples further down on the features. This would mean additional reduction of output shape $(None, 75, 5, 17, 75)$. It prepares data for flattening and subsequent dense layers, thus streamlining the feature.cartographic representations for subsequent analysis within the network.

4.11. TimeDistributed Layer (Flatten layer)

The TimeDistributed Layer, also known as the Flatten Layer, reshapes each feature map of size $5 \times 17 \times 75$ into a flattened vector of size 6375. This operation transforms the tensor from a shape of $(\text{None}, 75, 5, 17, 75)$ into a flattened output with shape $(\text{None}, 75, 6375)$. The reshaping process preserves the temporal dimension while collapsing the spatial dimensions into a single vector, which is crucial for subsequent attention mechanisms and fully connected layers.

4.12. MultiHeadAttention Layer

The MultiHeadAttention Layer is a critical component for capturing temporal dependencies in the data, utilizing 4 attention heads and a key dimension of 75. This layer performs attention by computing weighted representations of different parts of the input sequence.

4.13. Dense layer

The final Dense Layer in the model consists of a single unit with a Sigmoid activation function, producing a binary classification output. The probability of the positive class is represented by a value in the range of 0 to 1, which is the output of this layer.

5. Experimental results

5.1. Results and discussion

Fig. 2 represents an AxesImage object created by Matplotlib. which is generated by a machine learning model, during evaluation. It shows different intensity levels represented by colors ranging from purple (low intensity) to yellow (high intensity). This is used for visualizing the output of convolutional layers in a neural network, attention maps and feature activations

Fig. 3 demonstrates the model output. The code utilizes TensorFlow functions to decode numerical representations into readable text. Specifically, the line

```
tf.strings.reduce_join([num_to_char(word) for word in sentence])
```

```
In [68]: print('~'*100, 'PREDICTIONS')
         [tf.strings.reduce_join([num_to_char(word) for word in sentence]) for sentence in decoded]

Out[68]: [<tf.Tensor: shape=(), dtype=string, numpy=b'bin red with t five soon'>]
```

Fig. 3. Tensorflow text decoding output.

```
1/1 [==============================] - 3s 3s/step
Model Accuracy: 96.00%
```

Fig. 4. Model evaluation.

Table 2
Accuracy of deep learning models for lip reading.

Model	Accuracy
3D-RMA (Proposed)	**96.00%**
LipNet	88.6%
3D-2D-CNN-BLSTM with Word-CTC	91.4%
GMM-HMM with LDA-DCT	86.4%
LSTM-based model	79.6%

for sentence in [sample[1]] illustrates the process of concatenating characters to form words for a given sentence. The resulting printed output, bin red with t five soon', represents a decoded string from the TensorFlow tensor, showcasing a sample of the text generated or predicted by the model

Fig. 4 illustrates the outcome of the model training process within a machine learning framework. The model achieved an accuracy of 96.00%, indicating robust performance and effective learning. To compute this accuracy, the total number of correct predictions is compared to the total number of predictions made. Each dataset sample is processed to obtain model predictions, which are decoded into readable text. These predictions are then compared to the true labels, and a counter for correct predictions is incremented for each match. Accuracy is calculated as the ratio of correct predictions to total predictions, providing a measure of the model's effectiveness in accurately identifying the true labels.

In the proposed 3D Residual Multi-Head Attention-based Lip Reading (3D-RMA) model, sentence-level accuracy is adopted as the primary evaluation metric to assess transcriptional correctness under weakly aligned spatiotemporal constraints. The model employs the Connectionist Temporal Classification (CTC) loss function, which generates a posterior probability distribution over the label set for each input frame. During inference, the most probable output sequence $\hat{y}$ is derived via a decoding scheme that marginalizes over all valid alignments π using the forward–backward algorithm [18]. The decoded sequence is obtained as (11),

$$\hat{y} = \arg \max_{\pi \in \mathcal{B}^{-1}(y)} p(\pi \mid x) \tag{11}$$

where $\mathcal{B}^{-1}(y)$ denotes the set of all alignment paths that map to a valid output sequence y under the CTC collapsing function $\mathcal{B}$, and $p(\pi \mid x)$ is the probability of alignment path π given the input sequence x.

To quantitatively evaluate the model, accuracy A is computed as in (12),

$$A = \frac{1}{|\mathcal{Y}|} \sum_{i=1}^{|\mathcal{Y}|} \mathbb{1}(\hat{y}_i = y_i) \tag{12}$$

where $\mathcal{Y}$ is the set of all ground truth sequences, $\hat{y}_i$ is the decoded output sequence for the ith sample, y_i is the corresponding reference label, and $\mathbb{1}(\cdot)$ is the Kronecker delta function that returns 1 if the sequences match exactly and 0 otherwise. This exact-match criterion enforces strict correctness over entire output sequences. The model achieves a peak accuracy of 96.00%, validating its efficacy in modeling long-range temporal dependencies through Conv3D operations, residual connections, BiLSTM layers, and multi-head attention mechanisms.

5.1.1. Comparision with state-of-the-art-methods

Table 2 presents a comparative analysis of accuracy achieved by various deep learning models for lip reading on the GRID corpus Dataset [17]. The proposed model, 3D-RMA, achieves the highest accuracy of 96.00%, demonstrating the effectiveness of its advanced 3D convolutional architecture combined with multi-head attention mechanisms. This significantly outperforms existing approaches such as the 3D-2D-CNN-BLSTM with Word-CTC, which attains 91.4% [19], and LipNet [10], which achieves 88.6%. Traditional models like the GMM-HMM with LDA-DCT and LSTM-based model demonstrate lower accuracies of 86.4% [20] and 79.6% [21], respectively. These results underscore the advantages of leveraging both spatial and temporal features through deep residual learning and attention-based mechanisms in the proposed 3D-RMA model.

The proposed 3D-RMALR model demonstrates strong potential for real-world applications, particularly in enhancing communication accessibility for individuals with hearing or speech impairments. With a accuracy of 96.00%, it can be effectively integrated into assistive technologies such as silent speech interfaces, real-time transcription tools, and security systems that rely on visual speech

recognition. Compared to existing models, the significant improvement in accuracy (up to 10.7% higher than LSTM with CTC) suggests a measurable advancement in performance. This makes the model highly suitable for deployment in noisy environments where traditional audio-based speech recognition systems may fail.

6. Conclusion

This study introduces the 3D Residual Multi-Head Attention (3D-RMA) model for lip reading, a novel architecture that seamlessly integrates spatial and temporal feature extraction through Conv3D layers, residual connections, Bidirectional LSTMs, and Multi-Head Attention mechanisms. The use of Connectionist Temporal Classification (CTC) loss enables effective sequence prediction without requiring frame-level alignment, making the model well-suited for real-world lip-reading tasks. The model achieved a sentence-level recognition accuracy of 96.00% on an unseen test set, demonstrating strong generalization and robustness. Although accuracy was the principal metric used in this study, the model architecture supports further assessment through more granular metrics such as Word Error Rate (WER). These will be explored in future work to gain deeper insight into the transcription quality and to address possible insertion, deletion, and substitution errors in decoded sequences. The integrated use of residual 3D convolutions and attention-enhanced sequence modeling distinguishes the proposed method from prior work that typically isolates spatial and temporal processing. This unified design improves the model's ability to capture complex lip dynamics and phonetic variation over time. This architecture has practical relevance for real-time visual speech recognition in assistive technologies, particularly in environments where audio-based methods are ineffective. Its computational efficiency makes it suitable for deployment on edge devices, enabling broader use in applications such as silent speech interfaces and communication aids for the hearing-impaired. In conclusion, the 3D-RMA model delivers a technically robust and scalable approach to lip reading, with demonstrated performance and extensibility for future advancements in visual speech processing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the "University of Debrecen Program for Scientific Publication".

Data availability

Data will be made available on request.

References

[1] Pu G, Wang H. Review on research progress of machine lip reading. Vis Comput 2023;39(7):3041–57. http://dx.doi.org/10.1007/s00371-022-02511-4.
[2] Pujari S, Sneha S, Vinusha R, Bhuvaneshwari P, Yashaswini C. A survey on deep learning based lip-reading techniques. In: Proc. 2021 third international conference on intelligent communication technologies and virtual mobile networks. ICICV, 2021, p. 1286–93. http://dx.doi.org/10.1109/ICICV50876.2021.9388569.
[3] Nambeesan AS, Payyappilly C, J. C E, J. P J, Alex MS. LIP reading using facial feature extraction and deep learning. Int J Innov Sci Res Technol 2021;6(7):92.
[4] Lu Y, Li H. Automatic lip-reading system based on deep convolutional neural network and attention-based long short-term memory. Appl Sci 2019;9(8):1599. http://dx.doi.org/10.3390/app9081599.
[5] Shrestha K. Lip reading using neural network and deep learning. Department of Computer Science, Earlham College, Richmond, in, USA, (N/A). Email: kshres15@earlham.edu.
[6] Hao M, Mamut M, Yadikar N, Aysa A, Ubul K. A survey of research on lipreading technology. IEEE Access 2020;8:204518–44. http://dx.doi.org/10.1109/ACCESS.2020.3036865.
[7] Peyton S, Rosenfeld R. Deep learning approaches for visual speech recognition. IEEE Trans Pattern Anal Mach Intell 2020;42(5):1046–60. http://dx.doi.org/10.1109/TPAMI.2019.2917337.
[8] Akhter N, Ali M, Hussain L, Shah M, Mahmood T, Ali A, Al-Fuqaha A. Diverse pose lip-reading framework. Appl Sci 2022;12(19):9532. http://dx.doi.org/10.3390/app12199532.
[9] Oghbaie M, Sabaghi A, Hashemifard K, Akbari M. Advances and challenges in deep lip reading. 2021, http://dx.doi.org/10.48550/arXiv.2110.07879, arXiv preprint arXiv:2110.07879.
[10] Assael YM, Shillingford B, Whiteson S, de Freitas N. LipNet: End-to-end sentence-level lipreading. 2016, http://dx.doi.org/10.48550/arXiv.1611.01599, arXiv preprint arXiv:1611.01599.
[11] Zhang G, Lu Y. Research on a lip reading algorithm based on efficient-GhostNet. Electronics 2023;12(5):1151. http://dx.doi.org/10.3390/electronics12051151.
[12] Van Houdt G, Mosquera C, Nápoles G. A review on the long short-term memory model. Artif Intell Rev 2020;53(8):5929–55. http://dx.doi.org/10.1007/s10462-020-09838-1.
[13] Lip YK, Michael YRW, Chan WC. End-to-end sentence-level lipreading with transformer networks. In: Proc. 2020 IEEE int. conf. acoust. speech signal process.. ICASSP, Barcelona, Spain; 2020, p. 6039–43. http://dx.doi.org/10.1109/ICASSP40776.2020.9053296.
[14] Aripin, Setiawan A. Indonesian lip-reading detection and recognition based on lip shape using face mesh and long-term recurrent convolutional network. In: Applied computational intelligence and soft computing. Hindawi, London, UK; 2024, p. 1–13. http://dx.doi.org/10.1155/2024/6479124.

[15] Isaac E, Tomy SK, Joy B, Thomas EA, Ahmed HN, Raybin J. Lip reading using deep learning and convolutional neural network. In: AIP conference proceedings. Vol. 3134, AIP Publishing; 2024, 020011. http://dx.doi.org/10.1063/5.0230447.
[16] El Ogri O, EL-Mekkaoui J, Benslimane M, Hjouji A. Automatic lip-reading classification using deep learning approaches and optimized quaternion meixner moments by GWO algorithm. 304, Elsevier; 2024, 112430. http://dx.doi.org/10.1016/j.knosys.2024.112430,
[17] Cooke M, Barker J, Cunningham S, Shao X. An audio-visual corpus for speech perception and automatic speech recognition. 120, (5):Acoustical Society of America; 2006, p. 2421–4. http://dx.doi.org/10.1121/1.2229005,
[18] Graves A, Fernández S, Gomez F, Schmidhuber J. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd international conference on machine learning. 2006, p. 369–76.
[19] Margam DK, Aralikatti R, Sharma T, Thanda A, Pujitha AK, Roy S, Venkatesan SM. LipReading with 3D-2D-CNN BLSTM-HMM and word-CTC models. 2019, https://doi.org/10.48550/arXiv.1906.12170, arXiv preprint arXiv:1906.12170, URL https://arxiv.org/abs/1906.12170, submitted for publication to Interspeech 2019.
[20] Kaur M, Rastogi D, Sharma A, Dahiya A, Nagrath P. Crime investigation using lip reading. In: SCCTT-2024: international symposium on smart cities, challenges, technologies and trends. Delhi, India; 2024, URL http://ceur-ws.org, Corresponding author: mansimran2703@gmail.com.
[21] Wand M, Koutník J, Schmidhuber J. Lipreading with long short-term memory. 2016, arXiv preprint arXiv:1601.08188. URL https://doi.org/10.48550/arXiv.1601.08188 in press at ICASSP 2016.