Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 8, No. 6, December, 2024

SlowFast-TCN: A Deep Learning Approach for Visual Speech Recognition

Nicole Yah Yie Ha [1], Lee-Yeng Ong [1*], Meng-Chew Leow [1]

[1] *Faculty of Information Science and Technology (FIST), Multimedia University, Jalan Ayer Keroh Lama, Melaka 75450, Malaysia.*

Abstract

Visual Speech Recognition (VSR), commonly referred to as automated lip-reading, is an emerging technology that interprets speech by visually analyzing lip movements. A challenge in VSR where visually distinct words produce similar lip movements is known as homopheme problem. Visemes are the basic visual units of speech that are produced by the lip movements and positions. Furthermore, visemes are typically having shorter durations than words. Consequently, there is less temporal information for distinguishing between different viseme classes, leading to increased visual ambiguity during classification. To address this challenge, viseme classification must not only extract lip image spatial features, but also to handle visemes of varying durations and temporal features. Therefore, this study proposed a new deep learning approach SlowFast-TCN. SlowFast network is used as the frontend architecture to extract the spatio-temporal features of the slow and fast pathways. Temporal Convolutional Network (TCN) is used as the backend architecture to learn the features from the frontend to perform the classification. A comparative ablation analysis to dissect each component of the proposed SlowFast-TCN is performed to evaluate the impact of each component. This study utilizes a benchmark dataset, Lip Reading in Wild (LRW), that focuses on English language. Two subsets of the LRW dataset, comprising of homopheme words and unique words, represent the homophemic and non-homophemic dataset, respectively. The proposed approach is evaluated on varying lighting conditions to assess its performance in real-world scenarios. It was found that illumination can significantly affect the visual data. Key performance metrics, such as accuracy and loss are used to evaluate the effectiveness of the proposed approach. The proposed approach outperforms traditional baseline models in accuracy while maintaining competitive execution time. Its dual-pathway architecture effectively captures both long-term dependencies and short-term motions, leading to better performance in both homophemic and non-homophemic datasets. However, it is less robust when dealing with non-ideal lighting scenarios, indicating the need for further enhancements to handle diverse lighting scenarios.

Keywords:

Visual Speech Recognition;
Temporal Convolutional Network;
Lip Reading in Wild;
SlowFast Network;
Homophemes.

Article History:

Received:	20	July	2024
Revised:	19	November	2024
Accepted:	24	November	2024
Published:	01	December	2024

1- Introduction

Visual speech recognition (VSR), commonly known as lip-reading, is the process of recognizing the content of speech to text-based solely on visual cues (lip movements) without relying on audio. It has gained significant attention in recent years due to its potential applications in silent communication, security and surveillance systems, and accessibility technologies [1]. However, despite significant advancements, VSR systems still face critical challenges, particularly in accurately distinguishing between visually similar words, known as the homopheme problem [2-4], which remains one of the major challenges in this domain. Homophemes refer to different words that appear similar in the lip movements even though they have different pronunciations [5, 6], leading to recognition failure due to visual ambiguity.

* **CONTACT**: lyong@mmu.edu.my

DOI: http://dx.doi.org/10.28991/ESJ-2024-08-06-024

Visemes are the basic visual units of speech, representing groups of phonemes (distinct units of sound) that look similar when pronounced. Each viseme corresponds to a specific lip movement and position. In terms of spatial features, visemes are characterized by the shape and motion of the lips. They are crucial for distinguishing between different sounds of speech visually. In terms of temporal features, a viseme typically has a shorter duration compared to the whole word. This temporal brevity means that there is less information available over time to distinguish between different visemes. Consequently, when homophemes produce identical visemes, it creates visual ambiguity. This is making it difficult for VSR systems to accurately recognize and differentiate between words based solely on visual input. Existing approaches often struggle with this challenge, suffering high misrecognition rates [7, 8]. This is partly because VSR systems typically rely on deep learning models that focus on either spatial or temporal features, but not both in a balanced manner.

With the evolution of deep learning and increased datasets availability on large vocabularies, the research in this field has shifted from recognizing simple digits and alphabets to word or sentence level. From traditional approaches in decoding simple digits and alphabets to using deep learning approaches in decoding complex words and sentences, there have been many breakthroughs in lip-reading [2, 4]. Deep learning technologies such as Long-Short Term Memory (LSTM) networks [9-11], transformer [12, 13], and Temporal Convolutional Networks (TCN) [14] have been commonly used in the development of VSR. The use of these deep learning technologies in VSR is driven by the need to effectively capture and process temporal features of lip movements since these models excel at handling the sequential data and learning long-range dependencies, matching the nature of speech. These models excel in capturing temporal sequences but may fall short in effectively handling the complex spatial dynamics of lip movements, particularly when homophemes are involved.

Like other video processing tasks, the VSR system also faces challenges related to varying lighting conditions in which it needs high-quality lighting conditions [15]. Changes in lighting can obscure and distort the visual features that are important for accurate lip-reading [16, 17], such as the shape and movement of the lips. This issue is particularly critical when dealing with homophemes, where minor lighting variations can further exacerbate the visual ambiguity [18].

This study focuses on single-modal lip-reading, utilizing only visual data and excluding multi-modal approaches that involve both visual and audio data. The primary aim is to address the homopheme problem, while challenges related to speaker independence and pose variations are not within the scope. The study is focused on the English language, specifically developing an approach for recognizing spoken words from visual lip movements. The experiments are conducted on recorded video data rather than live or real-time data.

To address the gap, in this study, a new deep learning approach is proposed by integrating the SlowFast network [19] with TCN [20] to enhance the performance of VSR. Specifically, the proposed approach attempts to solve the problems of recognizing homophemes and capturing subtle spatio-temporal dynamics in visual speech. The SlowFast Network is known for its ability to capture rich spatio-temporal dynamics by processing video frames at different speeds. TCNs are effective in handling sequential data, making them suitable for modeling the temporal dependencies of visual speech. By combining these two architectures, the proposed SlowFast-TCN approach improves recognition accuracy and has robustness against homophemes by leveraging the strengths of both spatio-temporal feature extraction and temporal sequence modeling. Additionally, the robustness of the proposed approach under varying illumination conditions is evaluated, further highlighting the model's applicability in different lighting environments. The contributions of this study are:

- The development of a new deep learning approach that integrates the SlowFast Network with TCN.
- The evaluation of the performance of the proposed SlowFast-TCN on varying illumination conditions to assess its robustness in different lighting environments.
- The analysis of the contribution of each component of the SlowFast-TCN model through an ablation study to provide insights into the effectiveness of the combined architecture.

The organization of this paper is as follows. Section 2 reviews existing approaches in VSR. Section 3 describes the methodology of the proposed SlowFast-TCN experiment pipeline. Section 4 presents the experimental results and their comparisons with the baseline models. Section 5 presents discussion and summarizes all the findings and limitations. Finally, Section 6 provides a concluding remark on the findings and future work of this study.

2- Related Works

2-1- Visual Speech Recognition

A VSR system normally consists of five steps of input, lip detection, feature extraction, classification, and output, as illustrated in Figure 1 [4, 21]. The input is the video that needs to be translated. It will go through the process of lip detection and the cropping of the lip region of interest (ROI). Visual features of the lips are then extracted by the frontend network and used by the backend network to classify and predict the output lexicons (such as digits, words, or sentences), depending on the purpose of the VSR system. The categorization of the lip movements helps in identifying the word produced to decode the spoken speech.

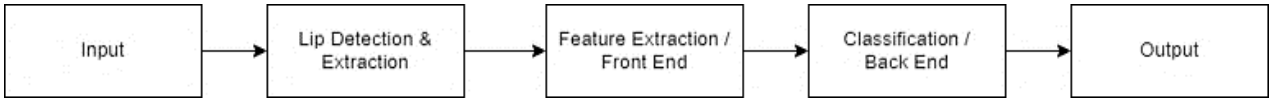

Figure 1. General pipeline in VSR system

With the emergence of deep learning, researchers discovered that it is possible to learn intricate features directly from the data. In general, deep learning approaches in VSR work end-to-end. These approaches self-learn the relevant lip features from the input videos and then classify them. The deep learning approaches can be categorized into two types, the frontend and backend networks. The commonly used approaches in frontend VSR are CNN and ResNet [12, 14, 22]. The widely used backend network models include LSTM [9-11] and TCN [14, 23, 24].

Recent research has also addressed the impact of environmental factor from lighting conditions [17, 18, 25], on the performance of VSR systems. Variations in lighting can significantly affect the visibility of lip movements [16], thereby influencing the system's ability to correctly classify speech. Despite these challenges, there has been limited exploration of how lighting specifically impacts homopheme recognition.

2-2- SlowFast Network

SlowFast [19] network is an architecture designed for video recognition tasks. It operates on a dual-pathway approach where it processes video frames at different temporal resolutions by capturing both slow and fast motion dynamics effectively. The 'Slow' pathway processes frames at a lower frame rate to capture spatial semantics, while the 'Fast' pathway captures motions at a higher frame rate, focusing more on the temporal information. This design allows the network to effectively integrate detailed spatial and temporal features. The SlowFast network has demonstrated superior performance in various video recognition tasks and outperformed traditional single-stream models in terms of accuracy and efficiency [19]. Previous studies have shown that models focusing solely on either high frame rate or low frame rate inputs often miss crucial temporal dependencies or detailed motion patterns, leading to lower accuracy rates [15, 26].

The SlowFast network has been successfully applied to a variety of video recognition tasks beyond VSR. For instance, it has achieved state-of-the-art accuracy in video action classification and detection tasks on major video recognition benchmarks like Kinetic, Charades and AVA [19, 27]. Wei et al. [27] have demonstrated significant improvements in video action recognition tasks, such as in the Kinetics-400 dataset, where integrating a cross-modality dual attention fusion module (CMDA) improved accuracy in recognizing complex actions. In tennis action classification [28], SlowFast has been used to identify and analyze various tennis shots, achieving a generalization accuracy of 74% on the THETIS dataset, demonstrating its effectiveness in handling fine-grained, fast-paced sports actions within a unified framework. Despite its success in general video recognition tasks, the SlowFast network has not been widely adopted in VSR systems. Traditional VSR systems often rely on single-stream models that process video frames at a uniform rate [21]. However, these single-stream models may not be as effective in capturing the nuanced spatio-temporal dynamics since they only learn spatial features without incorporating the corresponding temporal features [21].

A lip movement video typically contains spatial and temporal information. Since the spatial dimension typically has more information than the temporal one, it makes sense to handle them separately [29]. This aligns with the intuition behind the SlowFast network, which begins with a low frame rate pathway to capture detailed spatial information, followed by a higher frame rate pathway with reduced channel capacity that focuses on temporal dynamic features. Spatial features capture the shape, movement, and configuration of the lips during speech. These features are crucial for identifying the subtle differences in lip movements. However, relying solely on spatial information can lead to ambiguity, especially with homophemes. The lip positions for similar-sounding words might be nearly identical at certain moments, making it difficult to distinguish between them. Meanwhile, temporal features capture the sequence of lip movements in which the way of lip movements transition over time can provide additional clues that help differentiate homophemes [30]. Homophemes may have very similar spatial appearances at certain frames, but their temporal dynamics might differ slightly. Although this implementation of SlowFast seems promising, it has yet to be fully explored in a VSR system. Inspired by its success in various video recognition tasks, SlowFast network could potentially serve as the frontend network for extracting both spatial and temporal features in VSR.

2-3- Temporal Convolutional Network (TCN)

For sequence classification, TCN [20] has become popular as an alternative to RNN. Unlike RNN, TCN can process inputs in parallel rather than sequentially at each time step. This parallel processing capability is a key advantage of TCN. Another benefit of TCN is the flexibility in adjusting the receptive field size. This can be achieved by stacking more convolutional layers, using larger dilation factors, or increasing filter sizes. These adjustments allow for better control over how the model remembers information. Additionally, TCN mitigate issues like exploding or vanishing gradients, which can hinder training in traditional RNNs [20]. This is because TCN employs a backpropagation path that operates differently from the temporal direction of the sequence. Moreover, TCN typically requires less memory during training, especially when processing long input sequences [20]. This efficiency makes them particularly suitable for tasks involving extensive temporal data in lip-reading because they can handle long-range dependencies and large input sequences more effectively than RNN [14, 31]. A simple TCN architecture, as described in [20], has outperformed the RNN baseline approaches. Similarly, Martinez et al. [14] shown that multi-scale TCN (MS-TCN) outperformed the RNN approaches to word-level lip-reading.

2-4- Lip Reading in the Wild Dataset (LRW)

LRW [32] is a commonly used benchmark dataset. Each clip consists of 29 frames (1.16 seconds) of people speaking 500 words, with approximately 1,000 utterances per word. Each video is sourced from the BBC TV program, giving them the "in the wild" property, in contrast to the videos collected in a controlled laboratory setting. With the introduction of this large-scale English lip-reading dataset, the deep learning research of VSR has increased rapidly. This dataset is notable for its extensive coverage, featuring hundreds of words, thousands of examples per word, and videos from over a thousand distinct speakers. All the videos are of fixed duration, size and length. There are balanced classes, with each class having an identical number of samples. The primary goal of the LRW dataset is to evaluate lip-reading techniques for recognizing words in speaker-independent tasks [4, 33].

LRW has been widely used in the existing works to evaluate the performance of the models for VSR. Mudaliar et al. [34], who employed a 3D CNN and ResNet-18 for the frontend and a GRU model as the backend, has achieved a word accuracy of 88% on the LRW. Similarly, [22, 34], who used Bi-GRU and GRU networks with the same frontend setup, have achieved 82% and 88% accuracy, respectively. Martinez et al. [14], who applied MS-TCN with a SoftMax layer, has achieved 85.3% word-level accuracy. Building on the previous work, Ma et al. [35] used DC-TCN to replace MS-TCN and is able to reach a state-of-the-art performance of 93.4% on LRW.

2-5- Performance Metrics

The accuracy and loss are two important metrics used to evaluate the performance of a word-level VSR system. Accuracy measures how many words are correctly recognized, while the loss measures the difference between the predicted and actual word labels. The accuracy is defined as the ratio of the number of correctly recognized words to the total number of words in the test set, as defined in Equation 1:

$$Accuracy = \frac{Number\ of\ correctly\ recognized\ words}{Total\ number\ of\ words} \quad (1)$$

The loss is typically computed using a loss function such as cross-entropy or mean squared error. This function quantifies how much the predicted word labels differ from the actual labels, and the performance goal is to minimize this difference during training. The loss is defined in Equation 2:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}) \quad (2)$$

where $\boldsymbol{N}$ is the total number of samples, $\boldsymbol{M}$ is the number of classes, $\boldsymbol{y_{ij}}$ is the ground-truth label of the $\boldsymbol{i}$-th sample for the $\boldsymbol{j}$-th class, and $\boldsymbol{p_{ij}}$ is the predicted probability of the $\boldsymbol{i}$-th sample for the $\boldsymbol{j}$-th class.

3- Research Methodology

In this study, the research methodology is outlined in Figure 2. Firstly, the dataset is divided into two subsets with different difficulty levels based on the viseme sequences. After dividing the dataset into subsets accordingly, both subsets are used in training the models. The input video is first undergoing data pre-processing to get the cropped lip ROI. The training data undergoes random crop, horizontal crop, and time mask; validation data undergoes center cropping; testing data undergoes center cropping and gamma correction to simulate conditions in different lighting. The cropped training data ROI is then fetched to the SlowFast frontend network, followed by TCN used in the backend network for training the model. Finally, the trained model is evaluated on four experiments scenarios, which are elaborated in *Section 4.1*, *4.2*, *4.3*, and *4.4*.

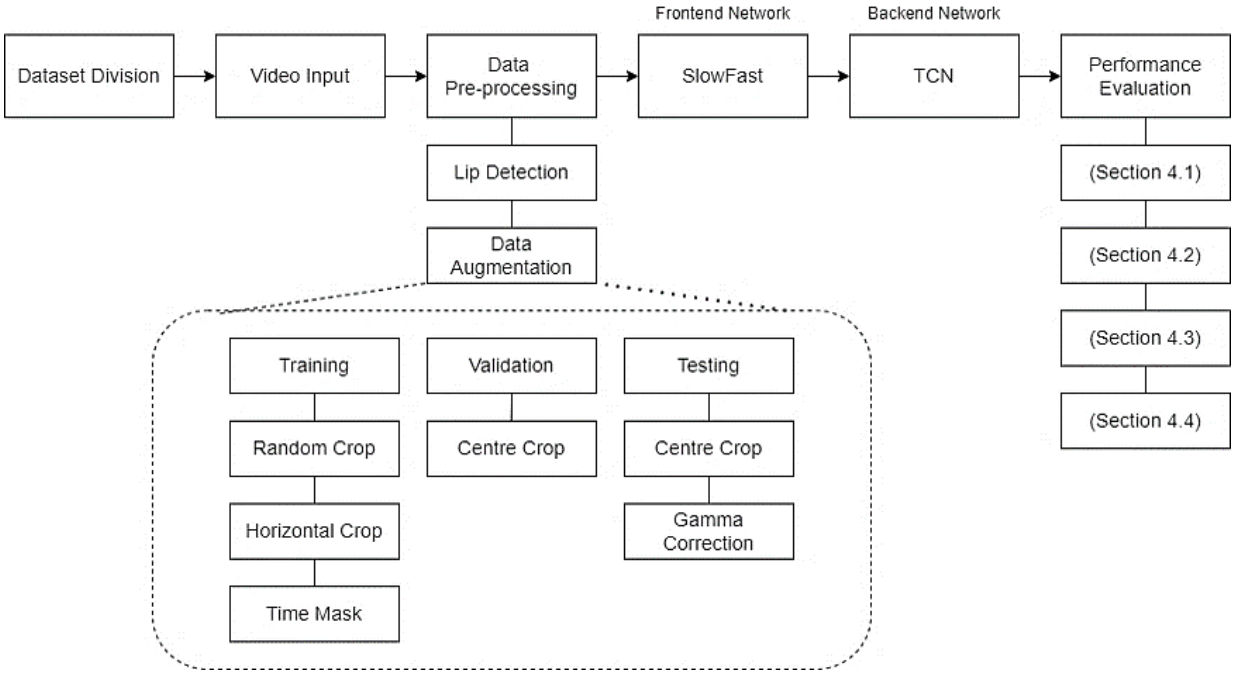

Figure 2. Flowchart of Research Methodology

3-1- Dataset

Two subsets are selected from the LRW based on their difficulty level, specifically considering the identical viseme sequences. Visemes are the basic visual units of speech that are produced by the lip movements and positions. Homopheme words refer to those that have identical viseme sequences. Table 1 shows the sample classes of the selected subsets, where the dataset is divided into subset A (easy) and subset B (hard). Firstly, the subset A consists of 23 classes of unique words without any homopheme pairs. Next, the dataset subset B consists of 48 classes of words that include homopheme pairs from the corpus. The viseme sequences of the homopheme pairs are presented in Table 2.

Table 1. Sample classes in two subsets

Non-homophemic Dataset A (23 Classes)	Homophemic Dataset B (48 Classes)	
BETTER	BETTER	NIGHT
BIGGEST	BIGGEST	NORTH
BILLION	BILLION	NOTHING
BRITAIN	BRITAIN	PRISON
BUILD	BUILD	PRIVATE
COUNCIL	COUNCIL	PROVIDE
COURSE	COURSE	RATHER
DIFFERENCE	COURT	RECENT
GIVING	DIFFERENCE	RESULT
HOUSE	DIFFERENT	SINCE
HOUSING	GIVING	SPEND
KNOWN	HOUSE	SPENT
LEAST	HOUSING	THEIR
MEETING	KNOWN	THERE
PRIVATE	LEAST	THINGS
RATHER	LIVING	THOSE
RECENT	LONDON	THOUGHT
SINCE	MATTER	THREAT
SPEND	MEANS	TRADE
THEIR	MEETING	WEATHER
THOSE	MILLION	WHETHER
THREAT	MINUTES	WHOLE
WORDS	MISSING	WORDS
	NEEDS	WORST

Table 2. Viseme sequence of the homopheme pairs

Homopheme Pairs		Viseme Sequence
BETTER	MATTER	P EY T ER
BIGGEST	MINUTES	P IY K AA T T
BILLION	MILLION	P IY K K AA K
BRITAIN	PRISON	P W IY T AA K
BUILD	MEANS	P IY K T
COUNCIL	LONDON	K AA K T AA K
COURSE	COURT	K AO W T
COURSE	NORTH	K AO W T
COURT	NORTH	K AO W T
DIFFERENCE	DIFFERENT	T IY F ER AA K T
GIVING	LIVING	K IY F IY K
HOUSE	NIGHT	K AA T
HOUSING	NOTHING	K AA T IY K
KNOWN	WHOLE	K AO K
LEAST	NEEDS	K IY T T
MEETING	MISSING	P IY T IY K
PRIVATE	PROVIDE	P W AA F AA T
RATHER	WEATHER	W EY T ER
RATHER	WHETHER	W EY T ER
RECENT	RESULT	W IY T AA K T
SINCE	THINGS	T IY K T
SPEND	SPENT	T P EY K T
THEIR	THERE	T EY W
THOSE	THOUGHT	T AO T
THREAT	TRADE	T W EY T
WEATHER	WHETHER	W EY T ER
WORDS	WORST	W ER T T

3-2- Data Pre-processing

The mouth ROI is extracted using lip detection, following the same strategy as [14], for all frames extracted from input video. Face detection and face alignment are performed to ensure consistent facial positioning across frames. Each frame undergoes alignment with respect to a standardized mean face shape. Then, the ROI are cropped with a size of 96 × 96 pixels from the aligned face image to ensure that the lip region remains approximately centered in the cropped images for all the frames. Lastly, the cropped ROI is transformed into grayscale. Among the 68 landmark points extracted from every frame with the Dlib library [36], only 20 points corresponding to the lip region are utilized in the proposed approach.

Data augmentations are performed during training based on [35], using 88 × 88 pixels on random cropping windows of the image with random horizontal flip on a probability of 0.5 flip ratio, in the aspect of the spatial augmentation. In term of temporal augmentation, N consecutive frames are selected for each training sequence, where N is randomly sampled from a uniform distribution ranging from 0 to N_{max}. Each of these masked frames is then replaced with the mean frame of its corresponding sequence. This technique of introducing temporal noise into the dataset is performed based on [35], where various forms of data augmentation are used to improve model robustness in situations with missing frames, as illustrated in Figure 3. The details are as shown in Table 3. For validation and testing, the data undergoes center cropping where the central patch of the images sequence is cropped to 88 × 88 pixels. The testing set undergoes preprocessing with gamma values of 0.5, 1.0 and 1.5, to represent different illumination levels, as illustrated in Figure 4.

Table 3. Details of data augmentation techniques

Data Augmentation Techniques	Details
Random Crop	• Resized to 88 × 88 pixels.
Horizontal Flip	• 0.5 probability flip ratio.
Time Mask	• N consecutive frames selected. • N randomly sampled from a uniform distribution (0 to N_{max}). • Masked frame replaced with mean frame.

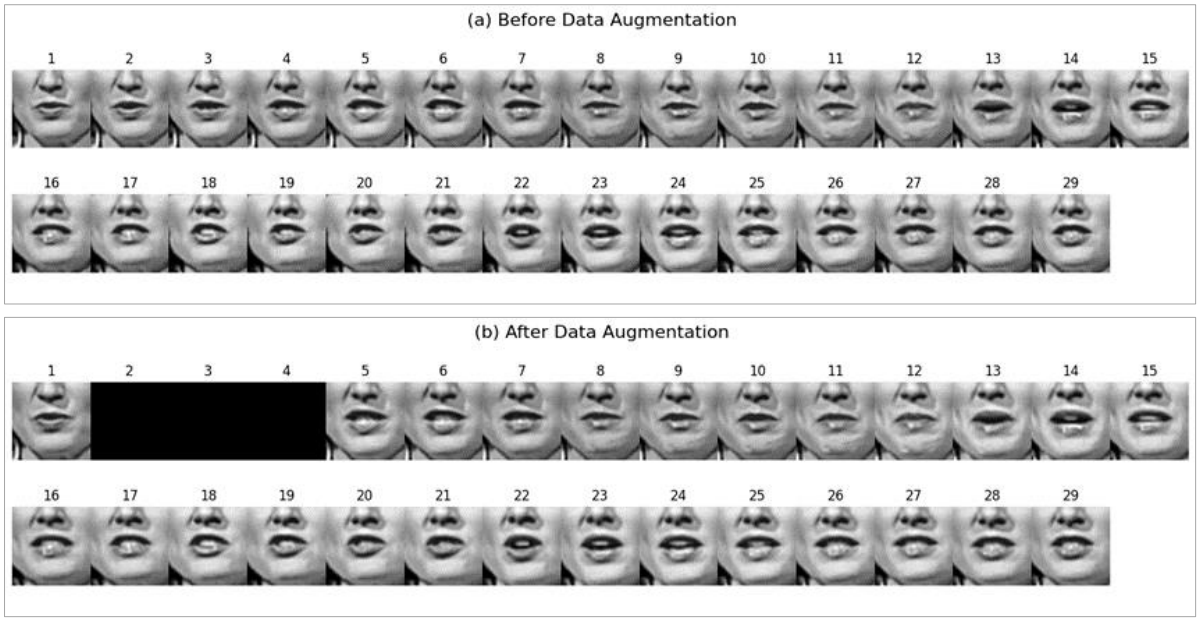

Figure 3. Example of 'BETTER' viseme (a) Before data augmentation (b) After data augmentation

Figure 4. Example of testing set with different gamma values representing varying illumination levels

3-3- SlowFast Architecture

The SlowFast architecture is designed to process video inputs through two pathways: the Slow pathway and the Fast pathway, each capturing different temporal resolutions. The network expects a tuple of two tensors, one for each pathway. In the Slow pathway, video inputs are processed by initially sampling a subset of frames sparsely from the input video sequence. This pathway focuses on processing low-frame-rate inputs to capture long-term dependencies. In this implementation, two frames are selected for processing in the Slow pathway from the video sequence of 29 frames. This selection is influenced by the stride parameter, which determines the temporal spacing between the sampled frames. A larger temporal stride allows the network to effectively capture long-term dependencies while processing fewer frames. Each frame undergoes a series of operations of 3D convolution, batch normalization, ReLU activation, and max pooling, as depicted in Figure 4. A 3D convolutional layer with a kernel size of (5, 7, 7) is applied, followed by batch normalization and ReLU activation, to standardize and enhance the non-linearity of the feature maps. Max pooling with a kernel size of (1, 3, 3) and a stride of (1, 2, 2) further reduces the spatial dimensions while preserving the essential features.

Concurrently, the Fast pathway operates with a higher temporal resolution, processing 14 frames from the same input video sequence of 29 frames. The choice of processing 14 frames in the Fast pathway is influenced by a smaller temporal stride, allowing for more frequent sampling of frames. To integrate information from high frame rate inputs and to capture short-term detailed motions, the Fast pathway operates concurrently with the Slow pathway. At the first fusion point, features from the Fast pathway are integrated into the Slow pathway using 3D convolution to align the temporal resolutions. This is followed by batch normalization and ReLU activation to maintain consistency in feature representation across both pathways. This fusion process repeats in subsequent stages (s2, s3, s4, s5) with subsequent fusion points, each involving multiple blocks of 3D convolution, batch normalization, and ReLU activation, where more features from both pathways are merged. In the final stage, additional layer of adaptive average pooling is used to reduce the spatial dimensions, before flattening the output to a 1D tensor. Table 4 shows the parameters that define the structure and fusion mechanism within the SlowFast architecture.

Table 4. SlowFast configuration

Parameter	Value
Width Per Group	64
Number of Groups	1
Fusion Conv Channel Ratio	2
Fusion Kernel	3
Alpha	8

3-4- Temporal Convolutional Network Architecture

The features output from the SlowFast network is then passed to a TCN backend. It uses the kernel sizes of 3, 5, and 7 that operate over four layers to capture the patterns over different temporal resolutions. Each temporal convolution has three branches, each with kernel size 3, 5, and 7 respectively, as depicted in Figure 5. Each convolution layer has different kernel sizes, followed by batch normalization and ReLU activation, similar to Martinez et al. research [14]. The final output from the TCN is averaged across the batch using a consensus function, which is then passed through a fully connected layer to produce the final classification scores. Table 5 shows the key parameters that define the architecture of TCN.

Table 5. TCN configuration

Parameter	Value
TCN Dropout	0.2
TCN Depthwise Pointwise Conv	False
TCN Kernel Size	[3, 5, 7]
TCN Num Layers	4
TCN Width Mult	1.0

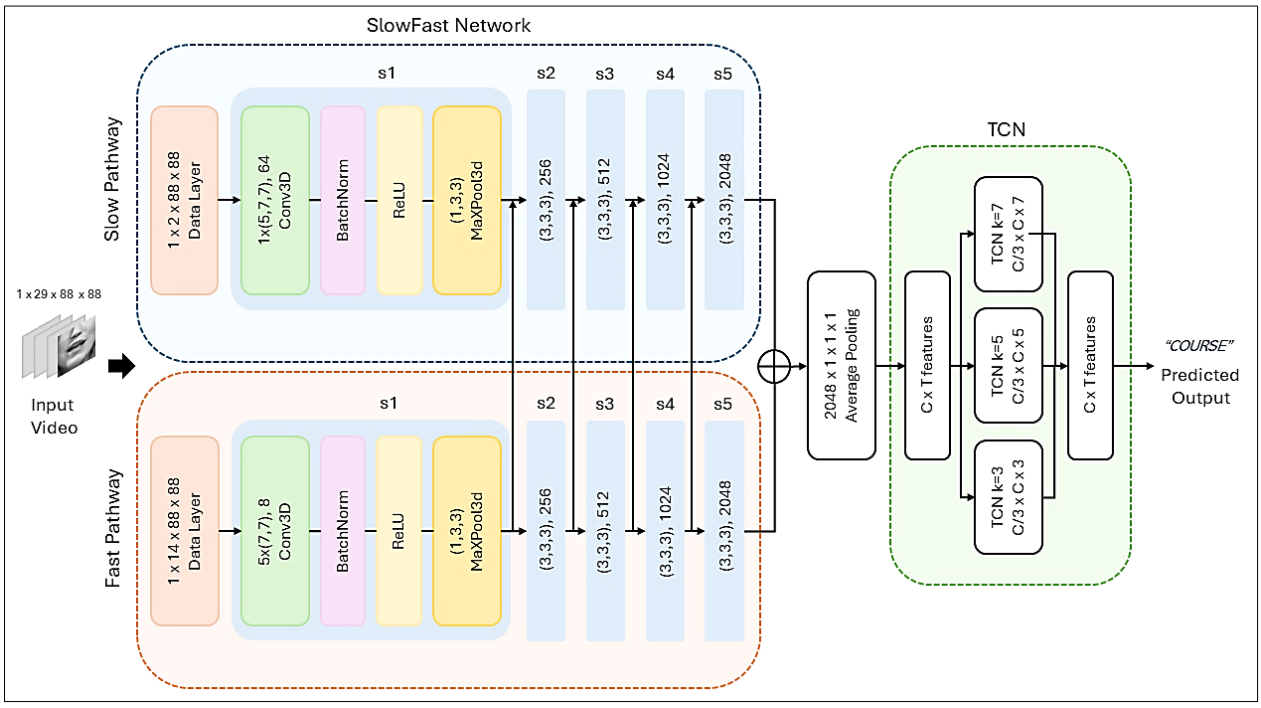

Figure 5. SlowFast-TCN model network architecture

4- Experiment Results

This section presents a series of the experimental results evaluating on the performance of the proposed SlowFast-TCN. Section 4.1 presents the experiments that were carried out on the subsets of the LRW dataset in comparison with the existing works employed using different models. The selected baseline models were chosen due to their proven effectiveness in VSR. These baselines models represent a range of approaches related to spatio-temporal feature extraction, which are suitable for evaluating the improvements of the proposed approach. Section 4.2 presents a detailed analysis on the performance of the proposed approach on the homopheme word pairs, specifically in subset B of the

dataset. Section 4.3 reports the results of the ablation experiments, highlighting the contribution of each component within the proposed approach. Finally, Section 4.4 presents the experiments that evaluate on varying illumination to assess the robustness of the proposed SlowFast-TCN approach across different environmental lighting scenarios.

4-1- Comparison with Existing Works

To evaluate the effectiveness of our proposed approach, its performance was compared with several existing models trained on the easy and hard datasets. Table 6 presents the comparison of the performance of the proposed approach with different models trained on the LRW subset A, consisting of 23 classes. The proposed SlowFast-TCN achieved the highest accuracy of 92.3% and the lowest loss of 0.2515, on an execution time of 33.058 seconds. Following that, the combination of the 3D CNN-ResNet-18 with MS-TCN achieved the second highest accuracy of 92.17% and 0.3363 loss, on an optimal execution time of 34.062 seconds. Using the same frontend network with a different variant of TCN, the combination of 3D CNN-ResNet-18 with DC-TCN showed a slight reduction to 91.91% accuracy and 0.3336 loss, on an increased computational time of 43.624 seconds. In contrast, the 3D CNN-ResNet-18 with LSTM exhibited a substantial drop in accuracy to 76.1%, with the highest loss at 0.5913 and the highest execution time of 92.077 seconds, indicating the poorer performance. The ShuffleNetV2 with TCN achieved an accuracy of 89.65% and a loss of 0.4083, while attaining a shorter execution time of 29.974 seconds. The addition of MS-DC-TCN to ShuffleNetV2 improved the accuracy to 91.57% and reduced the loss to 0.3216 on the shortest execution time of 29.531 seconds. Although the proposed SlowFast-TCN approach was not the best in terms of execution time, it balanced the performance with the best results in accuracy and loss, along with a moderate execution time.

Table 6. Comparison of proposed approach with different models trained on LRW dataset subset A (23 Classes)

Experiment	Accuracy (%)	Loss	Inference Execution Time (s)
3D CNN-ResNet-18 + DC-TCN [24]	91.91	0.3336	43.624
3D CNN-ResNet-18 + MS-TCN [14]	92.17	0.3363	34.062
3D CNN-ResNet-18 + LSTM [10]	76.10	0.5913	92.077
ShuffleNetV2 + TCN [26]	89.65	0.4083	29.974
ShuffleNetV2 + MS-DC-TCN [26]	91.57	0.3216	29.531
Proposed SlowFast-TCN	**92.30**	**0.2515**	**33.058**

Furthermore, Table 7 presents the comparison of the performance of the proposed approach with different models trained on a harder dataset, subset B, that contains 48 classes with homopheme word pairs. Overall, all models showed poorer performances on this dataset, with a drop of approximately ±10%. The performance for each model followed the same trends as in the previous experiments. The proposed SlowFast-TCN attained the best performance in accuracy and loss of 82.02% and 0.6497, respectively. Similarly, ShuffleNetV2 with MS-DC-TCN had the least computational time of 31.501 seconds.

Table 7. Comparison of proposed approach with different models trained on LRW dataset subset B (48 Classes).

Experiment	Accuracy (%)	Loss	Inference Execution Time (s)
3D CNN-ResNet-18 + DC-TCN [24]	81.13	0.6637	54.715
3D CNN-ResNet-18 + MS-TCN [14]	81.46	0.6654	45.485
3D CNN-ResNet-18 + LSTM [10]	69.91	0.8361	80.043
ShuffleNetV2 + TCN [26]	79.29	0.7102	31.571
ShuffleNetV2 + MS-DC-TCN [26]	80.31	0.6521	31.501
Proposed SlowFast-TCN	**82.02**	**0.6497**	**36.741**

In both subsets of the LRW, the proposed SlowFast-TCN consistently outperformed the other models by achieving the highest accuracy and the lowest loss on a competitive execution time, while maintaining its robustness in handling more classes. The proposed approach balanced high accuracy, low loss, and optimal execution time, making it an effective approach for tasks requiring efficiency in practical applications, where both accuracy and computational efficiency are equally important.

4-2- Results of Proposed Approach on Homopheme Word Pairs

Since subset B is challenging for all models due to the inclusion of homopheme word pairs, a detailed analysis on the performance of the proposed approach on this subset is insightful. Figure 6 shows the confusion matrix of the proposed SlowFast-TCN evaluated on the 48 classes containing homopheme word pairs from Table 7. The results showed that

misclassification occurs the most on the word pairs ‘SPEND’ and ‘SPENT’, in which 17 ‘SPEND’ had been predicted as ‘SPENT’ and 10 ‘SPENT’ had been predicted as ‘SPEND’. Followed by that, ‘WORDS’ and ‘WORST’ came second with 3 samples of ‘WORDS’ predicted as ‘WORST’ and 15 samples of ‘WORST’ predicted as ‘WORDS’. Other than that, ‘LIVING’ and ‘GIVING’ had been heavily misclassified as well, with 14 samples of ‘LIVING’ predicted as ‘GIVING’ and 7 ‘GIVING’ predicted as ‘LIVING’. Also, misclassification occurred on homopheme word pairs such as ‘DIFFERENCE’-‘DIFFERENT’, ‘COURSE’-‘COURT’, and ‘WEATHER’-‘WHETHER’. Notably, although not considered as homopheme, but words with the same stem such as ‘HOUSE’ and ‘HOUSING’ were observed to be often misclassified too.

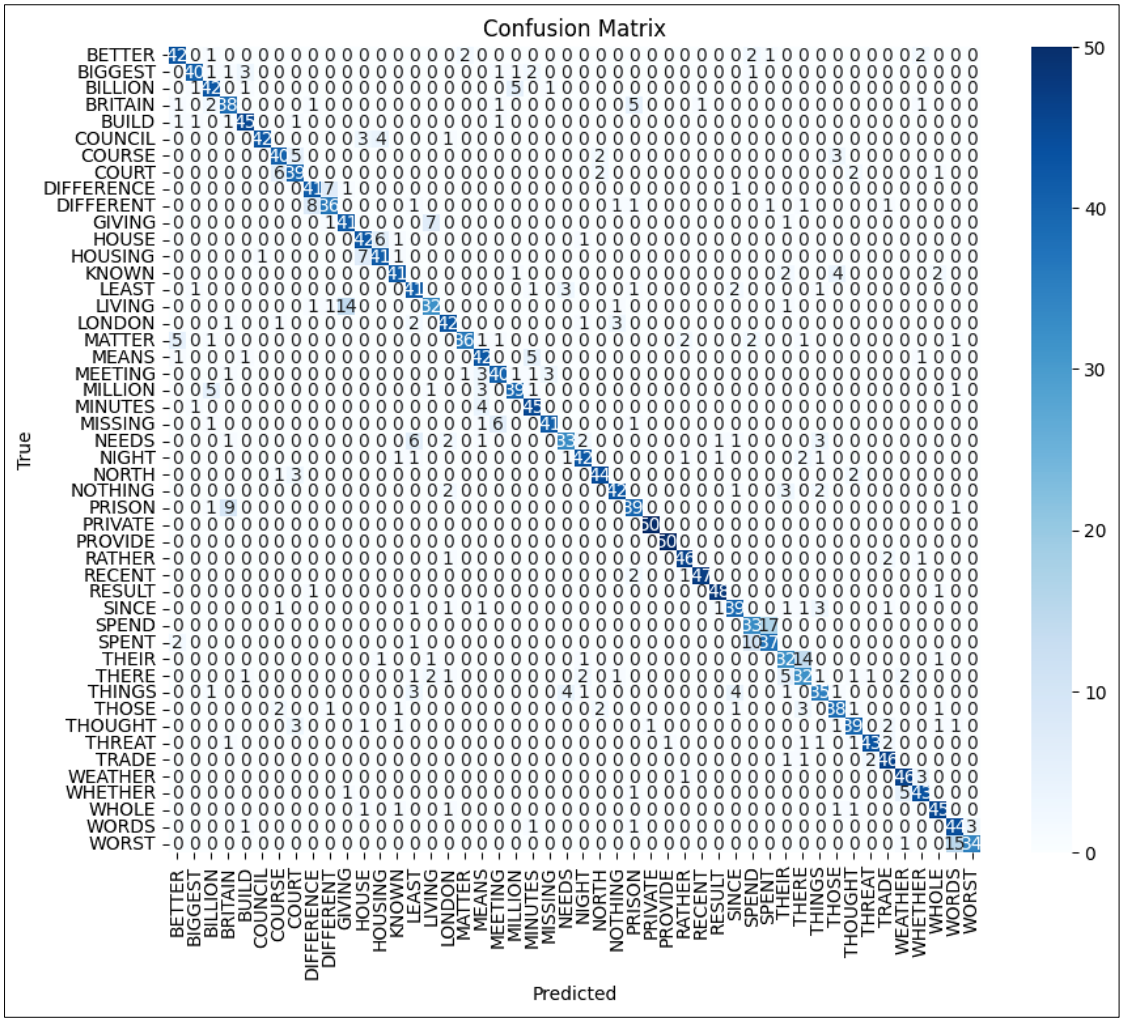

Figure 6. Confusion matrix of proposed SlowFast-TCN on homopheme word pairs.

From the highest misclassified homopheme word pair obtained from Figure 6, comparisons on the word precision of the proposed approach against a top performing baseline model (3DCNN-ResNet-MS-TCN) for these homopheme word recognitions are shown in Table 8. Each row represents a pair of homophemes along with their precision and the performance change. The baseline model was chosen based on its high accuracy and relatively low loss compared to other models in Tables 6 and 7.

The proposed approach showed substantial improvement for words like ‘WORST’ (+12.30%) and ‘WHETHER’ (+4.31%), suggesting enhanced accuracy in these cases. However, it had decreased precision for words like "SPEND" (-6.81%) and "WORDS" (-4.67%), indicating some decline in performance. While it improved significantly for certain words like ‘WORST’ and ‘WHETHER’, it also had decreased performance for others like ‘SPEND’ and ‘WORDS. For pairs like "DIFFERENCE" (+2.06%) and "DIFFERENT" (+2.07%), the changes were moderate and positive. Minor changes were seen in words like "SPENT" (+1.78%) and "GIVING" (-0.62%).

Overall, the proposed approach demonstrated more competent performance compared to the baseline model. Specifically, SlowFast-TCN achieved the highest precision of 91.89%, surpassing the best performance of the baseline model (85.54%). Conversely, even the lowest precision achieved by SlowFast-TCN, which is 66.07%, remained slightly higher than that of the baseline model (64.29%).

Table 8. Homopheme word precision and changes of proposed SlowFast-TCN compared to the top performing baseline (3DCNN-ResNet-MS-TCN)

Homopheme Pair		Word Precision (%)		Performance Change (%)
		Proposed SlowFast-TCN	Top Performing Baseline (3DCNN-ResNet-MS-TCN)	
1	SPEND	68.75	75.56	-6.81
	SPENT	66.07	64.29	+1.78
2	WORDS	69.84	74.51	-4.67
	WORST	**91.89**	79.59	**+12.30**
3	LIVING	74.42	70.45	+3.97
	GIVING	71.93	72.55	-0.62
4	DIFFERENCE	78.85	76.79	+2.06
	DIFFERENT	78.26	76.19	+2.07
5	COURSE	78.43	82.35	-3.92
	COURT	76.47	78.85	-2.38
6	WEATHER	85.19	**86.54**	-1.35
	WHETHER	84.31	80.00	+4.31

4-3- Ablation Results of Proposed Approach

To understand the contribution of each component in the final model, ablation experiments were conducted. Table 9 presents the results of these experiments, where each component was removed one at a time. Removing random crop only marginally decreased the accuracy to 92.2%, suggesting its importance but also its compensability by other augmentations. Similarly, excluding horizontal flip resulted in a minimal accuracy drop to 92.23%, indicating its relatively minor contribution when combined with random crop and time mask. However, excluding time mask led to a more noticeable decrease to 91.8%, indicating its significant role in handling temporal variations within sequences. Removing the SlowFast architecture resulted in an accuracy drop to 90.4%, underscoring its contribution, albeit less critical compared to TCN. Notably, excluding TCN caused the most significant accuracy reduction to 89.1%, emphasizing its role in capturing temporal dependencies in model performance. Overall, the results showed that the proposed approach with the full configuration achieved the highest accuracy. This may indicate that while all components and data augmentation techniques contributed positively, TCN and time mask were particularly more pronounced in enhancing the model's ability to handle temporal complexities in the data, thereby improving the overall performance.

Table 9. Results of ablation experiments on proposed approach

Data Augmentation			SlowFast	TCN	Accuracy (%)
Random Crop	Horizontal Flip	Time Mask			
✓	✓	✓	✓	✓	92.30
	✓	✓	✓	✓	92.20
✓		✓	✓	✓	92.23
✓	✓		✓	✓	91.80
✓	✓	✓		✓	90.40
✓	✓	✓	✓		89.10

4-4- Results of Proposed Approach on Varying Illumination

While the proposed SlowFast-TCN demonstrated optimal performance under standard conditions, its robustness in practical applications under varying environmental lighting was also being assessed. Since lighting variation is a well-known challenge related to video processing, it is worth investigating how the changes of illumination impact the model's ability to distinguish between visually similar words. Table 10 shows the results of the proposed SlowFast-TCN and the top-performing baseline (3DCNN-ResNet-MS-TCN) tested on different illumination settings to simulate different lighting conditions in practical application. For a gamma value of 0.5, SlowFast-TCN achieved an accuracy of 79.13%, while the baseline model achieved an accuracy of 77.17%. This indicates that the proposed approach performed better than the baseline under lower illumination, although both models experienced a decrease in accuracy compared to the standard illumination level. At the standard illumination level (gamma value of 1), SlowFast-TCN achieved its highest

accuracy of 81.46%, significantly outperforming the baseline model, which maintained an accuracy of 77.17%. This demonstrates the superior performance of the proposed approach under normal lighting conditions. When the gamma value was increased to 1.5, the accuracy of SlowFast-TCN dropped to 80.21%, while the baseline model accuracy slightly decreased to 77.04%. Despite the decrement, the proposed approach still maintained higher accuracy than the baseline under varying illumination.

The proposed SlowFast-TCN consistently achieved higher accuracy compared to the top-performing baseline model across different illumination conditions. However, the proposed approach exhibited a more significant decrease in accuracy when moving away from the standard illumination level (gamma = 1). Specifically, the accuracy dropped by 2.33% (from 81.46% to 79.13%) with gamma = 0.5 and by 1.25% (from 81.46% to 80.21%) with gamma = 1.5. In contrast, the baseline model showed minimal variation, with accuracy remaining around 77.17% across all gamma values, indicating a drop of only 0.13% at gamma = 1.5.

While the proposed SlowFast-TCN demonstrated higher absolute performance under all tested conditions, its sensitivity to changes in illumination was more pronounced compared to the baseline model. The model demonstrated better accuracy under standard illumination levels, where the lighting was neither too dark nor too bright, allowing for clearer visual cues of lip movements. In extreme lighting scenarios, the proposed SlowFast-TCN performed better under darker conditions than brighter conditions. However, its ability to accurately recognize homophemes significantly decreased in both conditions, indicating that extreme lighting conditions obscure the subtle differences necessary for distinguishing between visually similar words. This suggests that while SlowFast-TCN is more effective overall, it may require additional augmentation or normalization to maintain robustness across varying lighting conditions. The top-performing baseline, though less accurate overall, had more stable performance under different illuminations, indicating better inherent robustness to such variations. This comparison highlights the trade-offs between absolute performance and stability under various environmental conditions.

Table 10. Results of proposed approach and top performing baseline tested on different illumination

Gamma Value	Proposed SlowFast-TCN	Top Performing Baseline (3DCNN-ResNet-MS-TCN)
0.5	79.13	77.17
1.0	81.46	77.17
1.5	80.21	77.04

5- Discussion

This study has proven that the proposed SlowFast-TCN outperforms traditional baseline models in terms of accuracy while still maintaining an optimal execution time. The architecture of SlowFast reveals its ability to capture fine-grained temporal dynamics and motion at multiple timescales, which significantly contributes to its superior performance. The slow pathway processes low-frame-rate inputs, capturing long-term dependencies, while the fast pathway handles high-frame-rate inputs, focusing on short-term detailed motions. This dual-pathway approach allows the model to effectively leverage both detailed and contextual information, resulting in higher accuracy rates. This dual pathway performs better because it mitigates the limitations of a single pathway (either high frame rate or low frame rate inputs), which often fails to simultaneously capture detailed temporal dependencies and motion patterns. By integrating both slow and fast pathways, our approach addresses these shortcomings and enhances the model's ability to comprehend the complex temporal dynamics inherent in lip movements. Furthermore, TCNs are known for their ability to handle sequential data effectively, providing a more nuanced understanding of temporal dependencies [20]. The ablation study (Table 9) highlights the role of TCN in handling temporal variations. Specifically, the inclusion of both SlowFast architecture and TCN enhances the model's ability to capture and process temporal dependencies within video sequences. The key differences between the SlowFast-TCN model and the baseline models lie in their approach to handling spatio-temporal feature extraction. For instance, 3D CNN-ResNet-18 combined with various temporal networks (DC-TCN, MS-TCN, LSTM) primarily relies on separate spatial and temporal processing stages, while the proposed SlowFast-TCN integrates these processes to capture both slow and fast motion dynamics simultaneously. When compared to traditional models that often struggle with temporal dependencies in video data, the design of the proposed approach proves to be more efficient and accurate in capturing the nuances of lip movements.

In terms of the execution time, it is shown that ShuffleNetV2 has the shortest inference execution time, while the SlowFast network is just slightly slower compared to it. ShuffleNetV2 is known for its exceptionally fast execution time due to its lightweight architecture. It utilizes a channel split and shuffle mechanism that reduces the computation cost

significantly. By splitting the input channels into groups and shuffling them, ShuffleNetV2 minimizes the memory access cost and ensures efficient use of computational resources [37]. While the SlowFast network is slightly slower compared to ShuffleNetV2, it still maintains a competitive performance with only a marginal difference of 4 to 5 seconds. The dual-pathway of the SlowFast network requires more computational resources compared to the single-pathway design of ShuffleNetV2. However, the execution time remains optimal, which may be due to the parallel processing of the two pathways and the efficient handling of temporal dynamics [19].

The ablation study highlights the individual contributions of the data augmentation techniques, SlowFast and TCN components to the overall performance of the proposed approach. Time masking contributes more to the performance compared to other data augmentation techniques. This is because time masking involves randomly masking out portions of the input sequence, forcing the model to learn to predict the missing information. This technique is particularly useful in sequence learning tasks as it improves the model's ability to handle incomplete data and learns more robust temporal representations. The significant impact of time masking on accuracy highlights its importance in sequence learning to handle missing segments that allows it to become more resilient to variations and noise [38]. The TCN component contributes the most among all the components due to its superior capability in handling sequential data. TCN uses causal convolutions to ensure that predictions at a certain time step depend only on past time steps and preserve the temporal order of the input data. Additionally, TCN can capture long-range dependencies more effectively than traditional recurrent networks like LSTM or GRU due to their hierarchical structure [20].

Removing either the SlowFast or the TCN component results in a significant drop in accuracy, which indicates that both elements are crucial for the impressive performance of the proposed approach. Specifically, the ability of the SlowFast network to process temporal features across different time scales is essential for accurate lip-reading, while the TCN capability for handling sequential data enhances the model's understanding of temporal dependencies. The complementary nature of these components ensures that the proposed approach captures a wide range of features, from fine-grained movements to broader temporal patterns. This synergy explains why the proposed approach outperforms those lacking one of these critical components.

Despite its ability to have optimal accuracy under standard conditions, the SlowFast-TCN network shows reduced robustness in varying illumination conditions. This can be attributed to the model's reliance on visual features that are sensitive to lighting changes. While the SlowFast network effectively captures motion and temporal dynamics, it does not inherently adjust for variations in illumination, which leads to performance degradation in non-ideal lighting scenarios. The TCN component, although excellent for temporal sequence processing, does not compensate for changes in lighting. This limitation is common in VSR, where lighting inconsistencies can mislead the feature extraction process in the frontend network. For instance, shadows caused by uneven lighting can obscure crucial parts of the lips, while overexposure can wash out the details and lead to errors in VSR [39]. This limitation suggests a need for additional preprocessing steps or data augmentation techniques that can simulate different lighting conditions during training, thereby improving the model's robustness.

5-1- Limitations

Firstly, this study focuses only on the English language. However, lip movements vary across languages due to distinct viseme structures. Future works can extend the investigations to other languages, considering the unique phonetic characteristics and viseme mappings. This expansion would allow for more robust and language-agnostic VSR systems. Additionally, this study primarily addresses the homopheme problem, where different words share similar lip movements. While this is a critical challenge in VSR, other equally important factors are not explored in depth. For instance, speaker variability, which involves how different individuals articulate the same word, can significantly impact the recognition accuracy. Future work should investigate approaches to handle speaker-specific variations effectively.

Furthermore, another limitation of this study is the dataset constraints. This study relies on the LRW dataset for training and evaluation. While LRW provides a substantial collection of lip movements across various words, it may not fully represent the diversity of visual speech encountered in real-world scenarios. Also, the dataset is not being used in full due to computational limitations. Future work should consider using the complete dataset and incorporating additional datasets that cover different languages, accents, and environmental conditions to enhance the generalizability of the models.

While the proposed SlowFast-TCN network achieves better word recognition accuracy compared to baseline models, its real-time applicability remains unexplored. Real-time processing is important for practical deployment in applications like speech-to-text transcription, assistive communication devices, and security systems. Hence, future work can consider evaluating this.

Lastly, the current approaches employed in VSR primarily rely on computer vision techniques, which focus only on visual information when building the models for VSR. This approach may encounter performance limitations because VSR is a complex task and involves the understanding of language. To address this complexity, it may need to consider not only visual cues but also the logical, contextual, semantic, and syntactic aspects of language. Therefore, a limitation of this research is that it does not account for the contextual, semantic, and syntactic aspects of the language used.

6- Conclusion and Future Works

In summary, this study developed a novel deep learning approach called the SlowFast-TCN network for word-level VSR in English. By leveraging data augmentation techniques such as random crop, horizontal flip, and time mask to enhance training data variability, the combination of SlowFast and TCN networks can extract spatio-temporal features from video sequences and classify spoken words based on lip movements. A comparison of the proposed approach with several baseline models has been conducted. These experiments were evaluated on both subsets of the dataset with different difficulty levels of the homophemes. It is also evaluated on different illumination changes to test the robustness of the SlowFast-TCN. It is proven that the proposed approach attains the highest accuracy compared to other models. However, it is also shown that the proposed approach struggles to perform well on homophemes compared to unique words. A comparative experiment against the baseline model highlights that the proposed approach has a promising performance under standard conditions but is facing challenges in robustness for varying illumination conditions. Despite these strengths, the proposed approach still struggles with homopheme recognition due to the visual similarity between certain words, such as 'SPEND' and 'SPENT,' which exhibit nearly identical lip movements that create the challenge to differentiate them solely based on visual cues. Additionally, another challenge would be the contextual dependency. The model's reliance on spatio-temporal features without incorporating contextual understanding limits its ability to disambiguate homophemes. For example, the meaning of a homopheme can often be clarified by the surrounding words in a sentence, but the proposed approach is yet to leverage this information.

To address the complexity of VSR and its reliance on purely visual information, future work may integrate the use of language models for contextual understanding that could account for the bottleneck in homopheme aspects of VSR. As such, a language model can provide a context that helps the visual model to distinguish between similar-looking words by considering the surrounding words and their typical usage patterns. Developing a VSR system with both visual cues and language models has the possibility of resulting in a more accurate and context-aware recognition, thereby overcoming the limitations of current computer vision-focused approaches. In addition, future research should explore methods focusing on speaker-independent problems to effectively manage how different individuals articulate the same word. This could involve developing adaptive models that can learn and adjust to individual speaker characteristics or incorporating speaker-specific features into the training process. By tackling this variability, VSR systems can achieve higher precision and reliability across diverse speakers. To overcome the dataset limitations, future work should utilize the complete LRW dataset and integrate additional datasets that capture a broader range of visual speech. This includes datasets featuring different languages, accents, and environmental conditions. By enhancing the diversity and comprehensiveness of the training data, the generalizability and robustness of VSR models can be improved, making them more applicable to real-world scenarios. The real-time applicability of the proposed SlowFast-TCN model needs to be explored to ensure its practical deployment. Future work should focus on evaluating and optimizing the model for real-time processing, which is essential for applications like speech-to-text transcription, assistive communication devices, and security systems. Investigating the computational efficiency and latency of the model in real-time scenarios will help bridge the gap between research and practical use.

7- Declarations

7-1- Author Contributions

Conceptualization, N.Y.Y.H. and L.-Y.O.; methodology, N.Y.Y.H.; software, N.Y.Y.H.; validation, N.Y.Y.H, L.-Y.O., and M.-C.L. formal analysis, N.Y.Y.H., L.-Y.O., and M.-C.L.; investigation, N.Y.Y.H.; resources, N.Y.Y.H., L.-Y.O., and M.-C.L.; data curation, N.Y.Y.H.; writing—original draft preparation, N.Y.Y.H.; writing—review and editing, N.Y.Y.H., L.-Y.O., and M.-C.L.; visualization, N.Y.Y.H.; supervision, L.-Y.O. and M.-C.L.; project administration, L.-Y.O.; funding acquisition, L.-Y.O. and M.-C.L. All authors have read and agreed to the published version of the manuscript.

7-2- Data Availability Statement

Publicly available datasets were analyzed in this study. This data can be found here: *https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrw1.html.*

7-3- Funding

This work was supported by the Telekom Malaysia Research and Development under Grant RDTC/221073 (MMUE/230002).

7-4- Institutional Review Board Statement

Not applicable.

7-5- Informed Consent Statement

Not applicable.

7-6- Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

8- References

[1] Dixit, A., Sethi, P., Garg, P., Pruthi, J., & Chauhan, R. (2024). CNN based lip-reading system for visual input: A review. AIP Conference Proceedings, 3121(1), 40031. doi:10.1063/5.0221717.

[2] Hao, M., Mamut, M., Yadikar, N., Aysa, A., & Ubul, K. (2020). A survey of research on lipreading technology. IEEE Access, 8, 204518–204544. doi:10.1109/ACCESS.2020.3036865.

[3] Thapa, K. (2023). End-to-end Lip-reading: A Preliminary Study. Masters Thesis, London South Bank University, London, United Kingdom. doi:10.18744/lsbu.92zq5.

[4] Sheng, C., Kuang, G., Bai, L., Hou, C., Guo, Y., Xu, X., Pietikainen, M., & Liu, L. (2024). Deep Learning for Visual Speech Analysis: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 46(9), 6001–6022. doi:10.1109/TPAMI.2024.3376710.

[5] Kim, M., Yeo, J. H., & Ro, Y. M. (2022). Distinguishing Homophenes Using Multi-Head Visual-Audio Memory for Lip Reading. Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022, 36(1), 1007–1015. doi:10.1609/aaai.v36i1.20003.

[6] Fenghour, S., Chen, D., Guo, K., & Xiao, P. (2020). Disentangling homophemes in lip reading using perplexity analysis. arXiv preprint arXiv:2012.07528. doi:10.48550/arXiv.2012.07528.

[7] Jeon, S., Elsharkawy, A., & Kim, M. S. (2022). Lipreading architecture based on multiple convolutional neural networks for sentence-level visual speech recognition. Sensors, 22(1), 72. doi:10.3390/s22010072.

[8] Shi, B., Hsu, W. N., Lakhotia, K., & Mohamed, A. (2022). Learning audio-visual speech representation by masked multimodal cluster prediction. arXiv preprint arXiv:2201.02184. doi:10.48550/arXiv.2201.02184.

[9] Sepas-Moghaddam, A., Pereira, F., Correia, P. L., & Etemad, A. (2021). Multi-perspective LSTM for joint visual representation learning. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 16535–16543. doi:10.1109/CVPR46437.2021.01627.

[10] Stafylakis, T., & Tzimiropoulos, G. (2017). Combining residual networks with LSTMs for lipreading. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2017-August, 3652–3656. doi:10.21437/Interspeech.2017-85.

[11] Shashidhar, R., Patilkulkarni, S., & Puneeth, S. B. (2022). Combining audio and visual speech recognition using LSTM and deep convolutional neural network. International Journal of Information Technology (Singapore), 14(7), 3425–3436. doi:10.1007/s41870-022-00907-y.

[12] Fenghour, S. (2022). Viseme-based Lip-Reading using Deep Learning. Doctoral dissertation, London South Bank University, London, United Kingdom. doi:10.18744/lsbu.9280w.

[13] Ma, S., Wang, S., & Lin, X. (2020). A transformer-based model for sentence-level Chinese mandarin lipreading. Proceedings - 2020 IEEE 5th International Conference on Data Science in Cyberspace, DSC 2020, 78–81. doi:10.1109/DSC50466.2020.00020.

[14] Martinez, B., Ma, P., Petridis, S., & Pantic, M. (2020). Lipreading Using Temporal Convolutional Networks. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2020-May, 6319–6323. doi:10.1109/ICASSP40776.2020.9053841.

[15] Zhu, D., Han, C., Guo, J., & Sun, L. (2024). TWLip: Exploring Through-Wall Word-Level Lip Reading Based on Coherent SISO Radar. IEEE Internet of Things Journal, 11(19), 32310 - 32323. doi:10.1109/JIOT.2024.3427329.

[16] Chopadekar, G., Pandey, N., Rakhangi, N., Balsaraf, S., & Patil, V. (2024). Literature survey - lip reading model. International Research Journal of Innovations in Engineering and Technology, 8(4), 143. doi:10.47001/IRJIET/2024.804019.

[17] He, Y., Yang, L., Wang, S., & Liew, A. W. C. (2024). Lip Feature Disentanglement for Visual Speaker Authentication in Natural Scenes. IEEE Transactions on Circuits and Systems for Video Technology, 34(10), 9898-9909. doi:10.1109/TCSVT.2024.3405640.

[18] Lee, K. S. (2024). Improving the Performance of Automatic Lip-Reading Using Image Conversion Techniques. Electronics (Switzerland), 13(6), 1032. doi:10.3390/electronics13061032.

[19] Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). Slowfast networks for video recognition. Proceedings of the IEEE International Conference on Computer Vision, 2019-October, 6201–6210. doi:10.1109/ICCV.2019.00630.

[20] Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271. doi:10.48550/arXiv.1803.01271.

[21] Fenghour, S., Chen, D., Guo, K., Li, B., & Xiao, P. (2021). Deep Learning-Based Automated Lip-Reading: A Survey. IEEE Access, 9, 121184–121205. doi:10.1109/ACCESS.2021.3107946.

[22] Cheng, S., Ma, P., Tzimiropoulos, G., Petridis, S., Bulat, A., Shen, J., & Pantic, M. (2020). Towards Pose-Invariant Lip-Reading. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2020-May, 4357–4361. doi:10.1109/ICASSP40776.2020.9054384.

[23] Koumparoulis, A., & Potamianos, G. (2022). Accurate and Resource-Efficient Lipreading with Efficientnetv2 and Transformers. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2022-May, 5937–5941. doi:10.1109/ICASSP43922.2022.9747729.

[24] Ma, P., Wang, Y., Shen, J., Petridis, S., & Pantic, M. (2021). Lip-reading with densely connected temporal convolutional networks. Proceedings - 2021 IEEE Winter Conference on Applications of Computer Vision, WACV 2021, 2856–2865. doi:10.1109/WACV48630.2021.00290.

[25] Lahiri, A., Kwatra, V., Frueh, C., Lewis, J., & Bregler, C. (2021). LipsyNc3D: Data-Efficient Learning of Personalized 3D Talking Faces from Video using Pose and Lighting Normalization. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2754–2763. doi:10.1109/CVPR46437.2021.00278.

[26] Ma, P., Martinezy, B., Petridis, S., & Pantic, M. (2021). Towards practical lipreading with distilled and efficient models. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2021-June, 7608–7612. doi:10.1109/ICASSP39728.2021.9415063.

[27] Wei, D., Tian, Y., Wei, L., Zhong, H., Chen, S., Pu, S., & Lu, H. (2022). Efficient dual attention SlowFast networks for video action recognition. Computer Vision and Image Understanding, 222, 103484. doi:10.1016/j.cviu.2022.103484.

[28] Hovad, E., Hougaard-Jensen, T., & Clemmensen, L. K. H. (2024). Classification of Tennis Actions Using Deep Learning. arXiv preprint arXiv:2402.02545. doi:10.48550/arXiv.2402.02545.

[29] Calderó, M. S., Varas, D., & Bou-Balust, E. (2021). Spatio-temporal context for action detection. arXiv preprint arXiv:2106.15171. doi:10.48550/arXiv.2106.15171.

[30] Sheshpoli, A. J., & Nadian-Ghomsheh, A. (2019). Temporal and spatial features for visual speech recognition. Lecture Notes in Electrical Engineering, 480, 135–145. doi:10.1007/978-981-10-8672-4_10.

[31] Petridis, S., Stafylakis, T., Ma, P., Cai, F., Tzimiropoulos, G., & Pantic, M. (2018). End-to-End Audiovisual Speech Recognition. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2018-April, 6548–6552. doi:10.1109/ICASSP.2018.8461326.

[32] Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2017). Lip reading sentences in the wild. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, 2017-January, 3444–3450. doi:10.1109/CVPR.2017.367.

[33] Nemani, P., Krishna, G. S., Supriya, K., & Kumar, S. (2023). Speaker independent VSR: A systematic review and futuristic applications. Image and Vision Computing, 138, 104787. doi:10.1016/j.imavis.2023.104787.

[34] Mudaliar, N. K., Hegde, K., Ramesh, A., & Patil, V. (2020). Visual Speech Recognition: A Deep Learning Approach. 2020 5th International Conference on Communication and Electronics Systems, 1218–1221. doi:10.1109/icces48766.2020.9137926.

[35] Ma, P., Wang, Y., Petridis, S., Shen, J., & Pantic, M. (2022). Training Strategies for Improved Lip-Reading. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2022-May, 8472–8476. doi:10.1109/ICASSP43922.2022.9746706.

[36] King, D. E. (2009). Dlib-ml: A machine-learning toolkit. Journal of Machine Learning Research, 10, 1755–1758.

[37] Ma, N., Zhang, X., Zheng, H. T., & Sun, J. (2018). Shufflenet V2: Practical guidelines for efficient CNN architecture design. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 11218 LNCS, 122–138. doi:10.1007/978-3-030-01264-9_8.

[38] Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2019-September, 2613–2617. doi:10.21437/Interspeech.2019-2680.

[39] Butt, W. R., & Lombardi, L. (2021). Lip Detection and Tracking with Geometric Constraints under Uneven Illumination and Shadows. International Journal of Advanced Computer Science and Applications, 12(8), 17–24. doi:10.14569/IJACSA.2021.0120803.