

# Development and Evaluation of a 2D Convolutional Neural Network with Temporal Shift Module for Tagalog Visual Speech Recognition

Andrea Q. Ysais<sup>1\*</sup>, Mary Camille S. Santos<sup>1</sup>, Paolo Emerson L. Cordero<sup>1</sup>, Phil Andrei B. Sison<sup>1</sup>, Mikayla Darlene B. Domingo<sup>1</sup>

<sup>1</sup> College of Computing and Information Sciences, Systems Plus College Foundation, Angeles City, Philippines

\*Corresponding Author Email: [ysaisandrea827@gmail.com](mailto:ysaisandrea827@gmail.com)

## ABSTRACT

Visual Speech Recognition (VSR), or lip-reading, recognizes spoken content through visual information such as mouth and lip movements. This study will develop and evaluate a 2D Convolutional Neural Network (2D-CNN) combined with a Temporal Shift Module (TSM) for Tagalog visual speech recognition. The 2D-CNN will extract spatial features, while TSM will model temporal information across sequential video frames without using conventional 3D convolution. Connectionist Temporal Classification (CTC) will be used for sequence recognition and temporal alignment between visual features and corresponding Tagalog word sequences. The study will adopt an Explanatory Sequential Mixed-Methods Design with a predominantly quantitative approach involving dataset preparation, preprocessing, model development, training, validation, testing, and evaluation. Recognition performance will be assessed using Accuracy, Precision, Recall, F1-score, Matthews Correlation Coefficient (MCC), Character Error Rate (CER), and Word Error Rate (WER), while computational efficiency will be evaluated based on Training Time and Memory Requirements. A limited qualitative error analysis will then examine selected misclassified samples to identify observable mouth and lip movement patterns associated with recurring recognition errors. The study aims to provide empirical evidence on the recognition performance and computational efficiency of a 2D-CNN + TSM approach for Tagalog visual speech recognition.

**Keywords:** Visual Speech Recognition, Lip Reading, Tagalog, 2D-CNN, Temporal Shift Module, Deep Learning

## SECTION I — INTRODUCTION (THE PROBLEM AND ITS SETTING)

### 1.1 Introduction

Visual Speech Recognition (VSR), also known as lip-reading, is a technology that recognizes spoken words or word sequences from visible mouth and lip movements in video. Unlike conventional speech recognition, which relies on audio signals, VSR uses visual information to interpret speech. This makes it relevant in situations where speech audio is unavailable or unreliable and provides a basis for developing alternative forms of speech-related interaction.

Deep learning has enabled VSR systems to learn visual speech patterns from video sequences using different neural network architectures. 2D Convolutional Neural Networks (2D-CNNs) can extract spatial features from individual frames, while 3D Convolutional Neural Networks (3D-CNNs) can learn spatial and temporal features jointly. Other approaches combine convolutional networks with recurrent models, such as Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) networks, to model temporal information across frames. Temporal Convolutional Networks (TCNs) and Transformer-based approaches have also been explored for temporal modeling and sequence recognition. These approaches demonstrate that effective VSR

requires both the identification of visual speech features and the modeling of their changes over time.

However, the choice of architecture also involves computational considerations. Architectures that jointly or sequentially model spatial and temporal information may introduce additional computational requirements. The Temporal Shift Module (TSM) provides an alternative approach by enabling temporal information exchange within a 2D-CNN through channel shifting across neighboring frames. This allows temporal relationships to be modeled while retaining the use of 2D convolution for spatial feature extraction.

In the Philippine context, visual speech recognition research remains limited, particularly for Tagalog. Existing studies have explored different architectures for Filipino and Tagalog visual speech recognition, but there is limited empirical evidence regarding the recognition effectiveness and computational efficiency of a 2D-CNN with TSM for Tagalog visual speech recognition. Thus, this study focuses on the development and evaluation of a 2D Convolutional Neural Network with Temporal Shift Module for Tagalog Visual Speech Recognition, assessing its recognition performance through Accuracy, Precision, Recall, F1-score, MCC, CER, and WER, and its computational efficiency in terms of Training Time and Memory Requirements.

## 1.2 Background of the Study

Visual Speech Recognition (VSR), also known as lip-reading, is the process of recognizing spoken content through visual information, particularly movements of the lips. Human speech is inherently bimodal, consisting of both visual and auditory information, with visual cues contributing to speech perception alongside audio. VSR focuses on these visual cues to determine what is being spoken, making it useful when audio information is unavailable, unclear, or affected by unfavorable acoustic conditions. It can also complement conventional audio-based speech recognition by improving speech recognition and separation in noisy environments. Beyond speech recognition, VSR has potential applications in areas such as medical assistance, human-computer interaction, security, forensics, entertainment, and the transcription of silent archival films (Sheng, Kuang, Bai, Hou, Guo, Xu, Pietikäinen, & Liu, 2024).

Deep learning has contributed substantially to the development of automatic VSR systems. Oghbaie, Sabaghi, Hashemifard, and Akbari (2021) discussed the development of deep-learning approaches for automatic lip-reading and the increasing use of large-scale datasets for learning visual speech representations. CNNs can extract spatial features from individual video frames, while additional temporal modeling is needed to represent changes in mouth and lip movements across consecutive frames. Different architectures have therefore been explored to combine spatial and temporal information, including 2D CNNs, 3D CNNs, recurrent networks, temporal convolutional networks, and attention-based approaches.

The choice of architecture also involves computational considerations. 3D CNNs can process spatial and temporal information jointly but may require greater computational resources, while 2D CNNs primarily process spatial information and therefore require an additional mechanism to model temporal relationships. Lin, Gan, Wang, and Han (2022) described conventional 2D CNNs as computationally inexpensive while noting that they cannot directly capture temporal relationships. This distinction is relevant to the present study because the proposed architecture retains 2D convolution for spatial feature extraction while introducing TSM for temporal modeling.

The Temporal Shift Module (TSM) provides a mechanism for incorporating temporal information into 2D-CNN-based architectures by shifting selected feature channels across neighboring frames. Lin et al. (2022) introduced TSM as a method for exchanging information among neighboring frames through temporal channel shifting. Li, Mamut, Yadikar, Zhu, and Ubul (2021) investigated a channel-enhanced temporal-shift approach for efficient lip-reading. Li et al. (2022) further examined TSM in a 2D ResNet-18-based lip-reading configuration. These studies provide evidence that temporal shifting can be incorporated into CNN-based VSR architectures, while the specific effectiveness of the approach remains dependent on the architecture, dataset, and experimental conditions used.

Computational efficiency is particularly relevant to VSR because video-based models process multiple frames and may require substantial computational and memory resources. Ma, Martinez, Petridis, and Pantic (2021) identified a continuing gap between existing lip-reading methodologies and the requirements for practical deployment. Panagos, Sfikas, and Nikou (2024) likewise investigated approaches for developing resource-

efficient VSR architectures. These findings support the importance of evaluating computational requirements directly rather than considering recognition performance alone.

Language and dataset characteristics also need to be considered when evaluating VSR systems. Studies have investigated visual speech recognition using different languages, vocabularies, speakers, and recording conditions, making results from one dataset difficult to apply directly to another. Ma, Petridis, and Pantic (2022) examined visual speech recognition across multiple languages and in-the-wild conditions, demonstrating the importance of language and dataset diversity in VSR research. Therefore, evidence obtained from foreign datasets cannot by itself establish the recognition effectiveness or computational requirements of a particular architecture for Tagalog visual speech recognition.

In the Philippine context, visual speech recognition research has investigated Filipino, Tagalog, Cebuano, and Ilocano using different deep-learning architectures. Crespo and Recario (2024) investigated Filipino word lip-reading using CNN and GRU configurations. Villarica and Vinluan (2023) investigated a hybrid CNN-LSTM approach for Tagalog lip-reading. Pascual, Ayo, and Paglinawan (2026) developed a Tagalog lip-reading system using a 3D CNN combined with BiLSTM and CTC and reported recognition and computational challenges in their experiments. Cruz, Abadiano, Sangilan, Requillo, and Juanatas (2026) investigated a 2D ResNet-50 with LSTM and CTC for Cebuano and Ilocano. These studies establish that visual speech recognition has been investigated in Philippine-language settings, but they use architectures different from the proposed 2D-CNN + TSM approach.

Taken together, the existing literature establishes the importance of visual information in speech, spatial feature extraction, temporal modeling, computational considerations, and language-specific evaluation in VSR. However, limited empirical evidence is available regarding the recognition effectiveness and computational efficiency of a 2D-CNN with TSM specifically for Tagalog visual speech recognition. This provides the basis for the present study, which focuses on developing and evaluating a 2D-CNN + TSM architecture for Tagalog visual speech recognition using recognition-performance measures and computational-efficiency measures defined in the study.

The present study addresses an evidence gap concerning the limited empirical evidence on the recognition effectiveness and computational efficiency of 2D-CNN + TSM for Tagalog lip-reading. An evidence gap refers to insufficient or conflicting empirical data, where existing evidence is inadequate to support a clear conclusion. In this case, foreign studies have provided evidence supporting the use of 2D-CNN + TSM for visual speech recognition, but comparable evidence within the Philippine context, particularly for Tagalog, remains limited.

Foreign research has demonstrated the potential of 2D-CNN + TSM as a computationally efficient approach to visual speech recognition. Li et al. (2021) incorporated TSM with channel attention to enhance spatiotemporal feature extraction. Li, Yadikar, Zhu, Mamut, & Ubul (2022) investigated TSM with a ResNet-18-based architecture for lip-reading. Hao, Zhou, Li, Zhang, Xie, Wu, and Yin (2025) likewise explored a ResNet-18 and TSM-based approach for enhancing visual speech recognition. These studies provide evidence that TSM can enable 2D-CNN architectures to incorporate temporal information without relying solely on computationally intensive 3D convolution.

However, 2D-CNN + TSM remains underrepresented in Tagalog and Philippine lip-reading research based on the local studies reviewed. Villarica and Vinluan (2023) investigated a CNN-LSTM architecture for Tagalog lip-reading. Crespo and Recario (2024) investigated a 2D ResNet-LSTM approach for Filipino word lip-reading. Cruz, J. P. T., Abadiano, A. B., Sangilan, F. O., Requillo, E. G. T., & Juanatas, R. C. et al. (2026) investigated a 2D ResNet-LSTM approach for Cebuano and Ilocano. Pascual et al. (2026) investigated a 3D-CNN-Bi-LSTM architecture for Tagalog lip-reading. Among these, Pascual et al. (2026) provide a recent Tagalog baseline using 3D-CNN with Bi-LSTM but reported long training times and memory limitations. Consequently, while foreign studies provide evidence of the potential effectiveness and computational efficiency of 2D-CNN + TSM, sufficient local evidence is lacking to determine whether these benefits extend to Tagalog lip-reading.

To address this gap, the present study will develop and evaluate a 2D-CNN + TSM model for Tagalog lip-reading. Recognition performance will be evaluated using Accuracy, Precision, Recall, F1-score, MCC, CER, and WER, while computational efficiency will be assessed through Training Time and Memory Requirements.

### 1.3 Objectives of the Study

- **General Objective:** To develop and evaluate a 2D Convolutional Neural Network with Temporal Shift Module for Tagalog visual speech recognition
- **Specific Objectives:**
  1. To develop an alternative Tagalog visual speech recognition model using a 2D Convolutional Neural Network with Temporal Shift Module.
  2. To integrate the Temporal Shift Module into the 2D Convolutional Neural Network to capture both spatial and temporal features without relying on computationally heavier 3D convolution.
  3. To assess the recognition performance of the proposed model using accuracy, precision, recall, F1-score, Matthews Correlation Coefficient (MCC), Character Error Rate (CER), and Word Error Rate (WER).
  4. To evaluate the computational efficiency of the proposed model in terms of training time and memory requirements.

### 1.4 Scope and Limitation of the Study

- **Scope:**

This study focuses on the development and evaluation of a 2D-CNN with TSM for Tagalog visual speech recognition. The system will recognize Tagalog words from visual information obtained from sequences of mouth or lip movements in videos. Recognition will be evaluated at both the individual-word and word-sequence levels.

The study covers the preparation and preprocessing of Tagalog visual speech data, extraction of spatial features using a 2D-CNN, temporal feature modeling through TSM, sequence recognition using CTC, model training, validation, testing, and evaluation. The 2D-CNN will serve as the spatial feature extraction component, while TSM will serve as the temporal modeling component.

The recognition performance of the proposed system will be evaluated using Accuracy, Precision, Recall, F1-score, MCC, CER, and WER. Computational efficiency will be evaluated using Training Time and Memory Requirements. These measures will be obtained from the actual implementation and computational environment used in the study.

The study will focus on the Tagalog visual speech task defined by the prepared dataset, including its target vocabulary, speakers, and recording characteristics. Foreign benchmark datasets such as LRW and LRW-1000 will not be used as substitutes for the Tagalog dataset. Results from these datasets will instead be considered as related evidence and research context.

The study is delimited to the proposed 2D-CNN + TSM architecture and does not include the development and evaluation of alternative architectures such as 3D CNN, CNN-LSTM, CNN-GRU, TCN, or Transformer-based models as competing systems. Similarly, computational-efficiency evaluation is delimited to Training Time and Memory Requirements. Measures such as FLOPs, parameter count, inference latency, energy consumption, and hardware utilization will not be treated as primary computational-efficiency measures.

- **Limitation:**

The results of the study may be affected by the size, diversity, quality, and other characteristics of the prepared Tagalog visual speech dataset. Variations in the available speakers, vocabulary, video samples, and recording conditions may also affect the recognition results.

The findings may not automatically generalize to other languages, larger vocabularies, different speakers, or recording conditions that differ from those represented in the dataset. Results obtained from foreign VSR datasets cannot automatically be assumed to apply to Tagalog because language and visual speech characteristics may differ.

The computational-efficiency results may also be influenced by the specific hardware and software environment used during the implementation. Therefore, the measured Training Time and Memory Requirements may differ when the same model is implemented or trained under different computational

environments.

The study does not establish that the 2D-CNN + TSM architecture is universally superior to 3D CNN, CNN-LSTM, CNN-GRU, TCN, Transformer, or other visual speech recognition architectures. Such a conclusion would require direct and appropriately controlled comparative experimentation, which is not part of the present study.

The study does not claim to develop a production-ready, commercial, or complete assistive communication system. Its findings are limited to the empirical development and evaluation of the proposed deep-learning architecture for the defined Tagalog visual speech recognition task.

## 1.5 Significance of the Study

The study may provide useful contributions to visual speech recognition research, Philippine language technology research, and future development of deep-learning approaches for visual speech recognition.

### **Deaf, Mute, or Speech-Impaired Individuals (Accessibility Users)**

This group may potentially benefit from future visual speech recognition technologies developed from research such as this study. By investigating visual speech recognition for Tagalog, the study provides a technical foundation that may support future development of assistive communication tools, accessibility applications, and alternative interaction systems for individuals who cannot rely on audio speech.

### **Educators of Deaf, Mute, and Speech-Impaired Individuals**

Educators and special education professionals may potentially benefit from future localized visual speech technologies developed from this study and related research. Such technologies may eventually support supplementary instructional tools and interactive learning activities involving classroom communication, vocabulary training, and speech-reading for Filipino learners.

### **Researchers and Future Studies**

The study may serve as a reference for researchers investigating visual speech recognition, CNN-based architectures, temporal modeling, and language-specific VSR. In particular, it will provide recognition-performance and computational-efficiency results for a 2D-CNN + TSM architecture in a Tagalog visual speech setting.

The empirical results may provide information that future researchers can use when designing studies involving Tagalog or other Philippine languages. Future studies may build upon the findings by investigating different datasets, vocabularies, speakers, architectures, or temporal modeling approaches.

## 1.6 Definition of Terms

The following terms are defined both conceptually and operationally to clarify their meaning and their specific use in the study.

**Accuracy:** Accuracy is a measure of classification performance that represents the proportion of correct predictions among the total predictions (Pascual et al., 2026). In this study, accuracy is used to measure the overall proportion of correctly recognized Tagalog words by the proposed 2D-CNN + TSM model.

**Character Error Rate (CER):** Character Error Rate is a metric used to measure recognition errors at the character level by determining the corrections required to transform the predicted output into the ground truth text (Pascual et al., 2026). In this study, CER is used to measure character-level recognition errors in the output produced by the proposed Tagalog visual speech recognition model.

**Computational Efficiency:** Computational efficiency refers to the ability of a model to perform its intended computational task with reduced computational cost (Lin et al., 2022). In this study, computational efficiency refers to the computational requirements of the proposed 2D-CNN + TSM model, evaluated specifically through training time and memory requirements.

**Connectionist Temporal Classification (CTC):** Connectionist Temporal Classification (CTC) is a sequence-learning approach that enables a model to map input sequences to output sequences without requiring an explicit alignment between individual input frames and target output elements (Pascual et al., 2026). In this

study, CTC is used to train and decode the proposed 2D-CNN + TSM model for recognizing Tagalog word sequences from sequential visual speech features.

**F1-score:** F1-score is a measure calculated from precision and recall using their harmonic mean (Baloch, Ali, Hussain, Daud, & Dawood, 2026). In this study, F1-score is used to evaluate the recognition performance of the proposed model by combining its precision and recall.

**Lip Reading:** Lip reading refers to recognizing speech content through the sequence of lip movements (Li et al., 2021), while it can also involve recognizing words or sentences from the mouth movements of a speaking person (Dweik, Altorman, & Ashour, 2022). In this study, lip reading refers to recognizing Tagalog words from visual information showing the speaker's lip or mouth movements.

**Matthews Correlation Coefficient (MCC):** Matthews Correlation Coefficient (MCC) is a classification-performance metric that provides a balanced assessment of prediction performance, including in the presence of class imbalance (Tamura, Itaya, Hayashi, & Yamamoto, 2026). In this study, MCC is used as a recognition-performance metric to measure the agreement between the predicted and actual Tagalog word labels.

**Memory Requirements:** Memory requirements refer to the memory resources needed by a neural-network model during its operation and training, which are relevant to the resource demands of video-based models (Lin et al., 2022). In this study, memory requirements refer to the peak GPU memory allocated by the proposed 2D-CNN + TSM model during training, and are used as a computational-efficiency measure.

**Precision:** Precision is the proportion of true positive predictions among all positive predictions made by a recognition system (Pascual et al., 2026). In this study, precision is used to measure the proportion of correctly predicted Tagalog words among the words predicted by the proposed visual speech recognition model.

**Recall:** Recall is the proportion of true positive predictions among all actual positive instances (Pascual et al., 2026). In this study, recall is used to measure the proportion of actual Tagalog words that are correctly recognized by the proposed visual speech recognition model.

**Tagalog:** Tagalog is a Malayo-Polynesian Austronesian language from the Philippines (Barrios & Garcia, 2023). In this study, Tagalog refers to the language represented by the visual speech data and target words used for developing and evaluating the proposed visual speech recognition model.

**Temporal Shift Module (TSM):** A Temporal Shift Module (TSM) is a neural-network component that shifts feature information along the temporal dimension to facilitate temporal information exchange in lip-reading models (Li et al., 2021). In this study, TSM functions as the temporal modeling component of the proposed architecture by shifting selected feature information across sequential visual speech frames after spatial features are extracted by the 2D-CNN.

**Training Time:** Training time refers to the time required to train a machine-learning model and is an aspect related to computational efficiency (Lin et al., 2022). In this study, training time refers to the total elapsed time required to complete the training of the proposed 2D-CNN + TSM model and is used as a measure of computational efficiency.

**2D Convolutional Neural Network (2D-CNN):** A 2D Convolutional Neural Network (2D-CNN) is a convolutional neural network commonly used to extract spatial information from individual image frames (Li et al., 2022). In this study, the 2D-CNN functions as the spatial feature extraction component of the proposed architecture by extracting visual features from processed visual speech frames.

**Visual Speech Recognition (VSR):** Visual Speech Recognition is the process of recognizing speech content based on lip movements without relying on the audio stream (Ma et al., 2022). In this study, VSR refers to recognizing Tagalog words from sequential visual speech or lip movements using the proposed 2D-CNN + TSM model.

**Word Error Rate (WER):** Word Error Rate is a metric used to measure recognition errors based on the minimum number of word-level corrections required to align the predicted output with the ground truth (Pascual et al., 2026). In this study, WER is used to measure word-level recognition errors in the Tagalog output generated by the proposed visual speech recognition model.

## SECTION II — LITERATURE REVIEW

### 2.1 Introduction to the Review

This chapter reviews literature and studies related to Visual Speech Recognition (VSR), lip-reading, CNN-based spatial feature extraction, temporal modeling, Temporal Shift Module (TSM), computational efficiency, and language-specific VSR. The review is organized thematically to trace the development of VSR from spatial feature extraction toward spatiotemporal and computationally efficient architectures, while progressively narrowing the discussion from international research to Philippine and Tagalog applications.

The review first establishes the spatial and temporal requirements of VSR, then examines how CNNs, recurrent networks, temporal convolutions, Transformers, and TSM address these requirements. It subsequently focuses on TSM as a mechanism for introducing temporal interaction within CNN-based architectures. The review then considers computational efficiency and the factors that influence VSR performance, including dataset size, vocabulary, speaker variation, visual ambiguity, and language. The discussion then focuses on Philippine and Tagalog VSR to determine what has already been established locally and what remains insufficiently supported by empirical evidence.

The chapter concludes with a synthesis of the reviewed evidence and identifies the specific evidence gap addressed by the present study.

### 2.2 Review of Related Literature and Studies

#### 2.2.1 Foundations of Visual Speech Recognition and Lip-Reading

Visual Speech Recognition (VSR), commonly associated with lip-reading, involves recognizing spoken content from visual information, particularly the movements of the lips and surrounding facial regions. Ma et al. (2022) described VSR as recognizing speech content based on lip movements “without relying on the audio stream.” Their study investigated VSR across multiple languages in-the-wild, demonstrating the importance of model design, hyperparameter optimization, and data augmentation in improving recognition performance. This establishes the fundamental premise of the present study, which focuses on recognizing Tagalog speech from visual information alone.

The development of deep learning has contributed substantially to the advancement of automatic lip-reading. Oghbaie et al. (2021) stated that “Driven by deep learning techniques and large-scale datasets, recent years have witnessed a paradigm shift in automatic lip reading” (p. 1). This development provides a basis for using neural-network-based architectures in VSR rather than relying solely on manually designed visual features. Similarly, Sheng et al. (2024) reviewed the development of deep-learning methods for visual speech analysis. This supports the use of deep learning for extracting meaningful visual representations from sequences of mouth and lip movements. In the present study, this role is primarily performed by the 2D-CNN, which extracts spatial features from individual video frames.

“The huge improvement of neural networks in the visual task has resulted in a robust and reliable lip reading framework that can recognize speech from visual input alone” (Haq, Ruan, Cai, & Li, 2022, p. 53481). This supports the feasibility of recognizing speech content from visual input without relying on audio. “Lipreading involves recognizing spoken words by analyzing the movements of the lips and surrounding area using visual data” (Dweik et al., 2022, p. 1). This provides a basis for treating visible mouth and lip movements as the primary visual information processed by the proposed spatial feature-extraction component. The practical applications of visual speech technologies have also expanded beyond recognition itself. Innocente, Boemio, Lorenzetti, Pulito, Romagnoli, Saponaro, Marullo, Ulrich, and Vezzetti (2025) stated that “lip-reading technology, based on visual speech decoding and automatic speech recognition, offers a promising solution to overcoming communication barriers” (p. 1355). This illustrates the potential application of sequential visual-speech models in rehabilitation and communication-support contexts.

Language-specific challenges are also evident in the literature. Baloch et al. (2026) reported that there has been “limited success in developing an effective model for Urdu lip reading” (p. 1) because of dataset limitations and challenges encountered by earlier models. This directly supports the identified research gap, emphasizing that architectures proven on English datasets must be empirically re-evaluated under the specific linguistic

constraints of Tagalog.

Overall, these studies establish that VSR involves extracting visual information from speech-related movements and modeling that information to recognize spoken content. They also demonstrate the importance of deep learning, adequate datasets, and language-specific evaluation. These findings provide the foundation for examining CNN-based spatial extraction and temporal modeling in the succeeding sections.

### **2.2.2 CNN-Based Spatial Feature Extraction**

CNNs have been widely applied to visual speech recognition because of their ability to extract spatial representations from visual inputs. Bardhan, Singh, and Attri (2024) proposed a CNN-based model for detecting words from lip movements and converting them into text, demonstrating the applicability of CNN-based processing to visual word recognition. This provides a basis for using a 2D-CNN as the spatial feature-extraction component of the proposed architecture. However, recognition performance must also be considered alongside computational requirements. Ma et al. (2021) noted that “there is still a significant gap between the current methodologies and the requirements for an effective deployment of lipreading in practical scenarios.” This highlights the importance of considering computational efficiency when developing VSR systems intended for practical use. Their work demonstrates the potential of utilizing lightweight CNN architectures to preserve useful spatial representations while reducing computational requirements. In the present study, this concern is addressed by investigating an architecture based on 2D convolution and temporal shifting. Fu, Lu, and Ni (2023) emphasized this consideration by describing ShuffleNet as an “extremely computation-efficient Convolutional Neural Network (CNN) architecture” (p. 1). Their work demonstrates the continuing development of CNN architectures that attempt to maintain useful visual representations while reducing computational requirements. This supports the broader motivation for investigating computationally efficient CNN-based VSR architectures and underscores the importance of balancing spatial representation accuracy with architectural efficiency in modern lip-reading systems.

Spatial feature extraction alone, however, does not fully represent the dynamic nature of speech. Li et al. (2021) proposed “a channel enhanced temporal-shift module for efficient lipreading,” illustrating the use of temporal shifting alongside CNN-based processing. The approach is relevant to the present study because it demonstrates how temporal information can be incorporated into a CNN-based lip-reading architecture. Li et al. (2022) emphasized the need to capture both spatial and temporal characteristics, stating that researchers have focused on “extracting fine-grained spatial features while also obtaining temporal information of movement sequences” (p. 1). This distinction is particularly important in VSR because individual frames provide spatial information, while changes between consecutive frames provide information about speech movement.

More recent work has likewise emphasized the importance of combining spatial and temporal representations. Ma and Sun (2025) noted that “spatiotemporal feature enhancement mechanisms have recently gained attention in the field of lip-reading” (p. 1). This reinforces the principle that the proposed 2D-CNN, which processes frames spatially, must be paired with an additional module to model how those features change across a given time sequence.

Thus, the literature indicates that 2D-CNNs provide a suitable basis for extracting spatial features from individual frames, but additional temporal modeling is necessary to represent changes in mouth and lip movements over time. This provides the rationale for examining temporal modeling approaches in VSR.

### **2.2.3 Temporal Modeling in Visual Speech Recognition**

Temporal modeling is necessary because speech is expressed through sequences of movements rather than isolated frames. Various approaches have therefore been investigated to capture temporal relationships in visual speech.

Li, Hashim, Lin, Nohuddin, Venkatachalam, and Ahmadian (2024) discussed the continued development of AI-based visual speech recognition for applications such as realistic avatars and lip-reading. Their work reflects the broader expansion of VSR toward applications that require the processing of sequential visual information. Wang, Pu, and Chen (2022) proposed a “3D convolutional vision transformer (3DCvT), which combines vision transformer and 3D convolution to extract the spatio-temporal feature of continuous images” (p. 77205). This demonstrates how both convolution and attention-based mechanisms can be combined to capture spatiotemporal information. However, architectures that simultaneously process spatial and temporal information may introduce additional computational requirements. This consideration is relevant to the present

study because the proposed approach seeks to model temporal relationships while retaining the structure of 2D convolution. Temporal variation may also differ between speakers. Yeo, Kim, Kim, Rha, Han, Cheng, and Ro (2025) reported that “performance degrades when models are applied to unseen speakers due to their sensitivity to variations in visual information such as lip appearances” (p. 9472). This finding indicates that speaker-specific visual characteristics can affect lip-reading performance and provides a basis for considering speaker variation when evaluating VSR systems. In the context of the present study, this supports evaluating the proposed model using data that represent the speaker conditions included in the Tagalog dataset.

Overall, the literature demonstrates that temporal modeling can be implemented through different approaches, including convolutional, recurrent, and attention-based mechanisms. These approaches differ in their computational requirements and architectural complexity, motivating further investigation into efficient alternatives such as TSM.

#### **2.2.4 Temporal Shift Module and Efficient Spatiotemporal Modeling**

Lin et al. (2022) proposed the Temporal Shift Module (TSM) to address the limitation of conventional 2D CNNs in capturing temporal relationships. They described conventional 2D CNNs as computationally cheap but unable to capture temporal relationships, while 3D CNN-based methods can model spatiotemporal information but are computationally intensive. The key idea of TSM is “to shift part of the channels along the temporal dimension, thus facilitating information exchanged among neighboring frames” (Lin et al., 2022). TSM can therefore introduce temporal interaction into 2D CNNs without additional computation or parameters. This provides the architectural basis for examining a 2D-CNN + TSM approach in the present study.

TSM has also been directly investigated in lip-reading. Li et al. (2021) proposed a channel-enhanced temporal-shift approach for efficient lip-reading and evaluated it on LRW and LRW-1000. Li et al. (2022) further investigated TSM for word-level lip-reading and reported improved performance when TSM was added to a 2D ResNet-18 configuration. Li et al. (2022) reported LRW ablation results of 80.34% for the 2D ResNet-18 configuration, 83.83% after adding TSM, and 84.13% after adding SE-TSM. These results provide evidence that TSM can improve the performance of the particular 2D-CNN configuration investigated in that study. However, the complete architecture included additional components, so the findings should not be interpreted as the expected performance of a simple 2D-CNN + TSM model.

Hao et al. (2025) proposed LipGen, described as “a viseme-guided lip video generation framework for enhancing visual speech recognition.” They incorporated TSM into a ResNet-18-based visual speech recognition architecture, demonstrating that temporal channel shifting can be used within a 2D-CNN-based framework for spatiotemporal feature extraction. However, their approach also incorporated synthetic data generation, viseme-guided supervision, and a DC-TCN back-end, so its reported performance should not be attributed to TSM alone. The study therefore provides supporting evidence for the use of TSM in CNN-based VSR while also demonstrating that results depend on the complete architecture and experimental configuration.

Taken together, these studies establish the relevance of temporal shifting and other spatiotemporal mechanisms to visual speech recognition. TSM is particularly relevant to the present study because it provides a method for introducing temporal interaction while retaining 2D convolution as the primary spatial feature-extraction mechanism. Nevertheless, existing results are dependent on the datasets, architectures, and experimental conditions used in each study, making language-specific empirical evaluation necessary.

#### **2.2.5 Computational Efficiency and Lightweight Visual Speech Recognition**

Computational efficiency is an important consideration in the development and deployment of VSR systems. Panagos et al. (2024) observed that parameterized hypercomplex layers can reduce model size and parameter requirements while maintaining competitive recognition performance. However, they also reported that these layers caused a slight increase in memory demands during training because current hardware and implementations are optimized for conventional convolutions and matrix multiplications. This demonstrates that computational efficiency involves multiple aspects of computational cost and that reductions in model size do not necessarily translate directly into lower training-memory requirements. Wang et al. (2022) concluded that while advanced vision architectures raise recognition rates, they simultaneously trigger steep increases in computational cost. This highlights the practical need to directly measure the computational burden (training time and memory) introduced by complex temporal modeling within the experimental bounds of the study.

Local research further demonstrates the relevance of this issue. Pascual et al. (2026) reported memory

limitations while training a Tagalog lip-reading system based on a 3D convolutional neural network with bidirectional long short-term memory. This finding provides local evidence that the computational requirements of VSR architectures can become a practical constraint. It therefore supports the need to empirically evaluate the computational behavior of a 2D-CNN + TSM architecture under controlled experimental conditions.

#### **2.2.6 Language-Specific, Low-Resource, and Multilingual Visual Speech Recognition**

The expansion of VSR beyond English benchmarks has established language diversity as an important consideration in visual speech research. Studies involving Mandarin, Arabic, Italian, Urdu, Filipino, Tagalog, Cebuano, and Ilocano demonstrate that VSR can be adapted to different linguistic settings. Multilingual and in-the-wild research further demonstrates that language and environmental conditions significantly affect the visual speech recognition task (Ma et al., 2022).

These differences make cross-language performance comparisons difficult. Vocabulary size, number of speakers, recording conditions, dataset scale, and visual characteristics vary substantially among studies. For example, experiments have ranged from 10 Arabic words (Dweik et al., 2022) to 40 Urdu alphabet classes (Baloch et al., 2026) and 74 Tagalog words (Pascual et al., 2026). Accuracy obtained under one vocabulary and dataset structure therefore cannot be interpreted as directly equivalent to accuracy obtained under another. Low-resource settings present an additional concern because limited training data can restrict the visual patterns available for learning. Speaker and environmental variation also influence the generalizability of VSR, with personalized systems demonstrating the importance of adapting to individual lip-movement characteristics (Yeo et al., 2025). Thus, VSR performance reflects an interaction between the architecture and the visual speech conditions represented in the data, strongly supporting the need for language-specific empirical evaluation.

#### **2.2.7 Philippine, Filipino, and Tagalog Visual Speech Recognition**

The Philippine context provides more specific evidence concerning local visual speech recognition. "Apart from the low volume of training data, the features extracted by the CNN were not sufficient to predict the word being spoken" (Crespo & Recario, 2024). Their finding highlights the combined challenges of limited training data and visual feature representation in Filipino lip-reading. Specifically, they investigated Filipino word lip-reading using CNN and GRU configurations, with the best reported configuration achieving 17% test accuracy and 26% validation accuracy. These results indicate that Filipino visual speech recognition presents challenges involving data availability and feature representation and that spatial feature extraction must be considered together with temporal modeling.

Pascual et al. (2026) provide the closest comparison to the present study because they specifically investigated Tagalog visual speech using a 74-word vocabulary and a 3D CNN + BiLSTM + CTC architecture. Their reported results were 76.27% accuracy, 19.28% precision, 15.02% recall, 16.68% F1-score, 10.09% CER, and 24.08% WER. The study also reported approximately four hours of training per epoch and memory limitations that prevented the final three epochs from being recorded. In addition, class imbalance and visually similar visemes were identified as recognition challenges.

Cruz et al. (2026) provide evidence that a 2D ResNet-50 combined with LSTM and CTC can be applied to lip-reading in Philippine regional languages such as Cebuano and Ilocano and deployed on an edge device. This provides local evidence for the use of 2D CNN-based spatial processing in Philippine-language VSR. They investigated Philippine regional-language visual speech recognition using a 2D ResNet-50 with TimeDistributed processing, LSTM, and CTC for Cebuano and Ilocano. The study reported 74% Cebuano and 66% Ilocano lexicon accuracy and approximately 3–7 seconds of inference time on a Jetson Orin Nano. Although this demonstrates the use of 2D-CNN-based processing for Philippine-language recognition, it differs from the present study in language, task, temporal mechanism, and computational measure. Its inference-time results therefore cannot be treated as evidence of the training time or memory requirements of the proposed Tagalog VSR model.

These local findings connect the architectural and computational themes discussed earlier. Existing Philippine research has primarily combined CNN-based spatial processing with recurrent temporal models, while Pascual et al. (2026) demonstrate that a more computationally demanding 3D CNN-BiLSTM configuration can encounter practical training and memory limitations. The literature therefore provides a basis for investigating whether a different spatial-temporal combination can be empirically characterized in the same

broader Tagalog context.

Importantly, the local evidence does not indicate an absence of Tagalog VSR research. Instead, it shows that the specific combination of 2D CNN and TSM remains insufficiently investigated in the Philippine/Tagalog setting. The present study consequently extends existing local work rather than introducing VSR to a completely unexplored language context.

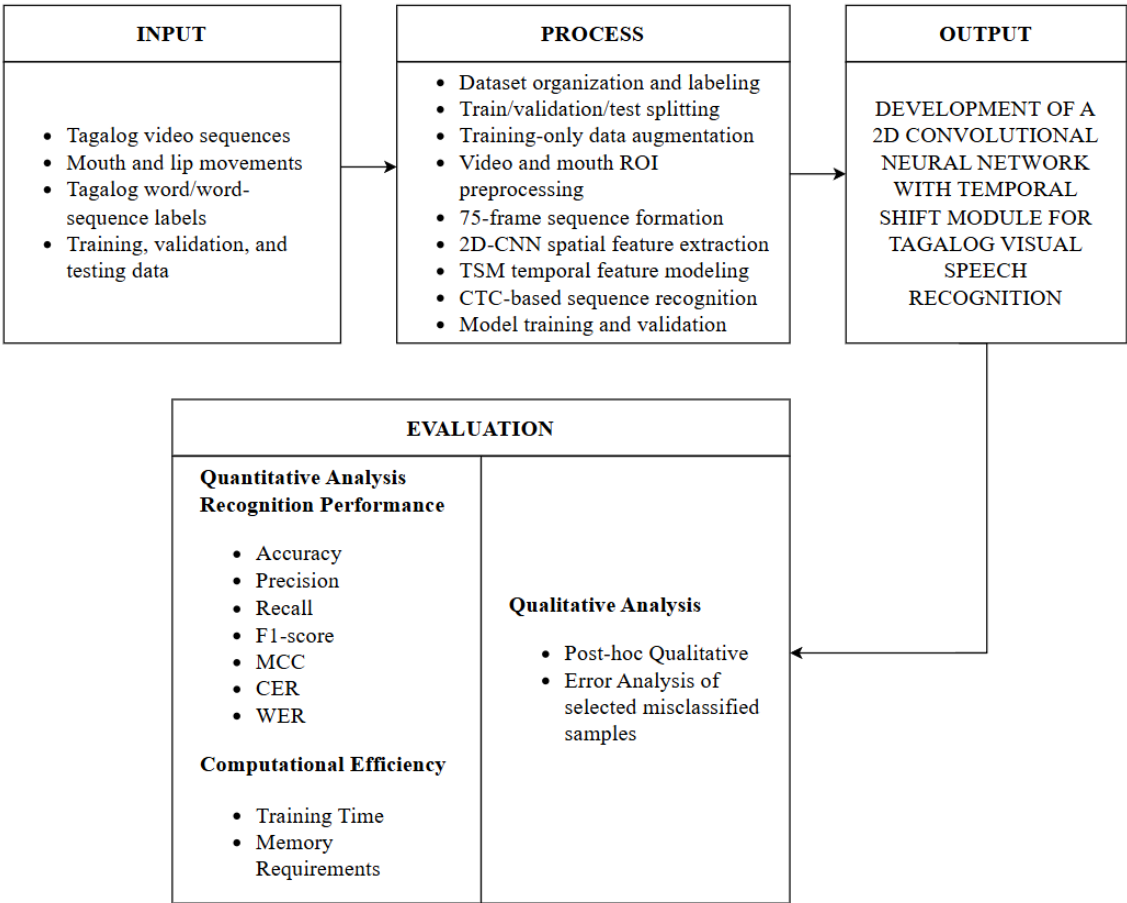
2.2.8 Recognition Challenges, Applications, and Research Trends

Visual ambiguity is a particularly persistent recognition challenge in VSR. Pascual et al. (2026) identified similar visemes in Tagalog. Innocente et al. (2025) reported difficulties with visually similar Italian words. Such cases illustrate a fundamental limitation of visual-only speech recognition: different speech units can produce similar visible mouth configurations. Consequently, model evaluation should consider not only aggregate accuracy but also the nature of recognition errors.

The literature also reflects a continuing movement toward practical and application-oriented VSR. While existing systems demonstrate potential for rehabilitation and automated word detection, research continues to heavily explore efficient architectures, including compact CNNs, parameter-efficient models, and TSM-based approaches. Taken together, these developments reveal three recurring directions in VSR research: improved spatiotemporal representation, greater computational awareness, and expansion to language-specific and multilingual settings. The reviewed studies do not establish a universally superior architecture; instead, they provide different solutions to the common requirements of representing visual speech across time while managing data and computational constraints.

2.3 Conceptual Framework

Figure 1  
Input–Process–Output Model of the Proposed 2D-CNN + TSM Tagalog VSR System



*Note.* The framework illustrates the processing flow from the Tagalog video sequence through training-only augmentation, preprocessing, 75-frame sequence formation, 2D-CNN spatial feature extraction, TSM temporal feature modeling, and CTC-based sequence recognition to produce the predicted Tagalog word or word sequence.

Figure 1 shows the Input–Process–Output model of the study and the evaluation of the proposed Tagalog visual speech recognition system. The framework illustrates how the visual speech data are prepared and processed through the proposed 2D-CNN + TSM architecture to produce the recognition model, which is then evaluated based on its recognition performance, computational efficiency, and selected recognition errors.

For input, it consists of Tagalog visual speech data, including video sequences containing mouth and lip movements, corresponding Tagalog word or word-sequence labels, and the prepared training, validation, and testing data. These inputs provide the visual information and target labels needed to develop the proposed visual speech recognition model.

For process, it involves dataset organization and labeling, train/validation/test splitting, training-only data augmentation, and video and mouth-region preprocessing. The processed videos are formed into 75-frame sequences and passed through the 2D-CNN for spatial feature extraction. The extracted features are then processed by the TSM for temporal feature modeling, followed by CTC-based sequence recognition to generate predicted Tagalog words or word sequences. The model is trained using the training data and validated using the validation data before final testing.

For output, it is the developed 2D-CNN + TSM-based Tagalog visual speech recognition model, which generates predicted Tagalog words or word sequences from visual speech input.

For evaluation, the developed model is assessed using recognition-performance measures, including Accuracy, Precision, Recall, F1-score, MCC, CER, and WER, and computational-efficiency measures, including Training Time and Memory Requirements. Selected misclassified samples are also examined through post-hoc qualitative error analysis to identify observable visual patterns that may help explain recognition errors.

## 2.4 Synthesis

**Evidence Gap.** The reviewed literature provides limited empirical evidence regarding the recognition effectiveness and computational efficiency of a 2D-CNN with Temporal Shift Module for Tagalog visual speech recognition. Although international studies demonstrate the application of TSM and other temporal mechanisms in VSR, and Philippine studies establish the feasibility of local-language VSR, comparable evidence specifically combining 2D-CNN + TSM for Tagalog remains limited.

The reviewed studies establish that VSR requires both spatial and temporal information. CNNs provide spatial representations of mouth and lip movements, while temporal mechanisms such as LSTM, GRU, TCN, Transformer, and TSM capture information across frames. This supports the use of a 2D-CNN with a complementary temporal mechanism rather than relying on frame-level spatial processing alone.

Among these approaches, TSM is particularly relevant to the identified gap because it introduces temporal interaction through feature shifting while retaining 2D convolution for spatial processing (Lin et al., 2022). Hao et al. (2025), although using a different approach, further illustrates the continuing importance of effective visual and temporal representation in VSR. However, these studies use different datasets and architectures, so their findings do not establish that TSM is universally superior to other temporal approaches.

The literature also shows that computational efficiency must be considered alongside recognition performance. Different measures can capture different aspects of computational cost, but the present study is specifically concerned with Training Time and Memory Requirements. This is particularly relevant to the local context because Pascual et al. (2026) reported approximately four hours of training per epoch and memory limitations in a Tagalog 3D CNN-BiLSTM system. These findings support the need to measure computational behavior directly rather than infer it from architecture or model size alone.

Philippine studies further establish that VSR and related visual-language recognition tasks are feasible for Tagalog and other Philippine languages using CNN-GRU, CNN-LSTM, 2D ResNet-LSTM, and 3D CNN-BiLSTM architectures. However, these studies do not sufficiently characterize a 2D-CNN + TSM architecture for Tagalog. Pascual et al. (2026) provide a particularly relevant local comparison through their Tagalog VSR system using a

74-word vocabulary and CTC-based sequence recognition, but its 3D CNN-BiLSTM architecture differs from the proposed approach. Thus, the existing literature provides the foundation for the proposed architecture while leaving its effectiveness and computational requirements in the Tagalog context insufficiently established.

The present study addresses this evidence gap by developing and evaluating a 2D-CNN + TSM architecture for Tagalog visual speech recognition. Recognition effectiveness will be assessed using Accuracy, Precision, Recall, F1-score, MCC, CER, and WER, while computational efficiency will be assessed using Training Time and Memory Requirements. The study does not assume that the proposed architecture will outperform other approaches. Instead, it aims to provide empirical evidence concerning its recognition performance and computational requirements within the Tagalog context.

## SECTION III — METHODOLOGY

### 3.1 Research Design

The study will use an Explanatory Sequential Mixed-Methods Research Design (QUAN → qual) with developmental and experimental components. It is predominantly quantitative because its primary purpose is to develop and empirically evaluate a 2D-CNN with TSM for Tagalog visual speech recognition. The quantitative phase will cover dataset preparation, video augmentation, preprocessing, model development, training, validation, testing, and quantitative evaluation, while the qualitative phase will provide a post-hoc explanation of selected recognition errors.

Recognition performance will be assessed using Accuracy, Precision, Recall, F1-score, MCC, CER, and WER, while computational efficiency will be assessed using Training Time and Memory Requirements. The qualitative phase will examine selected misclassified samples, particularly cases involving visually similar words or viseme patterns, based only on observable mouth and lip movements. It will not involve interviews, surveys, focus groups, expert panels, or additional participants.

The study will follow the sequence QUAN → qual, wherein quantitative testing will first establish recognition and computational results, selected errors will then be examined qualitatively, and the findings will subsequently be integrated. The proposed architecture will use the 2D-CNN for spatial feature extraction and TSM for temporal modeling, based on the mechanisms described in previous VSR and lip-reading studies. The study will not assume that 2D-CNN + TSM is superior to 3D CNN, CNN-LSTM, CNN-GRU, TCN, Transformer, or other approaches; instead, it will provide empirical evidence concerning its recognition performance and computational requirements for Tagalog visual speech recognition.

### 3.2 System Development Methodology

The study will follow a machine-learning development workflow consisting of the following sequential phases: first, dataset preparation; second, train/validation/test splitting; third, training-set data augmentation; fourth, video and visual data preprocessing; fifth, model development; sixth, model training; seventh, validation and iteration; eighth, testing; ninth, quantitative evaluation; tenth, qualitative error analysis; and finally, integration of findings. Dataset preparation will organize and label the Tagalog visual speech data and establish the training, validation, and testing partitions. Training-only augmentation will then be performed before mouth ROI extraction, followed by frame extraction, face detection, mouth landmark localization, ROI cropping, grayscale conversion, resizing or padding, normalization, and sequence formation. The resulting sequences will be processed by the proposed 2D-CNN + TSM model, trained and validated, and finally tested using unseen data. Quantitative results will measure recognition performance and computational requirements, after which selected errors will undergo qualitative examination and integration.

Unless otherwise stated, numerical implementation parameters not explicitly specified by cited studies will be treated as researcher-defined parameters and finalized during implementation according to the dataset, model behavior, validation results, and available computational resources.

### 3.3 Requirement Specification and Analysis

#### 3.3.1 Model and Training Requirements

The proposed model will consist of a 2D-CNN + TSM architecture. The 2D-CNN will extract spatial features from individual mouth-region frames, while TSM will introduce temporal information by shifting selected feature channels across neighboring frames. The recognition task uses a fixed vocabulary of Tagalog words that may occur within word sequences; therefore, the model will use a CTC-based output and CTC loss to learn the temporal alignment between the visual feature sequence and the corresponding Tagalog word sequence. LSTM, Bi-LSTM, GRU, TCN, Transformer, CBAM, SE, and other alternative components or architectures will not be required in the proposed model and will only be discussed as related approaches in the literature.

#### 3.3.2 Computational and Software Requirements

The model will be trained and evaluated using an NVIDIA GeForce RTX 3060 with 12 GB VRAM. The 12 GB represents available hardware capacity, not the model's measured memory requirement; the experiment will instead record actual peak GPU memory allocated during training. Software must support video decoding, frame extraction, augmentation, face and landmark detection, ROI processing, sequence preparation, 2D-CNN and TSM implementation, training and validation, checkpoint management, prediction generation, metric calculation, and GPU time and memory monitoring. The exact software frameworks, libraries, and versions will be documented according to the actual implementation.

#### 3.3.3 Input and Output Requirements

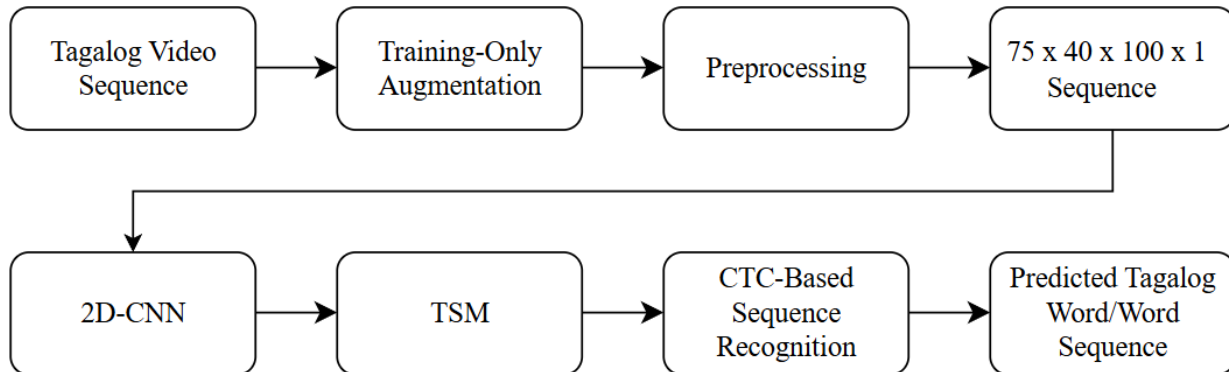
The conceptual input tensor will be represented as  $B \times T \times H \times W \times C$ , where  $B$  is batch size,  $T = 75$ ,  $H = 40$ ,  $W = 100$ , and  $C = 1$ . The model will output a temporal sequence of probability distributions over the CTC vocabulary, including the blank symbol. CTC decoding will convert these temporal predictions into the predicted Tagalog word or word sequence. The system will produce predicted Tagalog words or word sequences, which will then be used for recognition-performance evaluation, computational-efficiency measurement, and selection of misclassified samples for qualitative analysis.

### 3.4 System Design Specification

The proposed system will follow the architecture shown in Figure 2.

**Figure 2**

*Proposed 2D-CNN + TSM Architecture for Tagalog Visual Speech Recognition*



*Note.* The architecture illustrates the processing flow from the Tagalog video sequence through training-only augmentation, preprocessing, sequence formation, 2D-CNN spatial feature extraction, TSM temporal modeling, and CTC-based sequence recognition to produce the predicted Tagalog word or word sequence.

The augmentation stage will only apply to training data. After preprocessing, the sequence will be represented conceptually as  $B \times T \times H \times W \times C$ , with 75 grayscale frames of  $40 \times 100$  pixels. The 2D-CNN will process each frame spatially and learn features such as mouth shape, lip position, mouth opening, and other visible configurations. The same spatial feature extractor will be applied across the sequence, while temporal relationships will be handled by TSM.

Following 2D-CNN feature extraction, the features will be organized as  $B \times T \times C' \times H' \times W'$ . TSM will use a bidirectional partial shift in which 1/8 of the channels are shifted toward the previous frame, 1/8 toward the next frame, and the remaining channels remain unchanged. This specific channel allocation will be treated as a researcher-defined implementation choice based on the temporal-shift mechanism described by Lin et al. (2022). Li et al. (2021) provide supporting evidence for the use of temporal shifting in lip-reading. At the beginning and end of a sequence, positions without corresponding neighboring frames will be handled using zero-padding as defined in the proposed implementation.

The resulting temporally enhanced features will be passed to the CTC-based sequence recognition stage. The CTC output will produce predictions across the temporal dimension and learn the alignment between the visual feature sequence and the corresponding Tagalog word sequence without requiring frame-level alignment. CTC decoding will then generate the predicted Tagalog word sequence. No LSTM, GRU, TCN, Transformer, or 3D convolution will be added after TSM.

The proposed architecture differs from previous Philippine and regional-language approaches. Pascual et al. (2026) used 3D CNN + Bi-LSTM + CTC for Tagalog word recognition, Crespo and Recario (2024) used CNN-GRU configurations for Filipino words, and Cruz et al. (2026) used 2D ResNet-50 + LSTM + CTC for Cebuano and Ilocano. The present study instead evaluates 2D-CNN  $\rightarrow$  TSM  $\rightarrow$  CTC-based sequence recognition, establishing a methodological distinction without assuming superiority.

## 3.5 Implementation

### 3.5.1 Dataset and Data Preparation Requirements

The system requires a labeled Tagalog visual speech dataset in which each video contains visual speech corresponding to a sequence of three target Tagalog words. The dataset conditions will follow applicable characteristics reported by Pascual et al. (2026), including a 74-word Tagalog vocabulary, three words per line, 25 frames per second, and 75 frames per video. Each sample will retain its three-word sequence label and identifiable visual mouth or lip information.

Following Innocente et al. (2025), the dataset will be randomly divided into 90% development and 10% testing, with the development portion further divided into 90% training and 10% validation, resulting in an overall allocation of approximately 81% training, 9% validation, and 10% testing. The test partition will remain unseen during model development, parameter adjustment, and model selection. Video identifiers and available recording metadata will be retained.

### 3.5.2 Data Augmentation

Data augmentation will be performed only on the training partition and before mouth ROI extraction. Following Pascual et al. (2026), each training video will produce five additional variants through horizontal flipping, increased brightness, decreased brightness, increased contrast, and decreased contrast. Innocente et al. (2025) provide additional support for visual augmentation involving horizontal flipping and brightness/contrast modifications. Because Pascual et al. does not specify exact brightness and contrast factors, the study will use  $\pm 15\%$  as researcher-defined moderate adjustments. Transformations will be applied consistently across all frames of a video, preserving chronological order and avoiding frame insertion, deletion, shuffling, or reordering. Validation and test data will remain unaugmented.

### 3.5.3 Video and Preprocessing

The source video resolution reported by Pascual et al. (2026) is  $360 \times 288$  pixels, while the final CNN input will consist of grayscale mouth ROIs standardized to  $40 \times 100$  pixels. Each approximately three-second, 25-fps video will be converted into a chronological sequence of approximately 75 frames. Preprocessing will consist of frame extraction, face detection using Dlib, mouth landmark localization using landmarks 48–67, mouth ROI cropping, grayscale conversion, resizing or padding to  $40 \times 100$  pixels, normalization to  $[0,1]$ , chronological sequence formation, and association with the corresponding word label. Normalization will use pixel division by 255 because the cited Tagalog procedure specifies normalization without providing a more specific range. The resulting model input will therefore be  $75 \times 40 \times 100 \times 1$ .

This procedure is primarily based on Pascual et al. (2026). Depth-specific processing will not be included

because the proposed preprocessing pipeline converts the source video frames to grayscale before model input.

#### **3.5.4 2D-CNN and TSM Implementation**

The 2D-CNN will process each  $40 \times 100 \times 1$  frame using the same spatial feature extractor across the sequence. The resulting features will be reorganized as  $B \times 75 \times C' \times H' \times W'$  and passed to TSM. The temporal shift will move 1/8 of the channels toward the previous frame, 1/8 toward the next frame, and retain the remaining channels in place, allowing neighboring frames to exchange information without introducing a recurrent layer or 3D convolution. The temporally enhanced features will then be passed across the temporal dimension to the CTC-based output stage for sequence recognition.

#### **3.5.5 Model Training and Validation**

Training batches will pass through the 2D-CNN, TSM, and CTC-based output stage, after which CTC loss will be calculated between the predicted temporal output and the corresponding Tagalog word sequence. Model parameters will then be updated using the selected optimizer. The optimizer, learning rate, batch size, and number of epochs will be researcher-defined and finalized according to the dataset, model behavior, validation results, and available 12 GB GPU capacity. Total training time and peak GPU memory allocated will be recorded during training.

Validation data will remain unaugmented and will use the same preprocessing procedure. Validation loss and recognition behavior will be monitored for model learning and potential overfitting, and the final checkpoint will be selected using validation results. The test set will not be used for checkpoint selection or parameter adjustment.

#### **3.5.6 Testing and Model Output**

The selected model checkpoint will be loaded and applied to the unseen test sequences. Each sequence will pass through 2D-CNN  $\rightarrow$  TSM  $\rightarrow$  CTC-based sequence recognition, after which CTC decoding will generate the predicted Tagalog word or word sequence. Prediction-reference pairs will be stored for quantitative evaluation. The final model will be saved and loaded separately for testing to ensure that test results correspond to the selected checkpoint.

#### **3.5.7 Qualitative Error Analysis**

Selected misclassified samples will be examined after quantitative testing. The analysis will record the actual word, predicted word, visual sequence, observable error pattern, and possible visual explanation. The examination will focus on mouth shapes, mouth opening, lip positions, movement transitions, repeated configurations, and visual similarities between actual and predicted words. Particular attention will be given to visually similar viseme patterns, which Pascual et al. (2026) identify as a challenge in Tagalog lip-reading. The analysis will remain descriptive and will not make unsupported claims about pronunciation, speaker intention, or linguistic causes.

#### **3.5.8 Computational Environment**

The implementation will use an NVIDIA GeForce RTX 3060 with 12 GB VRAM, with the same hardware environment used for computational measurements. The 12 GB figure will be reported as available capacity, while peak memory allocation will be obtained experimentally. Actual software frameworks, libraries, versions, and other environment details will be documented from the implementation.

### **3.6 Testing and Evaluation**

The evaluation will address the study's recognition-performance and computational-efficiency objectives, followed by post-hoc qualitative analysis. Recognition performance will be measured using Accuracy, Precision, Recall, F1-score, MCC, CER, and WER. Accuracy will represent the proportion of correctly recognized Tagalog words, while precision and recall will measure the correctness and completeness of class predictions. F1-score will combine precision and recall using  $F1 = 2PR/(P + R)$ , and MCC will provide an additional overall classification-quality measure based on the multiclass prediction results.

CER will measure character-level errors using  $CER = (S + D + I)/N$ , where S represents substitutions, D deletions, I insertions, and N the number of reference characters. WER will measure word-level errors between

the predicted and reference Tagalog word sequences. Because the model uses CTC-based sequence recognition, the evaluation can account for substitutions, deletions, and insertions within the predicted word sequence.

Computational efficiency will be assessed using Training Time and Memory Requirements. Training time will be measured as the actual elapsed duration of the model-training run, with total training time as the primary reported value and per-epoch timing optionally recorded. Preprocessing and augmentation will not form part of the primary training-time measurement unless separately reported. Memory requirements will be represented by peak GPU memory allocated during training, measured using the RTX 3060 environment. The 12 GB VRAM capacity will not be reported as the experimental memory result.

Final testing will begin by loading the selected model and unseen test data without augmentation. The test videos will undergo frame extraction, face detection, Dlib mouth landmark localization, ROI cropping, grayscale conversion, resizing/padding, normalization, and 75-frame sequence formation. The resulting sequences will pass through the 2D-CNN, TSM, and CTC-based sequence recognition stage. CTC decoding will generate the predicted Tagalog word or word sequence. Predictions will be compared with ground-truth labels, prediction-reference pairs will be stored, and the seven recognition metrics will be calculated. Representative misclassified samples will then be selected for qualitative examination.

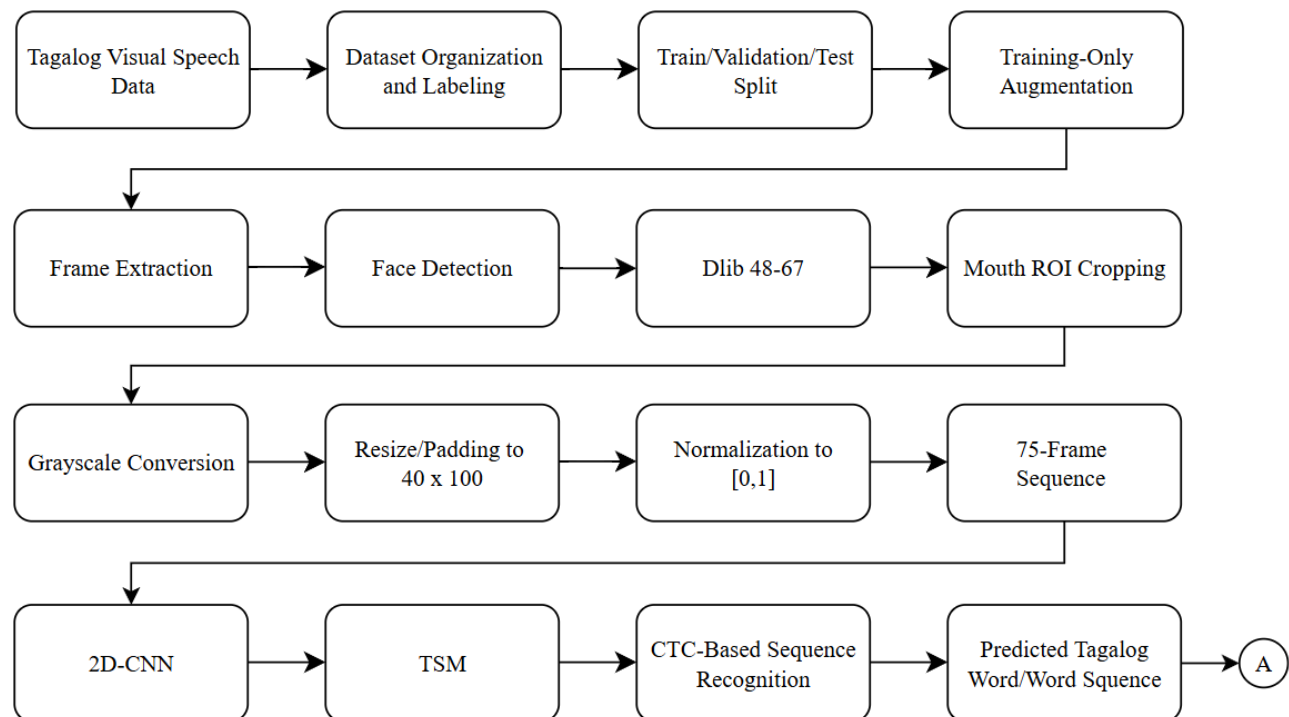
The qualitative analysis will examine selected 75-frame sequences frame by frame or as complete sequences, focusing on observable similarities in mouth shape, mouth opening, lip position, movement transitions, and other visual configurations. Recurring error patterns, particularly those involving visually similar words, will be described as possible visual explanations for the quantitative results. The qualitative analysis will not modify the model after final testing and will not replace the numerical evaluation.

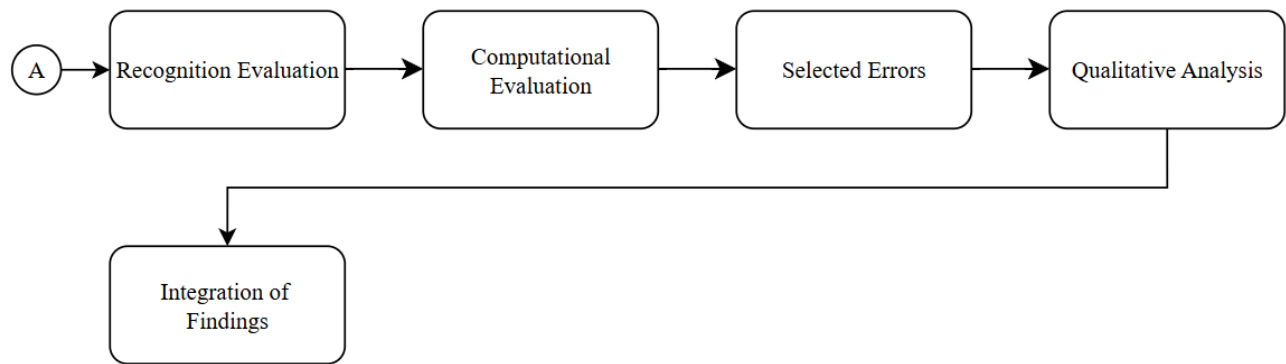
The quantitative and qualitative findings will then be integrated. Quantitative results will establish recognition performance and computational requirements, while qualitative observations will help explain selected recognition errors. The integration will therefore follow quantitative results → identification of notable errors → selection of representative samples → qualitative visual examination → identification of recurring visual patterns → integrated interpretation. No predetermined claim of success, efficiency, or superiority will be imposed before the experimental results are obtained.

Overall, the methodology follows Figure 3.

**Figure 3**

*Overall Methodology and System Flow for Tagalog Visual Speech Recognition*



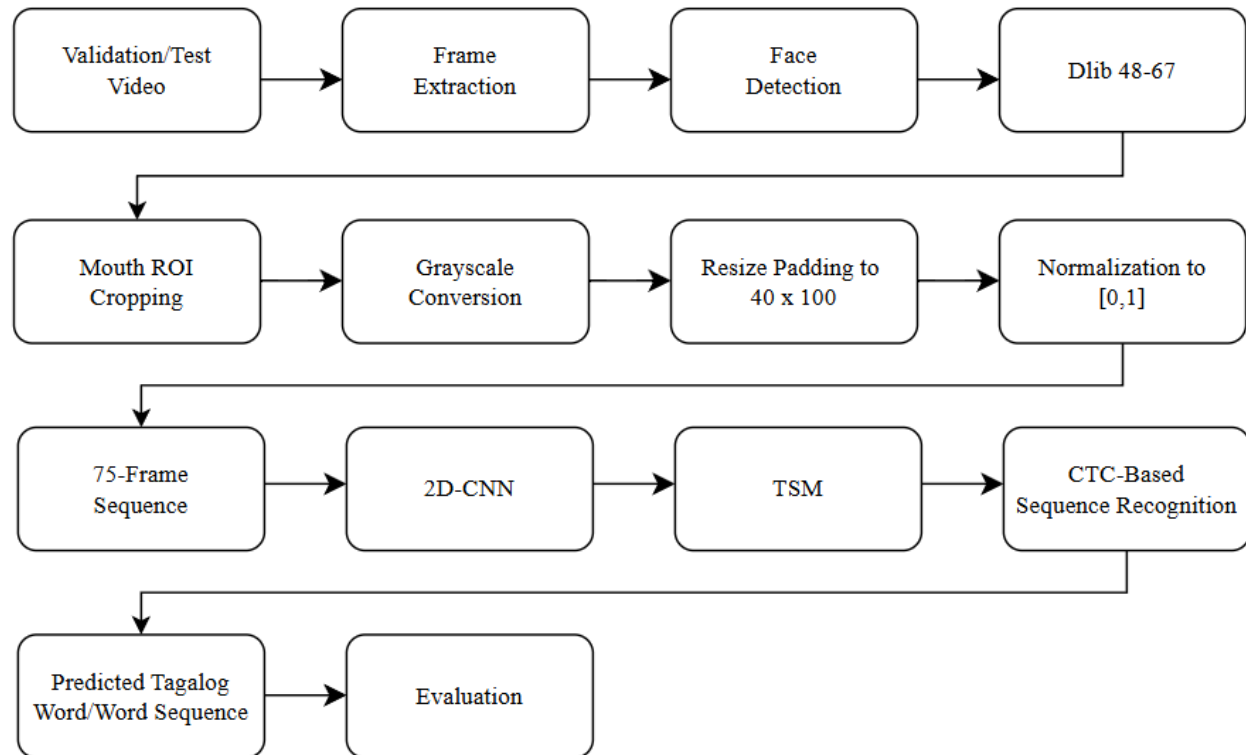


*Note.* The methodology illustrates the complete research and system flow, beginning with Tagalog visual speech data organization and labeling, followed by the train/validation/test split, training-only augmentation, video preprocessing, 75-frame sequence formation, 2D-CNN spatial feature extraction, TSM temporal modeling, and CTC-based sequence recognition. The flow continues to recognition evaluation, computational evaluation, selected-error identification, qualitative error analysis, and integration of findings.

For validation and testing, augmentation is omitted, illustrated in Figure 4.

**Figure 4**

*Validation and Testing Flow for Tagalog Visual Speech Recognition*



*Note.* The validation and testing flow illustrates the processing of non-augmented validation and test videos through frame extraction, face detection, Dlib 48–67 mouth landmark localization, mouth ROI cropping, grayscale conversion, resizing and padding to  $40 \times 100$ , normalization to  $[0,1]$ , 75-frame sequence formation, 2D-CNN spatial feature extraction, TSM temporal modeling, and CTC-based sequence recognition before evaluation.

The methodology is primarily grounded in Pascual et al. (2026) for the Tagalog video conditions, augmentation types and ordering, mouth localization, grayscale processing, spatial dimensions, and sequence preparation. Li et al. (2021) provide supporting literature for TSM-based lip-reading. Li et al. (2022) provide supporting literature for combining a 2D-CNN with TSM for lip-reading. Lin et al. (2022) provide the basis for the temporal-shift mechanism. Innocente et al. (2025) provide the basis for the dataset partitioning and additional

support for visual augmentation. The approach of Crespo and Recario (2024) provides a methodological comparison with a Philippine/Filipino VSR system. Cruz et al. (2026) provides a methodological comparison with a regional-language VSR system. These methods are adapted to the present study without reproducing the complete architectures of the cited studies. In particular, the proposed system uses 2D-CNN + TSM + CTC rather than the 3D-CNN and recurrent architectural configurations employed by several previous approaches.

Data augmentation remains restricted to the training partition, performed before mouth ROI extraction, and applied consistently across each complete sequence. The chronological order of frames is preserved, while validation and testing remain unaugmented to maintain an unbiased final evaluation.

## REFERENCES

- Baloch, A., Ali, M., Hussain, L., Daud, A., & Dawood, H. (2026). Lip reading systems for Urdu alphabets in diverse environments. *Scientific Reports*, 16, Article 27134. <https://doi.org/10.1038/s41598-026-53299-5>
- Bardhan, A., Singh, A., & Attri, S. H. (2024). CNN based real time detection of words from lip movements and automated into text. In K. Santosh, S. K. Sood, H. M. Pandey, & C. Virmani (Eds.), *Advances in artificial-business analytics and quantum machine learning* (Vol. 1191, pp. 87–102). Springer. [https://doi.org/10.1007/978-981-97-2508-3\\_7](https://doi.org/10.1007/978-981-97-2508-3_7)
- Barrios, A., & Garcia, R. (2023). Filipino Children's Acquisition of Nominal and Verbal Markers in L1 and L2 Tagalog. *Languages*, 8(3), 188. <https://doi.org/10.3390/languages8030188>
- Crespo, S. D. A., & Recario, R. N. C. (2024). Lip reading Filipino words using convolutional neural networks and recurrent neural networks. In K. Arai (Ed.), *\*Intelligent systems and applications: Proceedings of the 2023 Intelligent Systems Conference (IntelliSys)* (Vol. 2, pp. 771–790). Springer. [https://doi.org/10.1007/978-3-031-47724-9\\_51](https://doi.org/10.1007/978-3-031-47724-9_51)
- Cruz, J. P. T., Abadiano, A. B., Sangilan, F. O., Requillo, E. G. T., & Juanatas, R. C. (2026). Utilizing residual network 50 convolutional neural network architecture for enhanced Philippine regional language classification on Jetson Orin Nano. *Engineering Proceedings*, 134(1), Article 2. <https://doi.org/10.3390/engproc2026134002>
- Dweik, W., Altorman, S., & Ashour, S. (2022). Read my lips: Artificial intelligence word-level Arabic lipreading system. *Egyptian Informatics Journal*, 23(4), 1–12. <https://doi.org/10.1016/j.eij.2022.06.001>
- Fu, Y., Lu, Y., & Ni, R. (2023). Chinese lip-reading research based on ShuffleNet and CBAM. *Applied Sciences*, 13(2), Article 1106. <https://doi.org/10.3390/app13021106>
- Hao, B., Zhou, D., Li, X., Zhang, X., Xie, L., Wu, J., & Yin, E. (2025). LipGen: Viseme-guided lip video generation for enhancing visual speech recognition. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1–5). IEEE. <https://doi.org/10.1109/ICASSP49660.2025.10889163>
- Haq, M. A., Ruan, S.-J., Cai, W.-J., & Li, L. P.-H. (2022). Using lip reading recognition to predict daily Mandarin conversation. *IEEE Access*, 10, 53481–53489. <https://doi.org/10.1109/ACCESS.2022.3175867>
- Innocente, C., Boemio, M., Lorenzetti, G., Pulito, I., Romagnoli, D., Saponaro, V., Marullo, G., Ulrich, L., & Vezzetti, E. (2025). Deep learning-based lip-reading for vocal impaired patient rehabilitation. *Computer Modeling in Engineering & Sciences*, 143(2), 1355–1379. <https://doi.org/10.32604/cmescs.2025.063186>
- Li, H., Mamut, M., Yadikar, N., Zhu, Y., & Ubul, K. (2021). Channel enhanced temporal-shift module for efficient lipreading. In J. Feng, J. Zhang, M. Liu, & Y. Fang (Eds.), *Biometric recognition: 15th Chinese Conference, CCBR 2021, Shanghai, China, September 10–12, 2021, proceedings* (Vol. 12878, pp. 474–482). Springer. [https://doi.org/10.1007/978-3-030-86608-2\\_52](https://doi.org/10.1007/978-3-030-86608-2_52)
- Li, H., Yadikar, N., Zhu, Y., Mamut, M., & Ubul, K. (2022). Learning the relative dynamic features for word-level lipreading. *Sensors*, 22(10), Article 3732. <https://doi.org/10.3390/s22103732>
- Li, Y., Hashim, A. S., Lin, Y., Nohuddin, P. N. E., Venkatachalam, K., & Ahmadian, A. (2024). AI-based visual speech recognition towards realistic avatars and lip-reading applications in the metaverse. *Applied Soft Computing*, 164, Article 111906. <https://doi.org/10.1016/j.asoc.2024.111906>
- Lin, J., Gan, C., Wang, K., & Han, S. (2022). TSM: Temporal shift module for efficient and scalable video

- understanding on edge devices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5), 2760–2774. <https://doi.org/10.1109/TPAMI.2020.3029799>
- Ma, P., Martinez, B., Petridis, S., & Pantic, M. (2021). Towards practical lipreading with distilled and efficient models. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7608–7612). IEEE. <https://doi.org/10.1109/ICASSP39728.2021.9415063>
- Ma, P., Petridis, S., & Pantic, M. (2022). Visual speech recognition for multiple languages in the wild. *Nature Machine Intelligence*, 4, 930–939. <https://doi.org/10.1038/s42256-022-00550-z>
- Ma, Y., & Sun, X. (2025). Spatiotemporal feature enhancement for lip-reading: A survey. *Applied Sciences*, 15(8), Article 4142. <https://doi.org/10.3390/app15084142>
- Oghbaie, M., Sabaghi, A., Hashemifard, K., & Akbari, M. (2021). *Advances and challenges in deep lip reading*. arXiv. <https://doi.org/10.48550/arXiv.2110.07879>
- Panagos, I. I., Sfikas, G., & Nikou, C. (2024). Visual speech recognition using compact hypercomplex neural networks. *Pattern Recognition Letters*, 186, 1–7. <https://doi.org/10.1016/j.patrec.2024.09.002>
- Pascual, A. D. V., Ayo, T. J. G., & Paglinawan, C. C. (2026). Tagalog lip-reading system using 3D convolutional neural network with bidirectional long short-term memory. *Engineering Proceedings*, 134(1), 55. <https://doi.org/10.3390/engproc2026134055>
- Sheng, C., Kuang, G., Bai, L., Hou, C., Guo, Y., Xu, X., Pietikäinen, M., & Liu, L. (2024). Deep learning for visual speech analysis: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9), 6001–6022. <https://doi.org/10.1109/TPAMI.2024.3376710>
- Tamura, J., Itaya, Y., Hayashi, K., & Yamamoto, K. (2026). Statistical inference of the Matthews correlation coefficient for multiclass classification. *Journal of Applied Statistics*, 1–26. <https://doi.org/10.1080/02664763.2026.2666305>
- Wang, H., Pu, G., & Chen, T. (2022). A lip reading method based on 3D convolutional vision transformer. *IEEE Access*, 10, 77205–77212. <https://doi.org/10.1109/ACCESS.2022.3193231>
- Yeo, J. H., Kim, C. W., Kim, H., Rha, H., Han, S., Cheng, W.-H., & Ro, Y. M. (2025). Personalized lip reading: Adapting to your unique lip movements with vision and language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(9), 9472–9480. <https://doi.org/10.1609/aaai.v39i9.33026>