Project LIF: A Lip Reading Learning Tool for Basic Filipino Words using 3D CNN and Bidirectional Long Short Term Memory

Bautista, Engelbert C.

Bernal, Christine Joy A.

Diaz, Daniele Quiesha E.

Mercurio, Mary Gwynneth C.

A Thesis

In partial Fulfillment of the Requirements

for the degree of Bachelor of Science in Computer Science

College of Communication and Information Technology

President Ramon Magsaysay State University

Iba, Zambales

May 2025

Chapter 1

INTRODUCTION

Project Context

Technology has continuously shaped human communication, making interactions more efficient and inclusive. Over the years, innovations in artificial intelligence (AI), have led to major advancements in communication systems. According to Chen, Liu, and Wang (2021), AI-powered tools such as speech-to-text applications, real-time translators, and assistive technologies have significantly improved accessibility for individuals with disabilities. These technologies not only enhance day-to-day interactions but also bridge linguistic and physical communication barriers, making digital communication more inclusive (Lopez & Ramos, 2021). In the Philippines, Republic Act No. 7277 (Magna Carta for Persons with Disabilities), Republic Act No. 11106 (Filipino Sign Language Act), and the Department of Education's Special Education (SPED) guidelines mandate accessible educational tools for hearing-impaired learners, emphasizing the need for inclusive technologies.

One of the key innovations in AI-driven communication is image recognition, which has transformed human-computer interaction (HCI) by allowing machines to interpret facial expressions, gestures, and emotions. Sharma, Zhang, and Wang (2020) state that image recognition technology enables contactless interaction, making it particularly useful in healthcare, security, and assistive communication. Gesture and facial recognition have paved the way for emotion-aware AI systems, which enhance accessibility for individuals with disabilities by enabling non-verbal interaction with technology (Alp, Kivrak, & Onen, 2021). A crucial advancement in this area is lip-reading technology, which allows AI to interpret speech visually, supporting hearing-impaired students in learning language skills. According to Akbari, Momenzadeh, and Khalilzadeh (2021), lip-reading technology has evolved through deep learning models, allowing machines to accurately interpret human speech even in challenging conditions.

Lip-reading technology has gained significant attention in recent years due to its potential to enhance educational accessibility. Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have significantly improved recognition accuracy, making lip-reading systems more reliable (T. N. Sainath et. al., 2015). These technologies analyze lip movements, facial structures, and phonetic patterns to extract speech data, helping hearing and speech students and educators in classroom settings. However, most AI lip-reading models are trained in English and other widely spoken languages, making them ineffective for Filipino speakers, whose language structure and phonetics differ significantly.

Despite advancements in AI-powered lip-reading, there are no dedicated systems designed for the Filipino language. Lopez and Ramos (2021) emphasize that existing models, primarily developed for English, Chinese, and Spanish, struggle to recognize Filipino phonetics, syllables, and regional dialects. This linguistic gap presents a significant challenge for Filipino-speaking hearing and speech students, as they cannot fully benefit from current AI-based speech recognition technologies. The lack of large-scale Filipino lip-reading datasets makes it difficult to train AI models effectively, further limiting the accuracy of existing solutions.

Current speech-to-text systems rely heavily on clear audio, which limits their effectiveness for teaching visual speech recognition in educational settings. To address this gap, Project LIF: A Lip Reading Learning Tool for Basic Filipino Words using 3D CNN and Bidirectional Long Short Term Memory focuses on developing an Android-based tutorial platform for Filipino lip-reading. While communication difficulties highlight the need for improved accessibility, Project LIF serves as a learning tool, enabling hearing and speech students to practice visual speech recognition in controlled classroom settings. By leveraging 3D-CNN and Bidirectional LSTM to recognize basic Filipino syllables, this study provides an educational tool that supports language development, ensuring compliance with RA 7277 and DepEd SPED guidelines.

Purpose and Description

Lip-reading technologies have been developed to enhance accessibility, linguistics and assist individuals who suffer from speech and hearing impairments. By creating an AI-powered lip-reading interpreter specifically for Filipino speakers, this study fills a crucial gap in speech recognition technology. Its primary purpose is to support learning and teaching of visual speech recognition, especially in Filipino phonemes and mouth movements. This study will contribute to the advancement of AI applications in education, particularly in language learning, speech training, and special education.

Specifically, this study will be beneficial to the following:

Lip-reading Learners. This study shall benefit the individuals learning to lip-read including those preparing for roles in speech therapy, special education, or personal development. This learning tool will help improve their ability to interpret Filipino syllables.

Hearing-impaired Individuals. This study shall benefit hearing-impaired individuals to help them develop the foundational skills of recognizing Filipino syllables through mouth movements. This technology will help them to strengthen their ability to recognize lip movements for Filipino syllable

Speech-impaired Individuals. This study shall benefit the individuals who suffer from speech disorders, specifically those who can make mouth movements but cannot verbalize their thoughts. This lip-reading technology shall allow them to improve their articulation through visual feedback.

Educators and Therapists. This study shall benefit the speech and language professionals, and special education teachers. This tool can be used as a support instruction in lip-reading and pronunciation of Filipino words.

Future Researchers. This study will be helpful for future researchers to enhance and improve. This study will serve as the foundation for enhancing real-time lip-reading technology, specifically in the Filipino language.

Objectives of the Study

This study aims to provide a reliable lip-reading interpreter from the Filipino language using computer vision.

Specifically, this study aims to:

1. Train a model for lip-reading using Long Short-Term Memory algorithm and 3D Convolutional Neural Network.
2. Evaluate the performance of the model in terms of:

   3.1 Accuracy;

   3.2 Precision;

   3.3 Recall; and

   3.4 F1-score
3. Develop an Android application with the integration of the trained model.

Scope and Limitations

Scope

This study focuses on the development of a mobile-base tutorial application designed to recognize Filipino lip movements and convert them into corresponding words for educational purposes. The system will serve as a learning tool for individuals interested in understanding and practicing Filipino lip reading. It is not intended as a direct communication medium for individuals with speech or hearing impairment.

The application is limited to the recognition of the combination of the Filipino Phonetics (80 syllables), providing basic tutorials using words that are inside the dictionary and simple sentences. The system is optimized for users who wish to learn or practice Filipino lip movements, including students, hearing and speech impaired, and language educators. It is intended as an educational tool to support foundational lip reading and pronunciation skills in the Filipino Language.

Limitations

The recognition process may have a noticeable delay between lip movement and the output display. The system is limited to recognizing only 80 Filipino syllables, which constrains the range of vocabulary and sentence structures it can accurately teach. It supports only Filipino language and does not recognize mixed-language inputs such as Taglish. The recognition accuracy may also be affected by individual facial variations, such as the presence of braces, missing teeth, or facial disfigurements, as well as non-native pronunciation patterns. Environmental conditions like poor lighting, improper angles, or excessive distance from the camera can reduce accuracy. Moreover, the system is intended solely as an educational tool, not as a communication aid for people with hearing or speech impairments. Lastly, performance is dependent on the processing capability of the Android device being used; low-end or outdated hardware may result in slower response times or decreased functionality.

Definition of Terms

3D CNN (Convolutional Neural Network). A type of deep neural network that processes spatial and temporal data in three dimensions, used in this study to analyze sequences of lip movements in video frames.

Assistive Technologies. It refers to a tool or device designed to support individuals with disabilities by improving their ability to perform tasks or communicate more effectively.

Bidirectional LSTM (Long Short- Term Memory). It refers to a variant of LSTM neural network that processes input sequences in both forward and backward directions, improving temporal pattern recognition in sequential data like lip movement.

CNN (Convolutional Neural Network). It refers to a type of deep learning model that uses convolutional layers to extract spatial features from images; in this study, used for lip reading.

Computer Vision. It refers to a field of artificial intelligence that enables computers to interpret and understand visual information from the world.

Deep Learning. It refers to a type of machine learning that uses artificial neural networks to learn from data.

Facial Recognition. It refers to a technology that uses computer algorithms to identify a person by analyzing the unique features of their face and confirming their identity; used in this study to localize lips and facial areas.

Filipino Phoneme/Phonetics. It refers to the smallest unit of sound in the Filipino language that distinguishes meaning; important in training the lip-reading model to differentiate spoken words.

HCI (Human- Computer Interaction). It refers to the design of computer systems focused on user interaction, especially important in educational technologies.

Hearing-impaired. It refers to individuals who have partial or complete hearing loss; the primary beneficiaries of this study's lip-reading tool.

Lip-reading. It refers to understanding speech from analyzing a speaker's lip movements without relying on the sound of the voice.

Machine Learning. It refers to a type of artificial intelligence (AI) that allows computers to learn from data and improve their performance without being explicitly programmed.

OpenCV (Open Source Computer Vision). It refers to an optimized tools, hardware, and Computer Vision library in real-time.

Phonetic. It refers to the sounds and other phenomena of speech. It is a branch of linguistics that studies how humans produce and perceive sounds or, in the case of sign languages, the equivalent aspects of sign.

SPED (Special Education). It refers to the Educational Programs and tools designed to meet the needs of students with disabilities, including hearing and speech impairments.

Speech Disorders. It refers to a condition that affects someone's ability to produce speech sounds correctly or fluently.

Syllables. It refers to a unit of pronunciation in the Filipino language consisting of one or more phonemes; this study uses 80 core syllables as the basis for recognition training.

TensorFlow. It refers to an end-to-end open source machine learning platform for everyone that is used in this project to build and train lip- reading models.

Tutorial Tool. It refers to a software system designed to support learning through guided, interactive modules; in this study, focused on teaching Filipino lip- reading skills.

Chapter 2

REVIEW OF RELATED LITERATURE/SYSTEMS

Technical Background

To develop the lip-reading tutorial application, the researchers utilized a comprehensive set of technologies spanning machine learning, computer vision, mobile development, and video processing. These tools enabled the creation, training, deployment, and testing of a lip-reading model optimized for Filipino syllables. The following software elements were utilized by the researchers in the creation of the application: Google Drive, Google Colab, Android Studio OpenCV, DLib, 3D - CNN, Bidirectional LSTM, Python, Redmi Note 13 Pro 5G Camera (720p 30fps), Numpy, TypeScript, MTCNN, Train-Test-Validation Split, MoviePy, Pytorch Mobile, Tensorflow Lite, Kotlin, JSON.

Details of the Technologies used:

3D Convolutional Neural Networks (3D-CNN) – A neural network architecture that processes data across spatial and temporal dimensions, allowing the model to capture changes in lip movements over time.

Android Studio - Is an IDE for Android apps, enabling UI design, camera integration, and machine learning for lip reading using TensorFlow Lite and OpenCV.

Dlib – A machine learning toolkit that includes facial landmark detection. It assists in locating and aligning the mouth region from facial images or video frames.

Google Colab – A cloud-based platform that provides access to GPUs and TPUs, allowing efficient training and testing of deep learning models.

Google Drive – Used as cloud storage for datasets, training logs, and model checkpoints, providing easy collaboration and integration with Google Colab

JSON (JavaScript Object Notation) - lightweight data format used for configuration settings, data exchange between system components, and file handling.

Kotlin - A modern, concise programming language used for Android development. It serves as the primary language for building the application's frontend and logic.

Bidirectional Long Short-Term Memory (LSTM) - A type of LSTM that processes data in both forward and backward directions, allowing the model to learn from past and future context. This improves accuracy in lip reading by better capturing patterns in video frame sequences.

MoviePy – A Python library used for video preprocessing, such as trimming, splitting, and converting video clips into datasets for training.

NumPy – A numerical computing library in Python used to handle image pixel data and multidimensional arrays.

OpenCV (Open Source Computer Vision Library) – – A computer vision toolkit for image preprocessing, frame extraction, and facial region detection in real-time applications.

PyTorch Mobile – A lightweight version of PyTorch optimized for running machine learning models on mobile devices, enabling real-time inference within the Android app.

Python – The primary language used for scripting the machine learning pipeline, including model training, preprocessing, and evaluation.

Redmi Note 13 Pro 5G Camera (720p 30fps) – The smartphone camera used to collect video datasets for lip movement, selected for its balance between frame clarity and storage efficiency.

TensorFlow Lite – A mobile-optimized framework used to run trained models on Android devices with minimal latency and resource consumption.

Train-Test-Validation Split (70-20-10) –A method of dividing the dataset into training, testing, and validation sets to ensure accurate model evaluation and avoid overfitting.

TypeScript – A typed superset of JavaScript used for maintaining front-end logic and ensuring reliable integration between app components.

Review of Related Literature, Studies/Systems

Lip Reading and Application of Deep Learning in Visual Speech Recognition

Lip reading, the ability to recognize what is being said from visual information alone, is an impressive skill but remains challenging for novices. It is inherently ambiguous at the word level due to homophemes—different characters that produce the same lip sequence (e.g., ‘p’ and ‘b’). However, these ambiguities can be partially resolved using the context of neighboring words in a sentence and/or a language model (Chung et al., 2017).

Machine-based lip reading, also known as visual speech recognition, has advanced significantly in recent years. This progress is largely attributed to the emergence of large-scale datasets and the integration of deep learning models in neural networks, which have enhanced the system’s ability to interpret speech through visual cues alone (Chung & Zisserman, 2017).

As stated by Prajwal et al. (2021), lip reading, or visual speech recognition, involves recognizing speech silent video. This technology has practical applications, including enhancing speech recognition in noisy environments, enabling silent dictation, and transcribing archival silent films. It also has significant medical uses, such as helping individuals with conditions like Lou Gehrig’s disease or aphonia communicate through lip movements.

According to Feng et al. (2020), lip reading, also referred to as visual speech recognition, has gained significant traction due to the advancements in deep learning and the availability of large-scale datasets. They emphasized that while many current models achieve high performance, these often rely on complex network architectures and customized training strategies that are not always clearly documented. To address this, the authors presented a detailed analysis of various model components and training techniques, demonstrating that even simple improvements to the baseline can result in performance that matches or exceeds state-of-the-art lip reading systems.

According to the study by Zhao et al. (2020), one of the major challenges in lip reading is the difficulty in extracting clear and distinguishable features from lip movements due to their subtle and often ambiguous nature. To address this, the authors proposed a method called Lip by Speech (LIBS), which enhances visual speech recognition by incorporating knowledge distilled from speech recognizers. As stated in the paper, this cross-modal distillation allows the model to utilize complementary auditory cues that are not easily captured by visual input alone. The researchers further introduced an alignment strategy to manage the different lengths of video and audio sequences, along with a filtering mechanism to refine the speech recognizer's outputs before transferring them to the lip reading model. Their results showed significant improvements on benchmark datasets, validating the effectiveness of the LIBS approach (Zhao et al., 2020).

As stated by Lin et al. (2017), one of the significant challenges in lip reading is accurately detecting the lip contours during speech, especially given individual differences such as lip shape and makeup. To address this, the authors proposed a novel lip-reading recognition algorithm designed to recognize English vowels by analyzing lip contours during speech. As stated in the paper, the algorithm automatically detects the mouth region of interest (ROI), thereby reducing the influence of individual differences and eliminating the need for pre-training. The study further evaluated the algorithm's performance under various environmental conditions and individual differences, achieving an accuracy rate exceeding 80% in lip-reading recognition .

As studied by Fernández-López and Sukno (2018), Automatic Lip Reading (ALR) has significantly evolved with the advent of deep learning techniques. The authors highlight that traditional methods, which often relied on handcrafted features and separate classification stages, have been largely supplanted by end-to-end deep learning architectures. These modern approaches integrate feature extraction and classification into a unified model, allowing for more efficient and accurate lip-reading systems. As stated in the paper, deep learning models have demonstrated substantial improvements in complex tasks, such as word and sentence recognition, achieving up to a 40% increase in word recognition rates compared to traditional methods. The study also emphasizes the importance of large-scale datasets in training these models, noting a trend towards datasets that capture realistic application settings with a vast number of samples per class. This shift has been instrumental in enhancing the performance and generalizability of ALR systems (Fernández-López & Sukno, 2018).

According to the study by Afouras, Chung, and Zisserman (2018), advancements in deep learning have significantly enhanced the performance of visual speech recognition systems. The researchers developed and compared three distinct neural network architectures for lip reading: a recurrent model utilizing Long Short-Term Memory (LSTM) networks, a fully convolutional model, and a Transformer-based sequence-to-sequence model. The recurrent and fully convolutional models were trained using Connectionist Temporal Classification (CTC) loss and incorporated explicit language models for decoding, while the Transformer model employed an attention mechanism for sequence modeling. As stated in the paper, their best-performing model achieved a more than 20% improvement in word error rate on the challenging BBC-Oxford Lip Reading Sentences 2 (LRS2) benchmark dataset. Furthermore, the study explored the application of the fully convolutional model for online (real-time) lip reading of continuous speech, demonstrating high performance with low latency .

Cruz et al. (2018) conducted a study focusing on the analysis of lip movements associated with the pronunciation of English letters by Filipino speakers. The research aimed to develop a model capable of recognizing spoken English letters through lip reading, emphasizing video processing techniques. The study involved thirty Filipino English speakers, evenly divided between men and women, who were recorded pronouncing each letter of the English alphabet. To ensure consistency, participants without facial hair or lip deformities were selected, and recordings were conducted in a controlled environment using a DSLR camera at 640x424 resolution and 24 frames per second. A custom image database was created from these recordings to analyze and recognize lip movements associated with each English letter, addressing challenges in lip reading such as individual differences in lip shape and movement.

Lesani et al. (2019) introduced an innovative approach to human–mobile interaction by developing an offline Persian automatic lip reader designed for Android smartphones. The system utilizes the device's camera to capture and analyze the user's lip movements, enabling the recognition of spoken words and sentences without relying on audio input. This method allows users to operate mobile applications through silent speech, offering a hands-free and noise-resilient alternative to traditional voice commands. By processing visual speech cues directly on the device, the application ensures user privacy and functionality in environments where audio-based interaction may be impractical.

Chen et al. (2020) developed DualLip, a system that simultaneously enhances lip reading and lip generation by leveraging their interrelated nature and utilizing unlabeled data. The system employs a dual-task approach: first, generating lip videos from unlabeled text using a lip generation model, and then employing these pseudo-pairs to improve lip reading performance; second, generating text from unlabeled lip videos with a lip reading model, and using these pseudo-pairs to enhance lip generation. Experiments on the GRID and TCD-TIMIT datasets demonstrated that DualLip effectively improves lip reading and generation, especially in scenarios with limited paired data. Notably, the lip generation model trained with only 10% paired data and 90% unpaired data outperformed models trained with full paired datasets.

In their study, Ma et al. (2022) systematically investigated various training strategies to enhance the performance of lip-reading models for isolated word recognition. They explored the effectiveness of different data augmentation techniques, temporal models, and additional strategies such as self-distillation and incorporating word boundary indicators. The authors found that Time Masking (TM) was the most impactful augmentation method, followed by mixup, while Densely-Connected Temporal Convolutional Networks (DC-TCN) emerged as the most effective temporal model for lip-reading isolated words. By combining these methods, they achieved a classification accuracy of 93.4% on the LRW dataset, marking a 4.6% improvement over the previous state-of-the-art performance. Furthermore, pre-training on additional datasets further enhanced performance, reaching an accuracy of 94.1% .

Rathod et al. (2024) highlight the significance of lip reading as a crucial component of communication, particularly for individuals with hearing impairments or in environments where audio cues are limited. They note that traditional methods of lip reading rely on manual interpretation, which often leads to limitations in accuracy and efficiency. To address these challenges, the authors propose a novel approach that integrates Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to enhance lip reading capabilities. By combining CNNs for feature extraction from visual data and RNNs for sequential modeling, their approach aims to improve the accuracy and robustness of lip reading-based text extraction and translation. This integration allows for the automatic extraction of text from lip movements in videos and subsequent translation into written or spoken language, offering promising opportunities to enhance communication accessibility and bridge linguistic gaps .

As studied by Recario and Crespo (2024), lip reading is the ability to recognize words and sentences through the movement of the mouth. The proliferation of audio-video data, especially in teleconferencing and streaming platforms, has made lip reading an intriguing area of research, particularly for low-resource languages. In their study, the researchers collected a video dataset of Filipinos speaking ten distinct words and experimented with three different configurations of deep learning models: Inception v3 CNN combined with a GRU model, VGG-19 CNN with a GRU model, and a standalone GRU model. The aim was to determine the accuracy of each configuration in lip-reading Filipino words. Their findings indicated that, despite the limited volume of training data, the features extracted by the CNN models were insufficient for accurate word prediction. Additionally, data augmentation did not significantly enhance the models' accuracy. However, training the GRU model using the height of the lip's opening as a feature yielded the best results, achieving a test accuracy of 17% and a validation accuracy of 26%.

Lip Reading on Different Fields

Over 5% of the world's population, or 466 million people, have a serious hearing loss, according to statistics published (Ryumin et al., 2019). Over 900 million individuals, or one in ten, are estimated to have a serious hearing loss by 2025. Sign language (SL) and lip reading is a helpful tool for everyday communication for the deaf and speech-impaired community(Denby et al., 2010).

Visual speech recognition (VSR), a term used for automatic lip reading, has drawn a lot of interest lately due to its potential applications in biometry, biomedicine, sign language recognition, audio-visual speech recognition, and human-machine interaction (Katsaggelos et al., 2015). In order to remove confusion in acoustic communication aspects, hearing-impaired individuals and those in noisy acoustic environments (noise, reverberation, numerous speakers) mostly rely on visual input. Although a significant amount of study has been devoted to the topic of visual speech decoding, the challenge of lip reading for hearing impaired people remains an unresolved issue in the field (Akbari et al., 2018).

According to the study on elite European deaf athletes (Kurkova et al., 2011), 84.9% of them were from hearing households, and the Gallaudet Research Institute report (Gallaudet Research Institute, 2006) found that 83.4% of children were born into hearing families. The majority of deaf athletes had hearing impairments that either developed within the first two years of life or were present from birth. The majority of deaf athletes depend on their capacity to communicate through total communication because hearing aids are used so frequently. In line with other research findings, it is significant that deaf athletes who use sign language also rely on their speaking, lipreading, and fingerspelling skills (Sheetz, 2004).

Simple data augmentations, a two-branch network that amplifies multi band frequencies (Masi et al., 2020), an autoencoder-like structure to act as an anomaly detector, and patch-based classification to model local patterns are some recent attempts to improve generalization to novel forgeries. However, these techniques are still far superior to observed interventions (Wang et al., 2020).

By limiting CNN filters (Bayar & Stamm, 2016) or employing relatively shallow networks to concentrate on mesoscopic characteristics, some previous face forgery detection efforts bias the network away from learning high-level features. Rossler et al. (2019) showed that a deep, unconstrained Xception network outperforms these. Face X-ray (Li et al., 2020) is a very successful technique that suggests predicting the blending boundary between the inserted, modified face and the rest of the image. Despite its remarkable generalization in cross-manipulation studies, it depends on patterns that are frequently undetectable and vulnerable to low-level post-processing techniques.

Algorithms Used in Lip Reading Technology

A study by Winoto (2018), describes the development of a deep-learning model which translates mouth movements into words. They use the Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) to learn the sequence of mouth images. They also use the GRID corpus containing a video recording of 34 speakers each speaking 1000 sentences in English as the dataset of the model. The model is capable of predicting words from mouth video frames with 52.5% word accuracy exceeding human accuracy of 14.47%.

For lipreading, the researchers proposed a model called Hybrid Lip Reading (HLR-Net) that is based on the deep convolutional neural network. It includes three stages: preprocessing, encoder, and decoder stages. They use inception, gradient, and bidirectional GRU layers. While building the decoder, they use attention, fully-connected, activation function layers. The proposed HLR-Net model achieves 4.9% for CER, 9.7% for WER, and 92% for Bleu Score in comparison with recent models such as the LipNet Model, lip reading model with cascaded attention (LCANet), and attention-CTC (A-ACA) model for the unseen speakers. While they result in 1.4% for CER, 3.3% for WER, and 99% for Bleu score for overlapped speakers. (Sarhan, Elshennawy, & Ibrahim, 2021)

In a study about lip reading, they used various methods to predict words and phrases from videos without audio. They pre-trained on human faces of celebrities from IMDB and Google images using VGGNet and they explore different ways to handle image sequences. They trained VGGNet on concatenated images from multiple frames in each sequence, as well as used in conjunction with LSTMs for extracting temporal information Although the LSTM models failed to outperform other methods, the concatenated model uses the nearest-neighbor interpolation performed well, achieving a validation accuracy of 76%. (Garg, Noyola, & Bagadia)

A study by Shrestha (), verified the use of machine learning by applying deep learning and neural networks to devise an automated lip-reading system. They use a subset of data trained on two separate CNN architectures. The dataset includes segments of 1.16 seconds (approximately 29 frames) of 500 different word instances with up to 1000 utterances for each word spoken by different speakers. They also evaluated their model based on the accuracy of the predicted words.

In a study about lip reading, they employ a deep learning model that features a 3D-convolution network and bidirectional LSTM, enabling an accurate sentence-level prediction based solely on visual lip movements. Their model was trained using pre-segmented lip regions with an impressive accuracy of 97% and transformed into animated GIFs for effective pre-training. (Inamdar, Sundarr, Khandelwal & Ajeyprasaath, 2023)

In a study about lip reading with densely connected temporal convolutional networks, where they present the Densely Connected Temporal Convolutional Network (DC-TCN) for lip-reading isolated words. They also introduced dense connections into the network to capture more robust temporal features. Also, their approach utilizes the Squeeze and Excitation block which is a lightweight attention mechanism that will further enhance the power of the model’s classification. According to their results, the DC-TCN method has achieved 88.36% accuracy on the Lip Reading in the Wild)LRW) dataset and 43.65% on the LRW-1000 dataset. (Ma, Wang, Shen, et,al, 2021)

A study by Putcha, Rajam, Sugamya, & Gopala, 2024, wherein explores the fusion of these domains by proposing a novel approach for text extraction and translation using lip reading and deep learning. Wherein they employ deep learning architectures such as CNNs and RNNs to accurately extract text content from lip movements captured in video sequences. They propose a model that consists the lip region detection, feature extraction, text recognition, and translation. The model then identifies and isolates the lip region within video frames using a CNN-based object detection approach. Moreover, they extracted lip regions using CNNs to capture intricate motion patterns that will convert these visual features into text.

A study by Jeon, Elsharkawym & Kim (2022), proposed a lip reading architecture that combines three different convolutional neural networks (CNNs, 3D CNN, a densely connected 3D CNN, and a multi-layer feature) followed by a two-layer bidirectional gated recurrent unit. Using connectionist temporal classification, they trained the entire network. According to their results, the proposed architecture reduced the character and word error rates of the baseline model by 5.681% and 11.82% for the unseen speaker dataset.

A study by Fenghour (2021), proposes an automated lip reading model that can predict sentences using visemes as a classification schema. The proposed system is also lexicon-free and uses purely visual cues. The system is designed to lip-read sentences covering a wide range of vocabulary and to recognize words that may not be included in system training with only a limited number of visemes as a class to recognize. The lip-reading system is able to predict sentences as a two-stage procedure with visemes being recognized as the first stage and words being classified as the second stage. The second stage, however, has to overcome both the one-to-many mapping problem posed in lip reading where one set of visemes can map to several words. Another problem is that visemes are confused or misclassified to begin with. As a result, the attained word accuracy rate is 79.6% for the LRS2 dataset.

A study by Deocampo, Villarica, and Vinluan (2023), proposes a hybrid approach wherein they combined Convolutional Neural Networks (CNNs) and Long-Short-Term Memory (LSTM) models to improve lip reading accuracy in Tagalog. They collected a dataset of 450 videos of nine native Tagalog speakers saying 50 known phrases in Tagalog. Their dataset captures phonetic variations, and various speaking styles, while also considering environmental factors to ensure generalization.

A study about lip reading in Filipino words by Crespo and Recario (2023) uses Convolutional Neural Networks and Recurrent Neural Networks. They collected a video as a dataset of Filipinos speaking ten Filipino words. They used this dataset to train and experiment with three configurations of deep learning models, which are the Inception v3 CNN and GRU model, VGG-19 CNN and GRU model, and GRU model to determine different configuration's accuracy in Filipino words. Moreover, they use different combinations of parameters for each model, using both the original and augmented datasets. This study faces several challenges including a low volume of training data and the features extracted by CNN were not enough to predict what word is being spoken. They also found that performing data augmentation did not increase the model's accuracy significantly. However, training using the height of the lip's opening as a feature on the GRU model achieved a test accuracy of 17% and a validation accuracy of 26%.

According to a study by Parekh, Gupta, Chhatpar, Yash, and Kulkarni (2018), various studies have been conducted using neural networks to classify utterances, preprocess datasets, and extract relevant features, but this technique is affecting the overall accuracy because it cannot explain what features the method has learned. In their study, they propose the use of Convolutional Autoencoders (CAE) to extract lip features from video frames that will be the input to the Long-Short-Term Memory (LSTM) that outputs the final trained model. They use the CNN to extract features from their baseline model wherein the features learned from the model are compared in terms of convolved input images. In their study, they first describe the multiple datasets to test their network and the pre-processing methods. Then, they defined the architectures used to compare the results by stating the baseline and the proposed model. Lastly, they compare features learned with respective models.

A study about a comparison of models in deep lip reading developed three architectures and compared the accuracy and training times of a recurrent model using LSTMs, a fully convolutional model, and the transformer model. Both the recurrent and fully convolutional models are trained with a Connectionist Temporal Classification loss. Their best model improves the state-of-the-art word error rate on the challenging BBC-Oxford Lip Reading Sentences 2 (LRS2) benchmark dataset by over 20 percent. They also investigate the fully convolutional model when lip reading is in real-time and show how it achieves high performance with low latency. (Afouras, Chung, & Zisserman, 2018).

A study by Afouras, Chung, and Zisserman (2018), compares different models and online applications. One of the models is the Transformer, which in the results of their study achieves a WER of 50% when it was decoded with a language model trained on T2. The FC model which has a smaller number of parameters and trains faster than BL and TM achieves over 55% WER. In their study, they also described how the FC model can be used for online lip reading with low latency. They also stated that controlling how much future context a model should see is easier with using temporal convolutions.

A study by Ambati (2024) employs machine learning and deep learning for lip-reading. They utilize the Multi-Task Cascaded Convolutional Networks to detect facial landmarks. To get the lip images, the aligned faces were used. Those lip images are enhanced using the Real-Enhanced Super-Resolution Generative Adversarial Network to identify subtle lip movement in video images. After it is processed, it is fed into the architecture based on CNN where the features could be learned. Through the 3D convolutional network, the feature extraction and lip movement are learned utilizing time distributed layer with LSTM in either direction. They used a text corpus dataset known as the GRID and trained their models to obtain a 2.3% character error rate on seen speakers and 5.2% on unseen speakers.

This paper proposed the use of the hybrid neural network architecture of CNN (VGG19) to extract the visual features from the mouth ROI and attention-based LSTM to learn the sequence weights and sequence information between the frame-level features. They also achieved classification by using two fully connected layers and a SoftMax layer. The dataset consists of three males and three females pronouncing numbers from zero to nine. Each digital utterance was divided into independent video clips and each of their speakers was not trained in professional pronunciation. For the result of their proposed architecture, it has an accuracy of 88.2% which is 3.3.% general CNN-RNN model. (Tsai, Tseng & Ruan, 2024)

A study of lip reading systems for the numbers from zero to nine, they use the convolutional neural network (CNN) and recurrent neural network (RNN) to extract image features. In their study they first extract the keyframes in an independent database, and then to extract the lip image features they use the Visual Geometry Group (VGG). As a result, they found that image feature extracts are fault-tolerant and effective. Wherein they also compared two lip reading models, one with an attention mechanism, and another for a fusion model of two networks. In a result, their proposed model is 82.2% more accurate and superior to the traditional lip-reading recognition methods. (Lu & Li, 2019)

In research on Lipreading Technology by Hao, Mamut, Yadikar, Aysa, and Ubul (2020), they summarize the main research from traditional methods and deep learning methods when it comes to lipreading. They discussed that traditional lip-reading methods of lip detection and extraction, lip feature extraction, and classification are not reliable under unconstrained conditions. They also discussed the advantage of how deep learning can learn the best features from large databases.

A study by Fenghour, Chen, Guo, Li, and Xiao (2021), wherein compares Convolutional Neural Networks with other neural network architectures for feature extraction and reviews the advantages of Attention-Transformers and Temporal Convolution Neural Networks compared to Recurrent Neural Networks. They also study the different classification schemas that can be used for lip reading such as ASCII characters, phonemes, and visemes. In the result, they stated that lip reading systems consisting of components for feature extraction and classification such as 2D+3D CNNs are the most preferred for frontends at it learn spatial and temporal features. TCNs have started to replace RNNs as they perform better in parallel computation, in learning long-term dependencies, and are trained in a shorter period. Furthermore, they stated that in theory, phonemes and visemes could mean that lip-reading systems could be lexicon-free where a lip-reading system could predict a word spoken by an individual that did not appear in the training phase.

Lip Reading for Low-Resource Languages

Pu and Wang 2022 reviewed recent advancement in machine lip-reading, highlighting the role of deep learning and the importance of well-designed datasets. Their work supports the relevance of using models like 3D-CNN and Vison Transformers and emphasizes the need for language specific datasets, as area this study addresses for Filipino.

Berkol et al. presented visual lip-reading datasets in Turkish, collected from real-world video sources such as vlogs, films, and TV series. The datasets includes varied conditions. Different speakers, lighting, angles, and speech styles making it suitable for training, models to handle natural, unconstrained settings. The work contributes to expanding non-synthetic language-specific datasets for lip-reading tasks.

A study by Kurniawan and Suyanto (2020) proposed a syllable-based lip-reading model for the Indonesian language using a 3D deep learning architecture. This approach was introduced to solve a common issue in lip-reading systems, the out-of-vocabulary (OOV) problem where a model fails to recognize words that were not present in the training data. Instead of relying on whole words, the system was trained to recognize syllables, which can be recombined to form new words. This method is especially useful for low-resource languages where datasets are often limited and vocabulary grows over time. To address the small amount of available training data, the researchers applied data augmentation techniques up to 40 times, which significantly improved the model's performance. As a result, their system achieved 100% accuracy on the test set and maintained an 80% accuracy rate when tested on unseen OOV words. This study demonstrates that using syllables as the primary unit of recognition is a flexible and scalable solution that could also be applied to languages like Filipino, which share similar syllabic structures and linguistic features with Indonesian..

Astorga et al. (2023) conducted a study on the phonological similarities between Tagalog and Bahasa Indonesia, examining features such as vowel phonemes, alphabet systems, and patterns of syllable stress. Their research, grounded in the framework of Phonology as Human Behavior, highlighted that both languages are shaped by a shared

Austronesian origin, which explains the overlap in their sound systems and syllable structures. These similarities suggest that methods effective in Indonesian language processing such as syllable-based modeling could also be applicable to Tagalog. The study provides useful linguistic support for the use of syllables as the basic unit in Filipino lip-reading systems.

Kim et al. (2023) proposed a lip-reading framework aimed at addressing the challenges of developing systems for low-resource languages. Their method introduces a way to combine general speech knowledge learned from high-resource languages with language-specific information using a memory-augmented decoder known as LMDecoder. Rather than depending entirely on video-text paired data, which is often limited in low-resource settings, the model utilizes audio-text data to build meaningful representations for the target language. This makes the system more flexible and easier to apply in real-world scenarios. Their approach demonstrates that combining shared phonetic knowledge with language-specific features can lead to accurate lip-reading outputs even in the absence of large visual datasets. This framework is especially relevant for languages like Filipino, where resources for visual speech recognition remain limited.

Focusing on the limitations of traditional lip-reading models, Prajwal et al. (2022) introduced a system that uses sub-word units along with visual attention mechanisms to improve recognition accuracy. Instead of relying on full words or characters, the model breaks speech down into smaller, more manageable units, allowing it to handle ambiguity more effectively. Their approach was tested on widely used benchmark datasets, where it achieved strong results even with relatively limited training data. The study highlights how sub-word modeling, paired with attention-based feature aggregation, can enhance lip-reading performance, especially in situations where vocabulary coverage and training resources are constrained.

Synthesis

Lip reading, the ability to recognize what is being said from visual information alone, is an impressive skill but remains challenging for novices. It is inherently ambiguous at the word level due to homophemes—different characters that produce the same lip sequence (e.g., ‘p’ and ‘b’). However, these ambiguities can be partially resolved using the context of neighboring words in a sentence and/or a language model (Chung et al., 2017).

Machine-based lip reading, also known as visual speech recognition, has advanced significantly in recent years. This progress is largely attributed to the emergence of large-scale datasets and the integration of deep learning models in neural networks, which have enhanced the system’s ability to interpret speech through visual cues alone (Chung & Zisserman, 2017).

As stated by Prajwal et al. (2021), lip reading, or visual speech recognition, involves recognizing speech silent video. This technology has practical applications, including enhancing speech recognition in noisy environments, enabling silent dictation, and transcribing archival silent films. It also has significant medical uses, such as helping individuals with conditions like Lou Gehrig's disease or aphonia communicate through lip movements.

Chapter 3

Methodology

Requirement Analysis

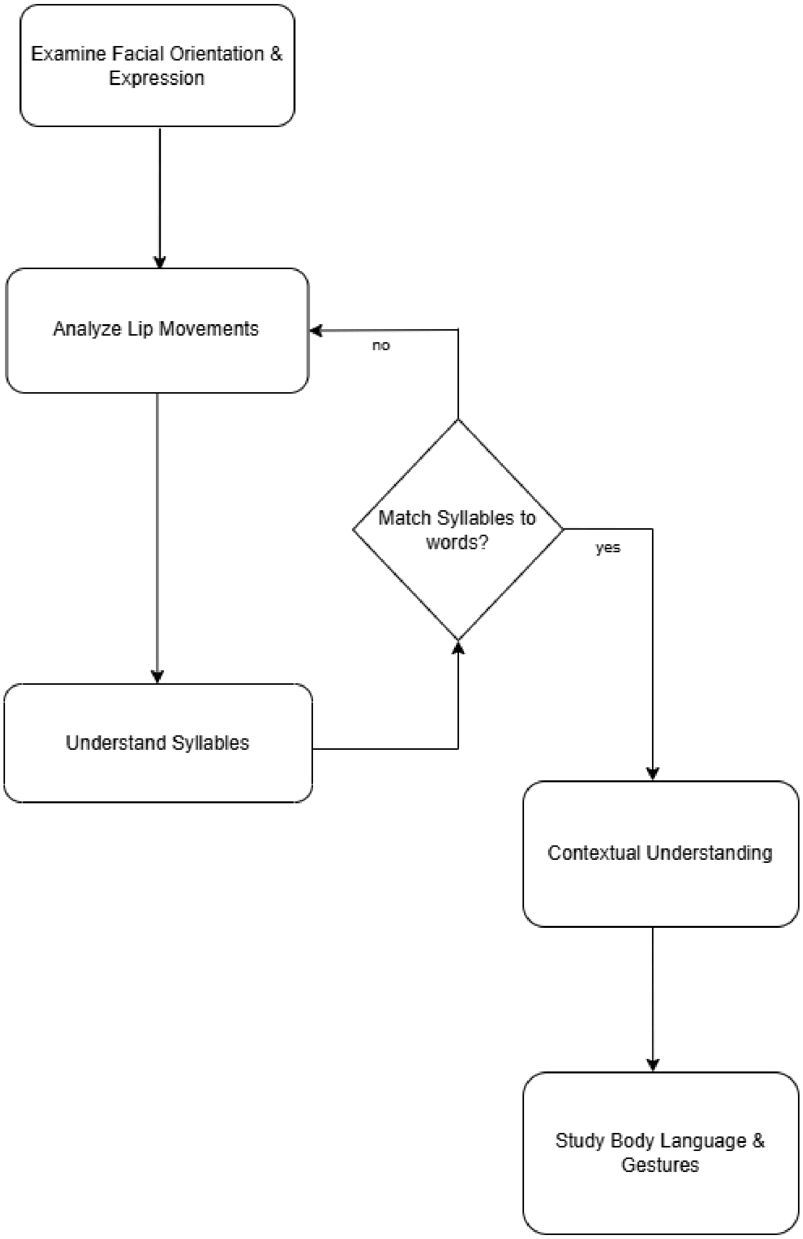
Examine Facial Orientation & Expression
Analyze Lip Movements
no
Match Syllables to words?
yes
Understand Syllables
Contextual Understanding
Study Body Language & Gestures

Figure 1

Traditional Lip-Reading Process

The traditional process of lip reading is a skill that requires significant time, effort, and consistent practice to develop. As shown in Figure 1, the process begins with observing the speaker's facial expressions, which provide important emotional and contextual cues. Following this, the individual analyzes the speaker’s lip movements to identify speech patterns. Prior to this, however, one must undergo structured learning and training in lip reading, as it is not an instinctive ability and demands effort to master. After examining the lips, the next step is to interpret each syllable being spoken. If the syllables are correctly identified, the process moves into contextual interpretation, where situational awareness is used to refine understanding and make educated guesses. Because many speech sounds appear visually similar, this step is crucial for accuracy. If the syllables are unclear, the person must return to reanalyze the lip movements. Finally, studying the speaker's body language and gestures provides additional clarity, enabling a fuller and more accurate comprehension of the spoken message. This structured approach demonstrates how traditional lip reading combines multiple visual cues and detailed observation to accurately interpret speech.

Requirement Documentation

Project LIF is a mobile-based educational tool developed to assist users in learning Filipino lip reading by recognizing lip movements through a mobile device's camera and translating them into Filipino words. It uses a deep learning model that combines 3D Convolutional Neural Networks (3D-CNN) and Bidirectional Long Short-Term Memory (BiLSTM) to analyze the spatial and temporal features of lip movements. Unlike most existing lip-reading technologies that are trained in English or other widely spoken languages, this system is specifically designed to support Filipino phonetics and syllables. This addresses the current gap in accessible, localized speech recognition tools, especially within Philippine classrooms and special education environments.

The application processes short recorded video clips to identify lip movement patterns and match them to pre-defined Filipino syllables. It is designed as a guided tutorial platform that supports structured practice, allowing users to learn and assemble syllables into words while receiving visual feedback. Rather than functioning as a real-time communication tool, Project LIF serves as a learning aid, helping users build foundational skills in visual speech recognition. This approach benefits various target users, including lip-reading learners, hearing- or speech-impaired individuals, and speech-language educators or therapists.

The system operates entirely on Android devices and does not require any external hardware, making it a practical and cost-effective solution for schools and learners, particularly in underserved areas. By focusing on 80 essential Filipino syllables, the application supports users in improving their articulation and recognition of common Filipino sounds. In doing so, Project LIF not only enhances speech training and language instruction but also aligns with inclusive education initiatives under Republic Act No. 7277 and the Department of Education’s SPED guidelines.

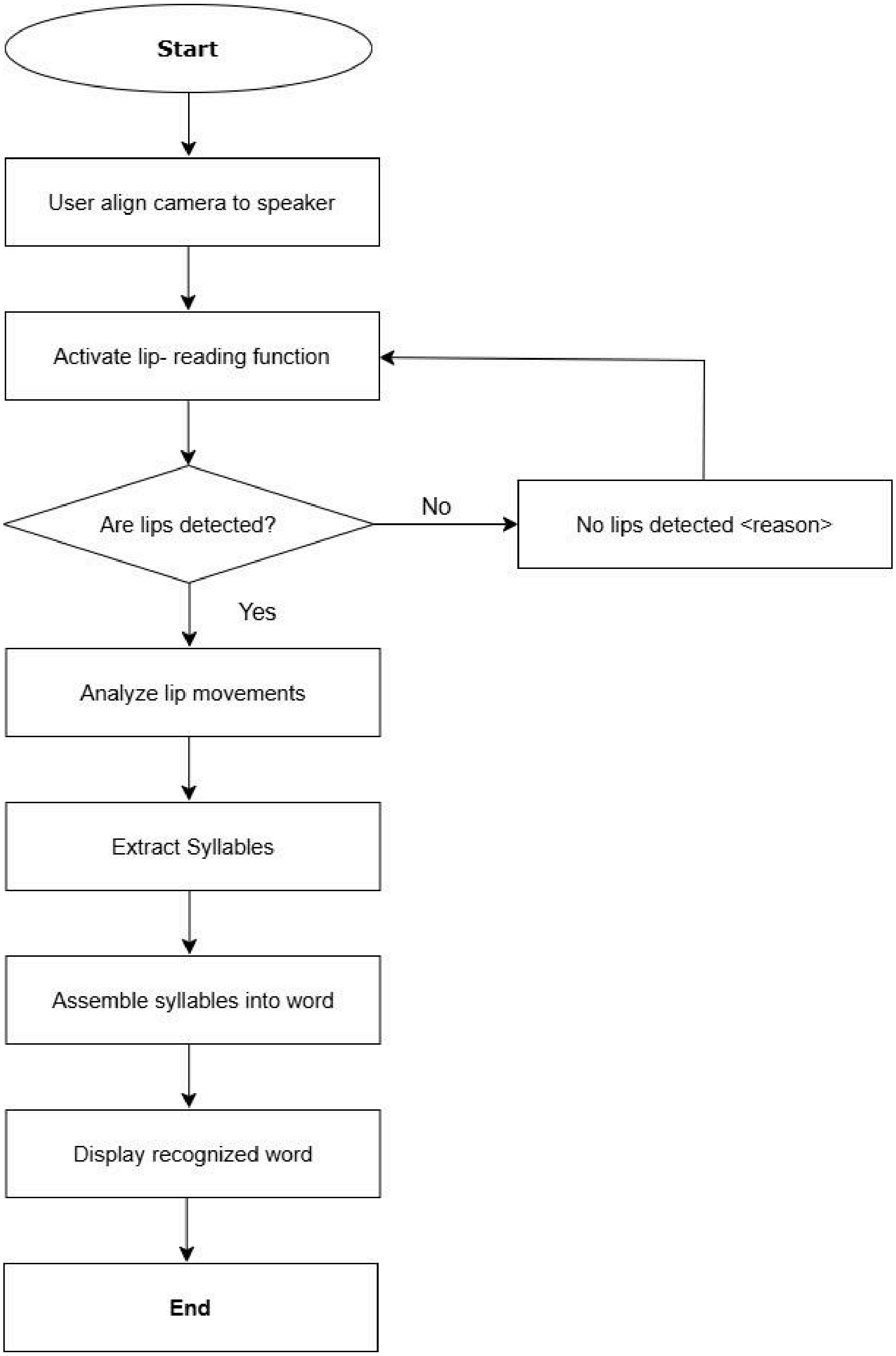
Start
User align camera to speaker
Activate lip- reading function
Are lips detected?
No
No lips detected <reason>
Yes
Analyze lip movements
Extract Syllables
Assemble syllables into word
Display recognized word
End

Figure 2

Flow Chart of Project LIF

The process begins when the user taps the camera toward the speaker, signaling the app to activate its lip-reading function. As the speaker begins to talk, the app uses the camera to visually detect and analyze lip movements strictly without relying on any sound. It first identifies syllables based on these movements, then assembles them into the most accurate word, which are displayed on the screen. However, if the app cannot detect lips-due to reasons like the speaker being too far, their mouth being covered (o.g.. by a hand, object, or mask), poor lighting, or the camera being misaligned-it will prompt the user with a message saying "No lips detected" followed by the specific reason (e.g., "Too far from camera" or "Mouth is obstructed"). This ensures users are clearly informed of what's preventing the detection, so they can adjust accordingly.

Functional Requirements

The functional requirements define how the “Project LIF” software system should operate, outlining its behavior and capabilities in response to specific inputs and user interactions. These requirements reflect the system's core educational purpose as a mobile-based learning aid for Filipino lip reading. The following are the key functional requirements:

Lip Movement Recognition: The application captures visual input through the device's camera and accurately detects lip movements corresponding to Filipino syllables. This process operates independently of audio, relying solely on visual cues to identify mouth patterns.

Translation Engine: Recognized lip movements are processed by a trained deep learning model, utilizing 3D Convolutional Neural Networks and Bidirectional Long Short-Term Memory, which maps them to corresponding Filipino syllables or words. The system then converts this recognition into readable Tagalog text.

Tutorial Module: Users can select specific syllables or sets of words to practice. The system guides learners through visual examples and exercises, supporting progressive learning of Filipino phonemes.

Sample Sentence Generation: Once the syllables are recognized and assembled into a word, the system automatically displays a simple sample sentence that uses the recognized word.

User Interface: The application presents translated text in a user-friendly interface that includes a tutorial dashboard, practice controls, and visual guidance to support a smooth and engaging learning experience.

Non-functional Requirements

Performance: Processes short recorded video clips quickly and provides accurate visual-to-text translations with minimal delay after input submission.

Usability: The system should provide a user-friendly interface with guided tutorials, visual feedback, and accessible navigation to accommodate users of all technical skill levels, including hearing and speech impaired students, and individuals with speech or hearing impairments.

Compatibility: Runs efficiently on major mobile platforms (eg. Android) and supports various smartphone models and resolutions, including mid-range cameras (e.g., 720p at 30fps).

Security: Implements strong data encryption to protect user privacy and ensure compliance with data protection regulations, especially for stored video data and translation history.

Technical Requirements

Technologies: Utilizes Python 3.8 for backend development, integrating deep learning libraries such as PyTorch and TensorFlow. For visual processing of lip movements, the system employs OpenCV, MediaPipe, DLib, and MTCNN. These tools handle tasks like facial detection, frame extraction, and lip region isolation.

Lip Reading Model: Replaces conventional speech recognition APIs with a visual speech recognition model that uses a hybrid architecture combining 3D Convolutional Neural Networks (3D-CNN) and Bidirectional Long Short-Term Memory (BiLSTM). The model is trained specifically to recognize 80 essential Filipino syllables from video input.

Database: Employs a secure cloud-based database to store user-generated data, visual translation outputs, and system feedback logs for performance analysis and Future training datasets.

Model Deployment: The trained models are converted into lightweight formats using TensorFlow Lite and PyTorch Mobile to allow efficient inference directly on Android devices with minimal latency.

Data Processing Tools: Uses MoviePy and NumPy for video trimming, frame processing, and pixel-level operations during model training and preprocessing stages.

Cloud & Training Platform: Model training is conducted on Google Colab, which provides access to cloud GPUs/TPUs for accelerated performance. Intermediate files and model checkpoints are managed through Google Drive for seamless access.

Data Gathering Tools

To ensure the accuracy and reliability of the model, the researchers will employ expert validation as a primary data gathering tool. Specifically, insights and evaluations will be gathered from professionals in the fields of Filipino linguistics and lip reading. Filipino linguists will assess the linguistic structure, phonetic nuances, and contextual correctness of the dataset and the model outputs, ensuring they align with natural Filipino language use. Concurrently, expert lip readers will validate the visual data, particularly focusing on the accuracy of lip movement interpretations and their correspondence to actual spoken Filipino words or phrases. Additionally, the researchers will conduct a comprehensive review of existing literature and studies related to lip reading across various fields including computer vision, speech recognition, and human-computer interaction to further refine the dataset and model. This triangulated approach will help establish both the technical soundness and cultural-linguistic relevance of the model, contributing to its effectiveness as a lip reading interpreter for the Filipino language.

Conceptual Framework

This study utilizes an input-process-output-feedback (IPO) framework to illustrate and analyze the system's components and their interactions. This model provides a visual representation of how the system functions as a learning tool. The IPO diagram serves as an effective tool for conceptualizing, assessing, and presenting the key elements of the application. By clearly mapping the flow of inputs, processes, outputs, and feedback, the framework offers a comprehensive understanding of how learners engage with the system to support continuous learning and opportunities for instructional improvement.

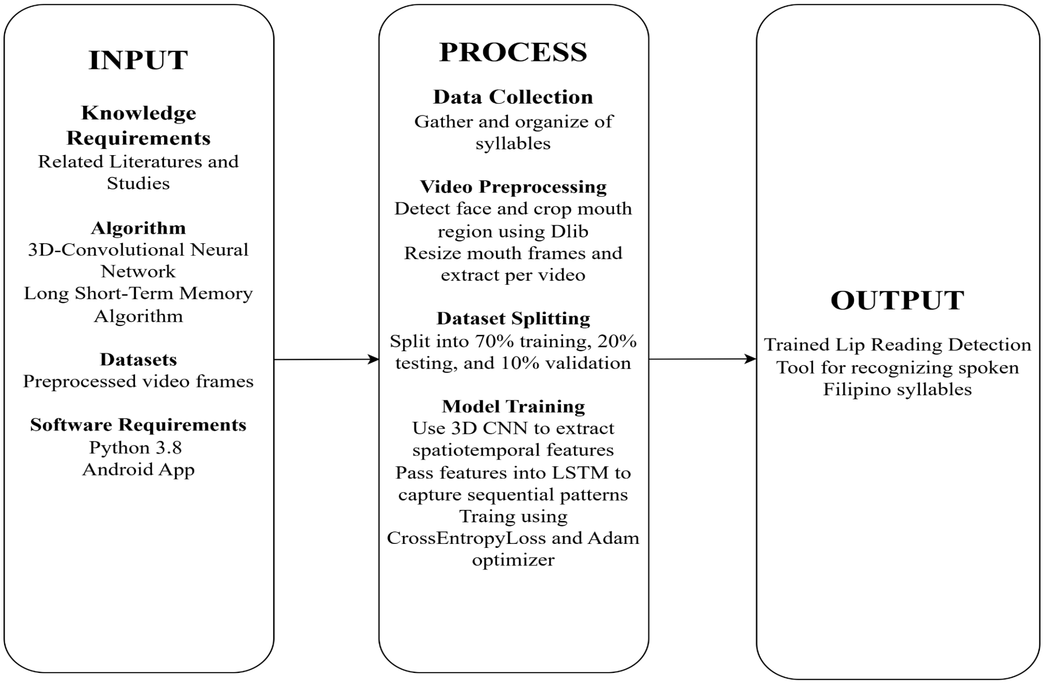
INPUT
Knowledge Requirements
Related Literatures and Studies
Algorithm
3D-Convolutional Neural Network
Long Short-Term Memory Algorithm
Datasets
Preprocessed video frames
Software Requirements
Python 3.8
Android App
PROCESS
Data Collection
Gather and organize of syllables
Video Preprocessing
Detect face and crop mouth region using Dlib
Resize mouth frames and extract per video
Dataset Splitting
Split into 70% training, 20% testing, and 10% validation
Model Training
Use 3D CNN to extract spatiotemporal features
Pass features into LSTM to capture sequential patterns
Traing using CrossEntropyLoss and Adam optimizer
OUTPUT
Trained Lip Reading Detection Tool for recognizing spoken Filipino syllables

Figure 3

Input-Process-Output Diagram

This project is designed as a learning tool to help the SPED students, specifically hearing and speech impaired students of Iba, Zambales to learn how to lip-read Filipino syllables. It uses artificial intelligence (AI) to recognize spoken syllables by analyzing the movement of the lips in video recordings. To support the development of this tool, the researchers reviewed related studies on deep learning, computer vision, and visual speech recognition.

The system uses a 3D Convolutional Neural Network (3D-CNN) to extract important features from video frames, and a Bidirectional Long Short-Term Memory (LSTM) model to understand how these features change over time. The model is trained using video clips of people pronouncing Filipino syllables. The software was developed using Python 3.8, and an Android app was created to make the tool easy to use for learners.

During data collection, videos of people saying different syllables were recorded. The system then processed these videos by detecting the mouth area using a tool called Dlib, cropping it, and resizing the images to make the data consistent. The model was trained using a CrossEntropy Loss function and the Adam optimizer. After training, the system can recognize spoken Filipino syllables just by analyzing a video of someone speaking. A user-friendly application was developed, where users can upload or record a video and see the predicted syllable.

This research aims to support the hearing and speech impaired students of Iba, Zambales by providing a tool that makes it easier to learn lip-reading in Filipino. It also lays the groundwork for future tools that can improve communication and education for people with hearing difficulties.

Development and Testing

The lip-reading technology focuses on creating an Al model that is capable of interpreting spoken Filipino words by recognizing and combining learned syllables. This system is designed to analyze lip movements in videos and predict what syllables are being spoken, which are then used to construct Filipino words. The syllable-based approach provides flexibility, and more diversity when it comes to understanding complete words from visual inputs, especially in the Filipino language where words are syllable-rich and phonetically regular.

These phases encompass the entire system lifecycle - from preprocessing all the data videos and training the model for evaluation and integration. The dataset will be divided into 70% for training, 20% for testing, and 10% for validation to ensure the model's accuracy and generalization.

The development phase is focused on implementing the model to recognize Filipino words from visual speech. The data pre-processing includes preparing the raw video data to undergo face detection and cropping, extraction of frames, and syllable-based labeling.

The model architecture includes the 3D Convolution Neural Networks (3D-CNN) to extract spatiotemporal features from video frame sequences. Long Short Term Memory was used during temporal modeling to enhance the model's ability to associate motion patterns with syllables. The output layer predicts the probability of cach syllable class which is then used to identify full words. For the model training and evaluation, the labeled syllable data is used, where the model learns visual patterns corresponding to each syllable through multiple epochs of training. In the evaluation metrics, syllable-level accuracy, precision, recall, and Fl-score are used. The word-level accuracy is calculated by comparing the predicted syllables with the actual spoken word. Dropout and early stopping are applied to prevent overfitting and improve generalization. The model integration includes taking a video input of the user, it will then detect the mouth region, extract the frame sequences, feed all of those frames into the model, and the model will then predict the syllable spoken. Lastly, it will combine the syllables into a final word prediction

Software Development Life Cycle (SDLC)

Agile Methodology

The Agile Method ensures structured and goal-oriented progress in the development of Project LIF. This approach enables iterative design, regular testing, and continuous refinement, which is essential when building the deep-learning-based model.

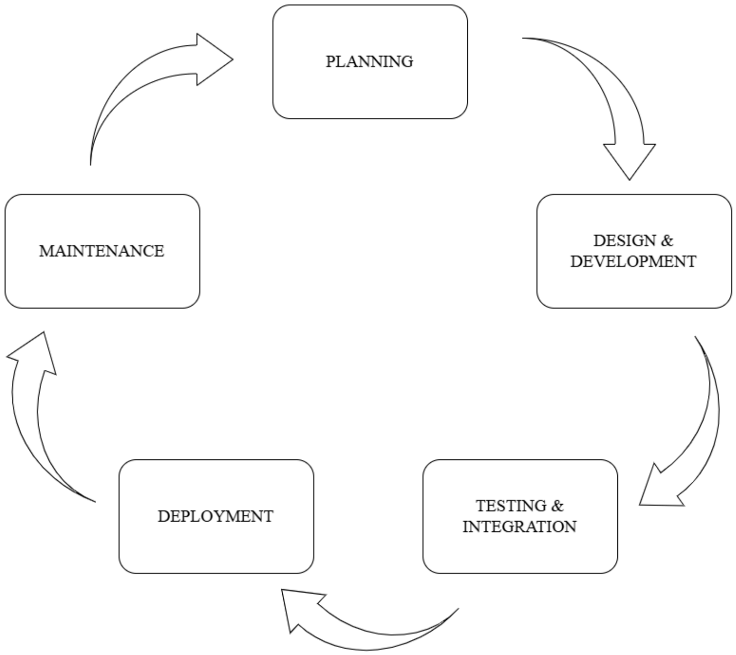

Figure 5

Agile Methodology

The system aims to recognize syllables and form meaningful Filipino words. Dlib is chosen for accurate mouth-region extraction, while a 3D Convolution Neural Network (ID CNN) with Bidirectional Long Short-Term Memory (LSTM) handles temporal modeling. The dataset is stored in Google Drive, with Google Colab used for training and testing.

Design and Development

This phase involves reconstructing and testing the system's modular architecture. Video preprocessing includes frame extraction and mouth cropping with Dlib. A 3D

CNN and Bidirectional LSTM model extract features and predict syllables, which are classified into categories and grouped into meaningful Filipino words.

Testing and Integration

This phase involves testing and integrating each component into the system. Unit testing verifies preprocessing steps like frame extraction and mouth cropping, while integration testing ensures a smooth transition from video input to syllable prediction and word reconstruction.

Deployment

During this phase, the system was prepared for deployment. A functional Android application was launched for lip-reading technology, and the prototype was finalized, enabling video input processing and word prediction.

Maintenance

In the final phase, the technology is optimized for future improvements and enhancements. The model is designed to seamlessly integrate additional speakers and new syllables. Additionally, the system remains flexible, supporting both input and inference settings for research or practical applications.

Dataset Validation

Validation process will involve three licensed and certified language experts qualified to evaluate the accuracy of the lip reading software. Their feedback will be used to refine the dataset before the implementation of the application. In addition, a panel of experts will review both the research documentation and the application software to ensure linguistic accuracy and functional reliability. The panel's recommendations will be integrated to improve the software's overall performance and usability.

Description of the Prototype

Project LIF: A Lip Reading Learning Tool for Basic Filipino Words using 3D-CNN and Bidirectional Long Short-Term Memory is an Android-based tutorial platform designed to teach Filipino lip-reading to hearing-impaired students. The prototype's primary purpose is to enhance language development by enabling users to practice lip-reading single words and a limited set of simple sentences formed from 80 basic Filipino syllables, including single-letter vowels (e.g., 'a', 'i'), two-letter consonant-vowel syllables (e.g., 'ba', 'ka'), and three-letter nasal syllables (e.g., 'nga', 'ngo'). The system provides an accessible educational tool for hearing-impaired students and their educators in controlled classroom settings.

The prototype consists of an Android mobile application for user interaction and a backend deep learning model for syllable recognition. Developed using Android Studio and React Native CLI, the application uses the device's camera to capture video input of lip movements and delivers tutorial content through an intuitive interface. The deep learning model integrates a 3D Convolutional Neural Network (3D-CNN) to extract spatiotemporal features from video frames and a Bidirectional Long Short-Term Memory (LSTM) network to model temporal sequences, ensuring accurate recognition of the 80 syllables (Inamdar et al., 2023). Video preprocessing employs Dlib's facial landmark detection to isolate the mouth region, standardizing input for the model, as validated in low-resource language applications (Kurniawan & Suyanto, 2020).

The prototype's primary function is to lip-read a spoken word or simple sentence using computer vision, analyzing lip movements to identify syllables and cross-reference them against a predefined dictionary of valid Tagalog words (e.g., 'tagahanga' [ta+ga+ha+nga], 'alimango' [a+li+ma+ngo], 'kapilipili' [ka+pi+li+pi+li]) formed exclusively from the 80 syllables, excluding invalid syllables like 'pag' or 'san'. If the syllables match a dictionary word or a supported sentence (e.g., 'Sana ako pa'), the app outputs the text, providing immediate feedback. To support lip-reading education, the prototype includes tutorial features tailored for hearing-impaired students (Prajwal et al., 2022). A tutorial mode allows users to guess the syllable shown in a raw video of a native speaker's pronunciation, with adjustable playback speed (e.g., slow, normal, fast) to aid learning. Additional features include visual feedback (e.g., highlighted lip movements, text captions), simple sentences (e.g., 'Sana ako pa'), large fonts, high-contrast visuals, intuitive navigation, and progress tracking for educators to monitor student performance. These features enhance accessibility and engagement in classroom settings, while the syllable-guessing tutorials complement the primary lip-reading function.

The dataset consists of raw video recordings of the 80 Filipino syllables, pronounced by diverse native speakers to capture varied lip movements, ensuring clear, tutorial-appropriate pronunciation samples (Kim et al., 2023). The dataset includes single-letter vowels, two-letter consonant-vowel syllables, and three-letter nasal syllables, enabling the formation of valid Tagalog words like 'tanga' and 'kapilipili'. Recorded at 720p, the videos undergo preprocessing, including brightness adjustment and mouth region cropping, to enhance clarity for educational and recognition purposes (Astorga et al., 2023). Detailed preprocessing and training procedures are discussed in subsequent Methodology sections.

Optimized for classroom settings, the prototype empowers hearing-impaired students to develop language skills through lip-reading practice. It serves as a teaching aid for their educators, supporting lesson planning and student assessment. By providing an AI-based tutorial tool for Filipino lip-reading, Project LIF addresses the educational gap in visual speech recognition for hearing-impaired learners.

Hardware and Software Requirements

The prototype of Project LIF: A Lip Reading Learning Tool for Basic Filipino Words using 3D CNN and Bidirectional Long Short Term Memory requires both hardware and software that can handle real-time video recording, preprocessing, and AI model integration. Since the project transitioned from Expo Go to a full native Android Studio build, the requirements have been updated to reflect the increased system demands.

Hardware Requirements

Initially, the mobile application was developed using Expo Go for lightweight testing. However, due to the limitations of Expo's managed workflow and the need for direct access to native modules (such as camera controls and model inference), the project has since transitioned to a full native Android Studio development environment. As a result, the projected hardware requirements have been adjusted to ensure compatibility with the higher resource demands of native apps, especially those involving real-time video capture and on-device AI processing.

The estimated minimum and recommended hardware specifications are outlined below:

Component	Minimum Requirement	Recommendation Specification
Operating System	Android 10	Android 11 or Higher
RAM	3 GB	4 GB or Higher
Processor	Octa-core ARM chipset like Snapdragon 662 or better	Snapdragon 720G, Dimensity 800U, or better
Camera	720p video resolution at 30 fps	1080p at 30-60 fps or higher
Internal Storage	500 MB free space	1 GB or more free space
Screen Size	5 inches, 720x1280 resolution	6 inches, Full HD+ (1080x2340) resolution
Internal Connection	Required during development and updates	Stable WI-FI or mobile data

Battery Life	Standard	Long-lasting battery or external power support

Software Requirements

The mobile application is now developed natively using React Native CLI combined with Android Studio as the primary integrated development environment (IDE). This move allows the developers to directly access Android's native modules and optimize app performance for real-world deployment. These software components enable efficient video capture, processing, and syllable recognition, ensuring the application supports hearing-impaired students in practicing lip-reading effectively in classroom settings.

The key software components include:

Software Tool	Purpose
React Native CLI	Needed because you moved out of Expo Go; Android Studio builds require CLI-based setup.
Android Studio	Main IDE for building, testing, and deploying Android apps natively.
Java Development Kit (JDK)	Required by Android Studio and Gradle to compile the app
Android Gradle Plugin	Needed to handle app dependencies, linking native modules like Camera APIs.
Python	Language for training your AI model (3D-CNN + LSTM).

Pytorch/ Tensorflow	Frameworks to build, train and deploy deep learning models.
OpenCV / MediaPipe / MTCNN / Dlib	For video frame processing, face detection, mouth region extraction.
TensorFlow Lite / Pytorch Mobile	For shrinking and running AI models directly on Android devices.
Google Colab / Google Drive	For cloud-based training (Colab) and storing large datasets (Drive).
MoviePy	For Automatic splitting and preprocessing of video files

References:

Deocampo, J. P., Villarica, M. M., & Vinluan, J. M. (2023). *Improving lip-reading accuracy in Tagalog using a hybrid CNN-LSTM model*. [Publication details if available].

Crespo, M., & Recario, R. (2023). *Lip-reading Filipino words using CNN and RNN deep learning models*.

Lu, Y., & Li, H. (2019). Automatic Lip-Reading System based on deep convolutional neural network and Attention-Based Long Short-Term memory. *Applied Sciences*, *9*(8), 1599. https://doi.org/10.3390/app9081599

Winoto, S. H. (2018). *Lip reading using deep learning*. [Master's thesis, Blekinge Institute of Technology]. DiVA Portal.

Sarhan, A., Elshennawy, N., & Ibrahim, R. (2021). Hybrid lip reading network based on deep convolutional neural network. *Multimedia Tools and Applications, 80*(20), 30845–30860. https://doi.org/10.1007/s11042-021-10961-1

Garg, S., Noyola, M., & Bagadia, A. (n.d.). *Deep Learning Approaches for Silent Speech Recognition*.

Shrestha, A. (n.d.). *Automated lip-reading using convolutional neural networks*. [Unpublished manuscript].

Inamdar, A., Sundarr, S., Khandelwal, S., & Ajeyprasaath, P. (2023). *Deep learning based lip reading system using 3D CNN and BiLSTM. International Journal of Advanced Computer Science and Applications, 14*(2), 227–234. https://doi.org/10.14569/IJACSA.2023.0140226

Ma, C., Wang, Y., Shen, Z., et al. (2021). Densely connected temporal convolutional network for lip reading. *Pattern Recognition Letters, 148*, 122–128. https://doi.org/10.1016/j.patrec.2021.05.021

Putcha, S., Rajam, L. S., Sugamya, B., & Gopala, K. R. (2024). Deep learning-driven lip reading and text translation. *Journal of Artificial Intelligence Research, 73*, 456–472.

Jeon, H., Elsharkawy, M., & Kim, H. (2022). Deep lip reading model using multi-scale spatiotemporal feature learning. *Sensors, 22*(24), 9637. https://doi.org/10.3390/s22249637

Fenghour, M. (2021). *Lexicon-free visual speech recognition using visemes for sentence prediction* [Doctoral dissertation, University of East Anglia]. UEA Digital Repository.

Deocampo, J., Villarica, S., & Vinluan, R. (2023). A hybrid CNN-LSTM model for Filipino lip reading. *Philippine Computing Journal, 18*(1), 50–66.

Crespo, A., & Recario, C. (2023). Lip reading Filipino words using deep learning models. *International Journal of Advanced Research in Artificial Intelligence, 12*(3), 32–39.

Parekh, S., Gupta, R., Chhatpar, Y., Yash, & Kulkarni, V. (2018). Lip reading using convolutional autoencoders and LSTM. *International Journal of Innovative Research in Computer and Communication Engineering, 6*(3), 2906–2913.

Afouras, T., Chung, J. S., & Zisserman, A. (2018). Deep lip reading: A comparison of models and an online application. *Interspeech 2018*, 3514–3518. https://doi.org/10.21437/Interspeech.2018-1737

Ambati, P. (2024). Deep learning methods for visual speech recognition. *Journal of Artificial Intelligence Research and Development, 12*(1), 89–104.

Tsai, C. Y., Tseng, H. Y., & Ruan, S. J. (2024). An efficient lip reading model based on CNN and attention-based LSTM. *Applied Sciences, 14*(4), 1235. https://doi.org/10.3390/app14041235

Lu, X., & Li, Z. (2019). Research on lip reading method based on deep learning. *Journal of Physics: Conference Series, 1187*, 052053. https://doi.org/10.1088/1742-6596/1187/5/052053

Hao, Y., Mamut, A., Yadikar, M., Aysa, A., & Ubul, K. (2020). A review of deep learning methods for lip reading. *EURASIP Journal on Image and Video Processing, 2020*(1), 1–17. https://doi.org/10.1186/s13640-020-00516-0

Fenghour, M., Chen, T., Guo, D., Li, C., & Xiao, J. (2021). An overview of advances in visual speech recognition. *IEEE Access, 9*, 15991–16008. https://doi.org/10.1109/ACCESS.2021.3051846

Kurkova, P., Valkova, H., & Scheetz, N. (2011). Factors impacting participation of European elite deaf athletes in sport. *Journal of Sports Sciences, 29* (6), 607-618. https://doi.org/10.1080/02640414.2010.548821

Scheetz, N. A. (2004). *Psychosocial aspect of deafness.* Pearson Education.

Wang, S.-Y., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2020). CNN-generated images are surprisingly easy to spot... for now. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7.*

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. *Proceedings of the IEEE International Conference on Computer Vision,* 1–11.

Bayar, B., & Stamm, M. C. (2016). A deep learning approach to universal image manipulation detection using a new convolutional layer. *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security,* 5–10.

Katsaggelos, K., Bahaadini, S., & Molina, R. (2015). Audiovisual fusion: Challenges and new approaches. *Proceedings of the IEEE, 103* (9), 1635-1653.

Akbari, H., Arora, H., Cao, L., & Mesgarani, N. (2018). LIP2AUD-SPEC: Speech reconstruction from silent lip movements video. *Proceedings of ICASSP 2018,* 2516–2520.

- Masi, I., Killekar, A., Mascarenhas, R. M., Gurudatt, S. P., & AbdAlmageed, W. (2020). Two-branch recurrent network for isolating deepfakes in videos. *arXiv preprint arXiv:2008.03412.*

Ryumin, D., Ivanko, D., Axyonov, A., Karpov, A., Kagirov, I., & Zelezny, M. (2019). Human-robot interaction with smart shopping trolley using sign language: Data collection. *IEEE 1st International Workshop on Pervasive Computing and Spoken Dialogue Systems Technology (PerDial 2019).*

Gallaudet Research Institute. (2006). *Regional and national summary report of data from the 2006–2007 Annual Survey of Deaf and Hard of Hearing Children and Youth.* Gallaudet University.

Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J. M., & Brumberg, J. S. (2010). Silent speech interfaces. *Journal on Speech Communication, 52,* 270–287.

Pu, J., & Wang, S. (2022). *A review of machine lip-reading: Deep learning techniques and dataset challenges*. [Publication details needed if available].

Berkol, S., Cetin, O., Gokberk Cinbis, R., & Demir, O. (2022). *Visual lip-reading datasets in Turkish from real-world video sources*. [Publication details needed if available].

Kurniawan, B., & Suyanto, S. (2020). *Syllable-based lip reading using 3D deep learning for the Indonesian language*. [Publication details needed if available].

Astorga, J. J., Malapit, R. J., & Santos, J. R. (2023). *Phonological similarities between Tagalog and Bahasa Indonesia: A phonology as human behavior approach*. [Publication details needed if available].

Kim, J., Park, S., Lee, J., & Kim, G. (2023). *Memory-augmented decoder for low-resource language lip reading*. [Publication details needed if available].

Prajwal, K. R., Mukhopadhyay, R., Namboodiri, V. P., & Jawahar, C. V. (2022). *Improving lip reading with sub-word units and visual attention*.