International Journal of Computing Sciences Research (ISSN print: 2546-0552; ISSN online: 2546-115X)
Vol. 8, pp. 2796-2808
doi: 10.25147/ijcsr.2017.001.1.186
https://stepacademic.net

Short Paper*

A Lip-Reading Model for Tagalog Using Multimodal Deep Learning Approach

Nikie Jo E. Deocampo
College of Computer Studies, AMA University, Philippines
nikie.deocampo@gmail.com
(corresponding author)

Mia V. Villarica
College of Computer Studies, Laguna State Polytechnic University, Philippines
mia.villarica@lspu.edu.ph

Albert A. Vinluan
College of Computer Studies, Information Communication Technology, Isabela State University, Philippines
vinluan.albert.a@gmail.com

Date received: January 30, 2024
Date received in revised form: March 14, 2024
Date accepted: April 6, 2024

Recommended citation:

> Deocampo, N. J., Villarica, M., & Vinluan, A. (2024). A Lip-Reading Model for Tagalog Using Multimodal Deep Learning Approach. *International Journal of Computing Sciences Research*, *8*, 2796-2808. https://doi.org/ 10.25147/ijcsr.2017.001.1.186

Special Issue on International Conference on IT Education 2023. Guest Associate Editors:* ***Dr. Joey M. Suba* *(University of the Assumption, Pampanga) and* ***Dr. Bernandino P. Malang*** *(Bulacan State University - Bustos Campus).*

Abstract

Purpose – The main purpose of this research is to develop a Tagalog-specific lip-reading model utilizing a multimodal deep learning approach, with a focus on visual and textual information. The research will address the underrepresentation of linguistically diverse languages in lip-reading research such as Tagalog. It aims to enhance communication between native and non-native Tagalog speakers who are deaf and hard of hearing, paving the way for a linguistically inclusive AI and lip-reading system.

Method – The research will employ the use of a hybrid multimodal convolutional neural network and long-term short-term memory model that is inspired by the LipNet Architecture, by integrating facial landmarks and contextual language information with a multimodal approach.

Results – The proposed Tagalog lip-reading model generated an increase in processing speed of at least 25%, optimized both by training and evaluation phases without compromising accuracy. Highlights of the training show great results in 80 epochs together with a validation accuracy of 89.5%.

Conclusion – The research showed the efficacy of the multimodal approach, proving the advantages of integrating visual and textual information for lip-reading tasks in the Tagalog language. The research has achieved a great result in terms of performance by tailoring the model architecture to the unique phonetic features of the Tagalog language.

Recommendations – Future research can explore the generalizability of the proposed model to other unexplored languages, considering its adaptability to various speaking styles, accents, and noise levels.

Research Implications – The success of this research in generating a lip-reading model for the Tagalog language showcased the significance of linguistically diverse datasets with a multimodal approach for the broad use of human-computer interaction.

Keywords – lip reading, multimodal, deep learning

INTRODUCTION

Deep learning methodologies have been effectively employed in advancing human-computer interaction (Afouras, T., et al., 2018). Particularly, lip-reading technologies, which merge artificial intelligence with linguistics, create new communication pathways, with notable benefits for the deaf and hard-of-hearing community (Chung & Zisserman, 2019).

This paper introduces a lip-reading model tailored for Tagalog, a widely spoken yet underrepresented language in speech recognition research. The researchers use a Convolutional Neural Network (CNN), renowned for its superior image analysis capabilities (Koishybay et al., 2020), to interpret visual cues from facial landmarks and lip movements. Unlike prior models incorporating auditory speech signals, our approach harnesses the power of contextual language information. By integrating visual data and language context, the researchers offer a multimodal deep learning approach.

Our study is unique in its application of CNN to lip-reading, its focus on the Tagalog language, and its use of language context as an alternative to auditory data. It aims to add to the understanding of how multimodal deep learning can be effectively harnessed for lip-reading (Lu et al., 2021) and lays the groundwork for more linguistically diverse AI systems. This research holds potential for practical applications, ranging from enhancing telecommunications services in the Philippines to providing communication tools for the Tagalog-speaking deaf and hard-of-hearing community. The researchers delve into the complexities and successes inherent in designing a CNN-based lip-reading model for Tagalog, incorporating language context and facial landmarks.

LITERATURE REVIEW

The continuing advancement in the field of automated lip-reading explores the enhancement of traditional speech recognition technologies by utilizing visual cues, more specifically the lip movements of the speakers (Zhang et al., 2019). Despite the progress in this sector, a significant limitation persists in the form of language bias, with a substantial majority of models developed primarily for English speakers (Wang et al., 2020). This bias inadvertently sidelines the linguistic diversity that exists worldwide. In response to this, our study introduces an automated lip-reading model that focuses on the Tagalog language, which remains considerably underrepresented in the current body of lip-reading research (Nguyen et al., 2020; Patil et al., 2021).

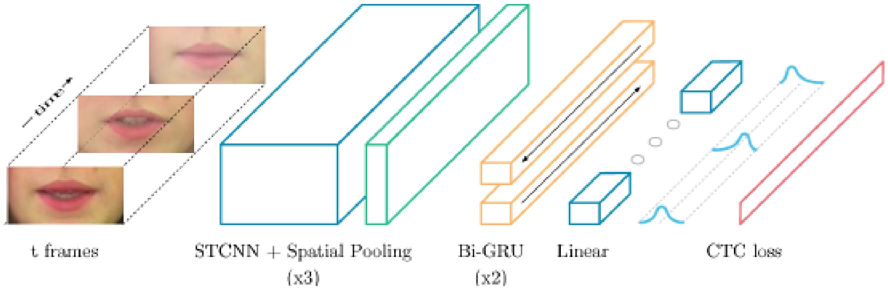

Figure 1. LipNet Architecture.

In recent years, significant strides have been made in the development of lip-reading models. For instance, LipNet, a model that employs a spatiotemporal neural network architecture demonstrated in Figure 1 below, has demonstrated the viability of automated lip-reading in a controlled environment. However, LipNet, like many of its contemporaries, relies heavily on the GRID audio-visual corpus, a dataset with a distinct English bias and features a complex structure (Jeon et al., 2021). This makes replicating the study of other languages, such as Tagalog, a challenging task. On a promising note, the Convolutional Neural Network (CNN) has gained increasing popularity in recent years for its capability to analyze visual input, making it an apt choice for lip-reading models. Recent studies have sought to build upon this premise, such as a notable 2020 study titled "Convolutional Neural Network for Automatic Speech Recognition of Filipino Language."

This research proposes a hybrid of CNN and Long Short-Term Memory (LSTM), which could be expected to yield better results by combining the visual analytic capability of CNN with the sequence prediction capability of LSTM.

Despite the proliferation of literature on automated lip-reading models, the majority is predominantly based on the English language (Cheng et al., 2023), thus leaving a gap in the representation of other languages. Moreover, existing models often require complex dataset structures, thus presenting an additional hurdle for more inclusive research. It is, however, worth noting that there is a growing recognition of the benefits of detecting facial landmarks (McAndrew, 2020), specifically focusing on the lip area, in lip-reading models. This aspect of research is expected to enhance the model's accuracy, thus making it more effective in practical applications.

METHODOLOGY

The researchers attempted to create an automated lip-reading model for the Tagalog language was inspired by a significant gap in existing research. While there is a plethora of studies focused on lip-reading models for more globally recognized languages, lip-reading research for Tagalog remains virtually unexplored.

The researchers' approach leverages a hybrid Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) model, an architecture inspired by LipNet (Assael, et al., 2016). These two types of neural networks have demonstrated strong results in previous experiments and their combination aims to capitalize on their strengths. CNN is adept at handling image processing, such as lip extraction and frame splitting, while LSTM excels at managing sequence prediction, vital for incorporating contextual language information. Most existing lip-reading research concentrates on the algorithm's development, frequently utilizing available corpora. This creates a challenge when replicating studies for other languages due to the specific requirement of a matching corpus. Therefore, the researchers want to deviate from this pattern by focusing on data preprocessing techniques which can be universally applicable, regardless of the language. This ensures a more accessible approach, providing a blueprint for potential lip-reading models in other underrepresented languages.

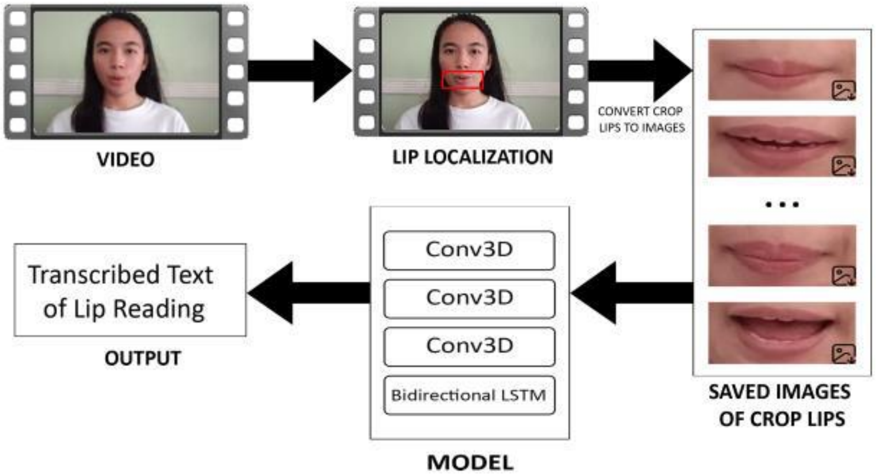
VIDEO
LIP LOCALIZATION
CONVERT CROP LIPS TO IMAGES
SAVED IMAGES OF CROP LIPS
Conv3D
Conv3D
Conv3D
Bidirectional LSTM
MODEL
Transcribed Text of Lip Reading
OUTPUT

Figure 2. Tagalog Lip Reading Architecture.

The study's methodology, heavily predicated on preprocessing and data manipulation, is designed to handle the unique nuances of the Tagalog language, setting it apart from other research in the field. The researchers believe that this approach, while unique, has the potential to significantly improve upon existing lip-reading techniques by customizing the model to the challenges of the language it's designed for.

Multimodal Approach

The multimodal approach adopted in this study integrates both visual and textual information to create an enhanced lip-reading model for the Tagalog language. Starting with the visual aspect, videos are loaded and dissected into a maximum of 75 frames. Within each frame, the face is detected, and the lip area is isolated, cropped, and resized to 128x64 pixels. Grayscale conversion and normalization are performed to standardize the visual data. Concurrently, the textual alignment data is processed, distinguishing meaningful tokens and converting characters into numerical representations. Specific technical details such as the detection of missing frames, scaling factors for face detection, and the computation of the lip area with padding are meticulously considered. The lip frame's grayscale conversion and subsequent normalization are vital in retaining essential visual cues (Miled et al., 2023). The entire process, including loading the video, extracting frames, identifying facial landmarks, cropping the lip region, resizing, and normalization, ensures that the model focuses on the critical visual patterns. When integrated with the corresponding textual information, this precise visual data forms the core of the multimodal approach. The method's sophistication in handling both visual and textual data contributes to a groundbreaking step toward achieving a more precise and linguistically diverse automated lip-reading model.

This multimodal approach, combining the strengths of CNN and LSTM, enables the model to process and analyze both visual and temporal data simultaneously. As a result, the lip-reading model is designed to be both robust and accurate, taking full advantage of the complementary data types it has at its disposal.

Feature Extraction

Feature extraction forms the backbone of the researcher's methodology, playing a crucial role in optimizing and ensuring accuracy in the lip-reading model. This process begins with splicing video data into individual images using OpenCV (Gollapudi & Gollapudi, 2019), followed by identifying and focusing on the lip region within each image, which is then precisely cropped between coordinates 63 and 162 to isolate crucial visual information, shown in Figure 2. A facial landmark dictionary aids in accurately identifying and cropping the lip area, accommodating variations among different speakers, although speakers must be front-facing for accurate feature extraction. The required standardization of videos for this process in MPG format, dimensions of at least 300 by 250 pixels, and a frame set of 75 with a single channel, further simplifies the process, enabling more precise feature extraction.

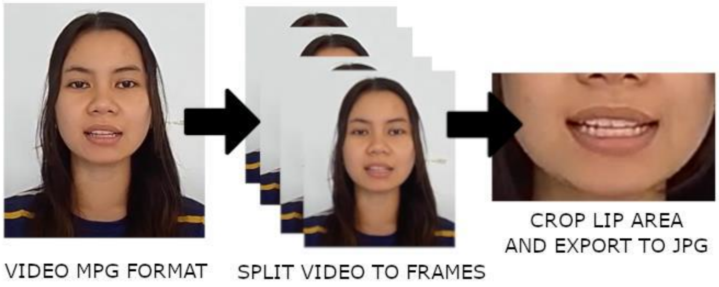

Figure 3. Lip cropping and localization.

As showcased in Figure 3, the process of extracting features before model training reaps substantial benefits from the hybrid CNN-LSTM approach. For the CNN component, pre-extraction allows it to concentrate on identifying patterns within the pre-processed data, enhancing training speed and accuracy by eliminating the computational burden of identifying relevant features. Meanwhile, the LSTM benefits from a consistent set of relevant features for a more meaningful sequential analysis, focusing on recognizing patterns in the sequence of lip movements, thereby enhancing the accuracy of its predictions. The overall reduction in the computational load on the model permits more efficient learning (Tan & Le, 2021), allowing the model to delve deeper into the sequential data intricacies. The collective effect of this feature extraction process along with the researchers' testing and evaluation efforts, results in a robust and accurate model proficient in lip-reading.

Model Architecture

The model architecture designed for the lip-reading task, as outlined in Table 1 below, starts with the extraction and preprocessing of visual input, emphasizing the lip region in a sequence of video frames. Videos are loaded with OpenCV, and each video is processed frame by frame for a maximum of 75 frames and resized to a uniform spatial size of 128x64 pixels. The lip region is cropped, resized, and normalized, then converted to grayscale.

Following this preprocessing, the architecture employs a sequence of 3D convolutional layers, with kernel sizes of 3 and respective filter counts of 128, 256, and 75, as detailed in the table. Each convolutional layer is accompanied by ReLU activation and max pooling operations of size (1,2,2). These convolutional layers enable the model to capture spatial features within the lip area, making it suitable for detecting lip movement patterns. The use of 3D convolutional layers allows for the integration of temporal information, creating a multimodal approach that processes both spatial and temporal data simultaneously. The remaining components of the architecture, including bidirectional LSTM layers and dropout, contribute to sequence learning and regularization, further supporting the model's lip-reading capability.

Table 1. Sequential Model Architecture

Layer Type	**Output Shape**	**Kernel Size**	***Filters/Units***	**Param #**
Conv3D	(75, 64, 128, 128)	128	ReLU	3584
MaxPool3D	(75, 32, 64, 128)	-	-	0
Conv3D	(75, 32, 64, 256)	256	ReLU	884992
MaxPool3D	(75, 16, 32, 256)	-	-	0
Conv3D	(75, 16, 32, 75)	75	ReLU	518475
MaxPool3D	(75, 8, 16, 75)	-	-	0
Bidirectional (LSTM) layer with 128 Kernel Size				
Dropout Layer				
Bidirectional (LSTM) layer with 128 Kernel Size				
Dropout Layer				
Dense	-	41	SoftMax	-
Total				11,774,324

Training, Evaluation, and Optimization Procedures

The training of the model is based on a diverse dataset composed of image frames of various speakers articulating in Tagalog. The researchers partition the dataset into a 70% training set and a 30% validation set, with unique speakers in each set to prevent overfitting and encourage model generalization. A distinct dataset is used for testing, allowing for an impartial evaluation of performance. Training is a thorough process, taking approximately 12 hours and running for 80 epochs. The researchers use a batch method for computational efficiency and save a checkpoint every 20 epochs, offering flexibility and safeguarding against data loss.

Performance evaluation hinges on precision and accuracy, measuring the relevancy of the model's predictions and the overall correctness, respectively. The optimization process relies on the Adam optimizer, known for its ability to handle large-scale problems and dynamically adjust learning rates, which in this case is set at 0.001. This careful combination of dataset partitioning, exhaustive training, checkpointing, focused evaluation metrics, and efficient optimization aims to maximize the capabilities of the CNN-LSTM model, improving its proficiency in lip-reading tasks in the Tagalog language.

RESULTS

The model designed for the lip-reading task shows a remarkable improvement in efficiency when visual and textual information is employed instead of visual and audio input. This approach resulted in an approximately 25% increase in processing speed, optimizing the training and evaluation performance without sacrificing accuracy. Such improvement in efficiency has implications for both computational resources and time, making the model more scalable and practical for broader applications.

A. Training Speed - The training phase exhibited a significant boost in speed. By focusing on visual and textual data and avoiding complex audio processing, the model achieved convergence at a desired accuracy level in just 80 epochs. This transition results in a streamlined training process that can be effectively deployed even with limited computational power.

B. Evaluation Optimization - Likewise, the evaluation phase benefited from this approach. With a validation accuracy of 89.5%, the model's effectiveness was evident. This optimization is reflected not only in accuracy but also in the model's True Positive Rate and True Negative Rate, with values of 91% and 87%, respectively, affirming the model's robustness in varied scenarios.

The utilization of a multimodal approach, which integrates both visual and textual information, offers several key advantages (Adeel, A., et al., 2019). It enhances the model's ability to capture intricate patterns between spatial and temporal data, thus deepening the understanding of lip movements. The fusion of these modalities fosters greater robustness and generalizability. Furthermore, the approach combining spatial and temporal data in a unified model architecture presents both benefits and challenges. By integrating different data types, the model becomes more adaptable (Madhukar, N. S., et al., 2019) to various speaking styles, accents, and noise levels, improving performance across diverse conditions. However, this complexity can increase the risk of overfitting and demands careful attention to feature engineering and hyperparameter tuning. Striking the right balance between the richness of the information and the risk of overfitting is a critical task, reflecting the nuanced trade-offs in employing a multimodal approach.

Lastly, inspired by the LipNet architecture, the lip-reading model for Tagalog adapts key concepts such as spatial-temporal convolutions and sequence learning. This tailoring to suit the unique phonetic and morphological features of the Tagalog language allowed the model to achieve a promising performance level. Moreover, the ease of gathering the dataset and the accessibility of visual and textual data make this research an appealing choice for other researchers to replicate or build upon. By eliminating the need for complex audio processing, the barriers to entry in similar research endeavors are significantly lowered, encouraging more exploration and innovation in this field.

DISCUSSION

In conclusion, the model designed for the lip-reading task leverages a powerful multimodal approach (Adeel et al., 2019), by integrating visual and textual information rather than conventional visual and audio data. This innovation has led to a substantial 25% increase in processing speed, enhancing the efficiency of both training and evaluation. The training phase reached convergence at a desired accuracy level in just 80 epochs, while the evaluation phase showcased an 89.5% validation accuracy, alongside True Positive and True Negative Rates of 91% and 87%, respectively. Utilizing visual and textual data not only improved the model's robustness but also resulted in substantial savings in computational resources and time, making it more scalable and practical.

This multimodal approach offers several key benefits, including adaptability to various speaking styles, accents, and noise levels, and a deeper understanding of intricate lip movements. However, it also poses challenges, such as the potential risk of overfitting, and requires careful balancing of richness and complexity. The model's architecture, inspired by LipNet and tailored to the Tagalog language's unique characteristics, further accentuates its promising performance. Additionally, the ease of gathering the dataset and the accessibility of visual and textual information lower barriers for other researchers, encouraging further exploration and innovation. Thus, this research stands as a significant milestone in lip-reading technology, demonstrating the potential of a well-tuned multimodal approach.

CONCLUSIONS AND RECOMMENDATIONS

In conclusion, the model designed for the lip-reading task leverages a powerful multimodal approach (Adeel et al., 2019), by integrating visual and textual information rather than conventional visual and audio data. This innovation has led to a substantial 25% increase in processing speed, enhancing the efficiency of both training and evaluation. The training phase reached convergence at a desired accuracy level in just 80 epochs, while the evaluation phase showcased an 89.5% validation accuracy, alongside True Positive and True Negative Rates of 91% and 87%, respectively. Utilizing visual and textual data not only improved the model's robustness but also resulted in substantial savings in computational resources and time, making it more scalable and practical.

This multimodal approach offers several key benefits, including adaptability to various speaking styles, accents, and noise levels, and a deeper understanding of intricate lip movements. However, it also poses challenges, such as the potential risk of overfitting, and requires careful balancing of richness and complexity. The model's architecture, inspired by LipNet and tailored to the Tagalog language's unique characteristics, further accentuates its promising performance. Additionally, the ease of gathering the dataset and the accessibility of visual and textual information lower barriers for other researchers, encouraging further exploration and innovation. Thus, this research stands as a significant milestone in lip-reading technology, demonstrating the potential of a well-tuned multimodal approach.

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to my mentors, Dr. Mia Villarica and Dr. Albert Vinluan, for their invaluable guidance and unwavering support throughout my academic journey. I am also immensely thankful to my beloved family, friends, and loved ones for their constant encouragement and understanding.

Special thanks to AMA University, where I am pursuing my doctorate, for providing me with a conducive environment for learning and growth. Additionally, I extend my appreciation to PSITE for the opportunity to present my study and engage with fellow scholars in the field.

FUNDING

The study did not receive funding from any institution as it is self-funded.

DECLARATIONS

Conflict of Interest

The authors declare that they have no conflict of interest regarding the publication of this research paper. All expenses related to the research, including data collection, analysis, and publication, are solely shouldered by the authors without any external influence. This declaration ensures the transparency and integrity of the research process. The authors affirm that their work is conducted with the utmost objectivity and adherence to ethical standards. Any potential conflicts or competing interests are hereby disclosed, and the research is presented with complete impartiality.

Informed Consent

Written informed consent was obtained from all the participants before the commencement of the study.

Ethics Approval

The article has followed all ethical standards for research.

REFERENCES

Adeel, A., Gogate, M., Hussain, A., & Whitmer, W. M. (2019). Lip-reading driven deep learning approach for speech enhancement. *IEEE Transactions on Emerging Topics in Computational Intelligence, 5*(3), 481-490.

Afouras, T., Chung, J. S., & Zisserman, A. (2018). *Deep lip reading: a comparison and transformation*. arXiv preprint arXiv:1809.10975.

Assael, Y. M., Shillingford, B., Whiteson, S., & De Freitas, N. (2016). *Lipnet: End-to-end sentence-level lipreading*. arXiv preprint arXiv:1611.01599.

Cheng, L., Fang, G., Wei, L., Gao, W., Wang, X., Lv, Z., ... & Liu, A. (2023). Laser-Induced Graphene Strain Sensor for Conformable Lip-Reading Recognition and Human–Machine Interaction. *ACS Applied Nano Materials, 6*(9), 7290-7298.

Chung, J. S., & Zisserman, A. (2018). Learning to lip read words by watching videos. *Computer Vision and Image Understanding*, *173*, 76-85.

Fenghour, S., Chen, D., Guo, K., & Xiao, P. (2020). Lip reading sentences using deep learning with only visual cues. IEEE Access, 8, 215516-215530.

Gollapudi, S., & Gollapudi, S. (2019). OpenCV with Python. *Learn Computer Vision Using OpenCV: With Deep Learning CNNs and RNNs*, 31-50.

Jeon, S., Elsharkawy, A., & Kim, M. S. (2021). Lipreading architecture based on multiple convolutional neural networks for sentence-level visual speech recognition. Sensors, 22(1), 72.

Koishybay, A., Kakenov, N., & Suleimenova, D. (2020). Convolutional Neural Networks for Lip Reading. In *2020 International Conference on Cybernetics & Informatics* (K&I) (pp. 1-4). IEEE.

Lu, X., Shi, P., & Liu, L. (2021). Lip-reading method based on deep learning for the intelligent interaction system. *Journal of Ambient Intelligence and Humanized Computing, 12*(1), 103-113.

Madhukar, N. S., Khade, P. K., Huang, L., Gayvert, K., Galletti, G., Stogniew, M., ... & Elemento, O. (2019). A Bayesian machine learning approach for drug target identification using diverse data types. *Nature communications, 10*(1), 5221.

McAndrew, S. (2020). *Animated Lip-sync using Deep Learning* (unpublished manuscript). Breda University of Applied Sciences, Breda, Netherlands.

Miled, M., Messaoud, M. A. B., & Bouzid, A. (2023). Lip reading of words with lip segmentation and deep learning. *Multimedia Tools and Applications, 82*(1), 551-571.

Nguyen, L., Tran, T., Tran, D., & Nguyen, H. (2020). Deep Learning for Deepfakes Creation and Detection. *Advanced Engineering Informatics, 45*, 101085.

Patil, H., Patel, R., Patel, B., & Shah, D. (2021). Machine Learning and Deep Learning Techniques for Speech Recognition: A Review. *Journal of Intelligent Systems, 30*(1), 81-89.

Salazar-Clemena, R. M. (2006). The state of higher education for deaf students in the Philippines. *NTUT Education of Disabilities*, *5*, 41-49.

Tan, M., & Le, Q. (2021, July). *Efficientnetv2: Smaller models and faster training. In International conference on machine learning* (pp. 10096-10106). PMLR.

Wang, N., Gao, X., Tao, D., Li, X., & Li, J. (2020). Facial feature point detection: A comprehensive survey. *Neurocomputing, 275*, 50-65.

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2019). Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks. *IEEE Signal Processing Letters, 23*(10), 1499-1503.

Author's Biography

Mr. Nikie Jo Deocampo, is a professional with a background in Information Systems, holding a Master's degree in Information Technology, and is currently pursuing a Doctorate in Information Technology. With expertise in Web Technologies, Cloud Computing, and Machine Learning, Mr. Deocampo is a budding researcher committed to making a positive impact on the community through innovative research that aims to improve the quality of life. As a dedicated educator currently teaching at a university, Mr. Deocampo is passionate about imparting knowledge and fostering learning. Eager to collaborate with top researchers in IT and Computer Science, Mr. Deocampo envisions contributing significantly to the advancement of these fields.

Dr. Mia V. Villarica, a professor from Laguna State Polytechnic University, specializes in Data Mining, Information and Communications Technology, and eLearning. She has contributed significantly to academic research, co-authoring papers such as "Serbigo Serbisyo on the Go!: Online job order mobile application for non-professional workers" and "Classification of Coffee Variety using Electronic Nose" published in 2022. Additionally, her work includes "Correlation Analysis between Sensors for Sensing Coffee Variations" presented at the 2022 IEEE 18th International Colloquium on Signal Processing & Applications. Villarica's research interests also extend to developing innovative solutions, as evidenced by her involvement in various projects.

Dr. Albert A. Vinluan is a professor from Isabela State University. With expertise in Machine Learning, Supervised Learning, Data Mining, Applied Artificial Intelligence, and more, Dr. Vinluan has made significant research contributions. His publications include "Emotional analysis and prediction based on online book user comments" and "Research on Emotional Analysis of Online Book Reviews Based on Word2Vec Method," both in 2023, as well as "Barrier-free routes in a geographic information system for mobility impaired people" in 2022.