Improving the Recognition Performance of Lip Reading Using the Concatenated Three Sequence Keyframe Image Technique

Lap Poomhiran
Faculty of Information Technology and Digital Innovation
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand
lap.p@email.kmutnb.ac.th

Phayung Meesad
Faculty of Information Technology and Digital Innovation
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand
pym@kmutnb.ac.th

Sumitra Nuanmeesri
Faculty of Science and Technology
Suan Sunandha Rajabhat University
Bangkok, Thailand
sumitra.nu@ssru.ac.th

Abstract-This paper proposes a lip reading method based on convolutional neural networks applied to Concatenated Three Sequence Keyframe Image (C3-SKI), consisting of (a) the Start-Lip Image (SLI), (b) the Middle-Lip Image (MLI), and (c) the End-Lip Image (ELI) which is the end of the pronunciation of that syllable. The lip area's image dimensions were reduced to 32×32 pixels per image frame and three keyframes concatenate together were used to represent one syllable with a dimension of 96×32 pixels for visual speech recognition. Every three concatenated keyframes representing any syllable are selected based on the relative maximum and relative minimum related to the open lip's width and height. The evaluation results of the model's effectiveness, showed accuracy, validation accuracy, loss, and validation loss values at 95.06%, 86.03%, 4.61%, and 9.04% respectively, for the THDigits dataset. The C3-SKI technique was also applied to the AVDigits dataset, showing 85.62% accuracy. In conclusion, the C3-SKI technique could be applied to perform lip reading recognition.

Keywords-concatenated frame images; convolutional neural network; keyframe reduction; keyframe sequence; lip reading

I. INTRODUCTION

Deep learning applications, especially Convolutional Neural Network (CNN) applications, have recently achieved impressive success in diverse object detection and recognition tasks [1], However, CNNs face some challenges, in particular in video recognition. A video may be incomplete and sound may be lacking during certain parts. If the audio at the crucial moment is missing, it may result in the video's contents being misunderstood [2]. These videos will be more useful if they were edited and the missing words or messages could be found. Most of the proposed solutions rely on the lip reading technique to help transcription by reading and observing the moving lips, including tongue and face to get the right words. Moreover, the process of transcribing or translating the speech obtained by lip reading is a skill that requires learning and practice until becoming proficient at recognizing the lip movement or lip pattern related to the pronunciation of each syllable.

In general, there are two popular multimodals or methods of supervised learning for lip reading: Visual Speech Recognition (VSR) and Audio-Visual Speech Recognition (AVSR). VSR uses a method to teach the machine with visual-only information from a video without using speech or audio for training [3]. On the other hand, AVSR trains the machine by applied images combined with audio data from a video to achieve greater accuracy [4]. Authors in [5] found that the use of Visual-Only (VO) data has better classification accuracy than that of Audio Visual (AV) and Audio Only (AO) data. There are currently two groups of research studies on lip reading recognition, the first group uses images in VO [2, 3, 5-14] while the other group uses both audio and video together [4, 15-25]. Both these groups have a model for extracting the features with a technique used in combination. Nevertheless, the latter group differs in that they have to combine audio feature data with visual features when it comes to machine learning. AV speech recognition is used commercially on various software or systems, but the recognition quality is reduced by environmental noise. This situation is not the same as using data from a quiet image, making lip reading a pivotal role in Automatic Speech Recognition (ASR) in harsh audio environments [6].

Neural networks are commonly used to help extract features and recognize lip reading patterns in machine learning, such as in CNNs [6, 11, 24], Long Short-Term Memory (LSTM) [3, 10, 21], or a combination of CNN with LSTM [2,

Corresponding author: Lap Poomhiran

12, 13]. As mentioned above, in machine learning for lip reading recognition, visual training is required. The images of the face or lips are an essential factor as the primary input. This makes the number of images arranged for use as a large input. In order to sort the images of the video frame, the lip movement is determined by the number of frames, for example, 40 frames per word [7], or the number of frames in seconds such as 0.2 seconds [8, 16] or 1 second, or maybe determined using every frame in the video wherein one video will have only one word and a short duration. The methods mentioned above also use multiple frames in machine learning processing, which consumes more resources and processing time. There are also real world problems, e.g. where a speaker speaks at a slowly speaking speed, resulting in a longer video file length while the words are spread out. Therefore, limitations on the number of frames or the duration may differ from the words or messages conveyed. However, if there is a representation of the images or keyframes, it will reduce the number of images and cost used in machine learning.

Most lip reading studies are applied to English datasets of digits, alphabets, words, phrases, and sentences. The AVDigits [26] is a popular dataset for testing and improving the performance of a model. There is a total of 540 videos with a resolution of 1920×1080 pixels (px). It consists of six speakers who speak numbers from 0 to 9 in English and each repeats the numbers 9 times. Greek [12], Myanmar [14], Spanish [27], and Czech [28] have been also studied, but there are no Thai language datasets created for lip reading. It is difficult for the Thai language to be in some words with similar lip patterns but with different meanings. Since the vowels are pronounced similarly, the patterns of lip movements are similar. There is a pattern of intonation in Thai language resulting from a combination of tones, resulting in intonation in five tonal sounds. Therefore, this research aims to create a dataset in the Thai language and reduce the image dimensions and the number of frames by finding keyframes to replace syllables or words for Thai lip reading recognition using CNNs.

II. RESEARCH METHODOLOGY

The research methodology for the improving performance recognition of lip reading using C3-SKI consists of 1) dataset preparation, 2) face detection and lip localization, 3) C3-SKI creation, 4) model development, and 5) model effectiveness evaluation.

A. Dataset Preparation

The Thai digit dataset is called THDigits. It was created as a Thai video file containing numbers from 0 to 9, which are repeated three times with three different speaking speeds (slow, regular, and fast) by using 100 mixed-gender speakers. A total of 3,000 video files with length between 1 and 4 seconds were constructed. These videos have four different resolutions: 1920×1080, 1280×720, 960×540, and 720×404 px and were recorded with smartphones, regardless of model and brand. The Thai numbering is shown in Table I.

B. Face Detection and Lip Localization

Before processing face detection and lip locating, each video prepared in the Thai digit dataset was separated into individual frames, ordered from the first to the last frame. After that, the face was detected in each frame using the Viola-Jones [29, 30] technique based on the Haar-like feature. The hypothesis T represents any distinguishing characteristic, h is the distinguishing characteristic, and β represents the percentage of error classification. The characteristic $C(x)$ is given as (1) [29]:

$$C(x)=\begin{cases}1, & \sum_{t=1}^{T}\alpha_t h_t(x)\geq\frac{1}{2}\sum_{t=1}^{T}\alpha_t\\ 0, & otherwise\end{cases} \quad (1)$$

where $\alpha_t=\log\frac{1}{\beta_t}$.

The difference of pixel sums or features on an image is compared using the Haar-like feature with the black and white filter that assigned I and P values to represent N by N images as in (2) [30]:

$$\sum_{1\leq i\leq N}\sum_{1\leq j\leq N} I(i,j)1_{P(i,j)=white}-\sum_{1\leq i\leq N}\sum_{1\leq j\leq N} I(i,j)1_{P(i,j)=black} \quad (2)$$

After face detection comes the lip positioning based on 68 facial landmarks [31]. Lip localization determines the lip area or Region Of Interest (ROI) to leave only the desired feature area by cropping in each frame to the lip area only, illustrated in Figure 1.

TABLE I. THAI NUMBER PRONUNCIATION IN THDIGITS DATASET

Numbers	Thai pronunciation	Meaning
0	S̄ūnỳ	Zero
1	H̄nùng	One
2	S̄xng	Two
3	S̄ām	Three
4	S̄ì	Four
5	H̄ā	Five
6	H̄k	Six
7	Cĕd	Seven
8	Pæd	Eight
9	Kêā	Nine

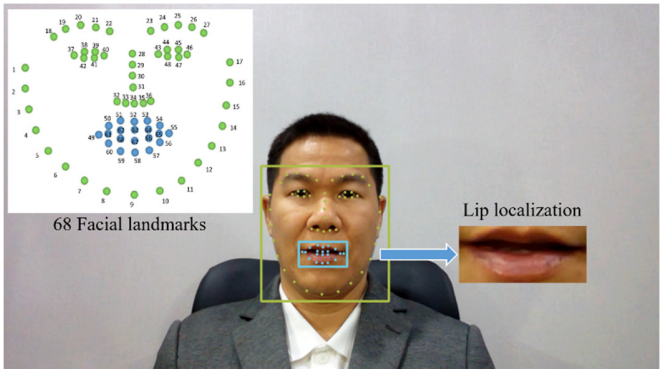

Fig. 1. Face detection and lip localization on each image frame.

C. The Concatenated Three Sequence Keyframe Image

Recent research favorably replaces any word or syllable by concatenating frames as a single image before training. For example, some research uses frames together as a single image

in 5 rows with 5 frames per row, in a total of 25 frames per image [6]. As mentioned above, if the speaker speaks slowly or extensively, the number of frames will be increased, but only one word or syllable will be conveyed. In this study, only the keyframes with pronounced lip movement were selected, relying on measuring the outer lip dimensional from the mouth's height (h) and width (w) for each frame x as shown in Figure 2.

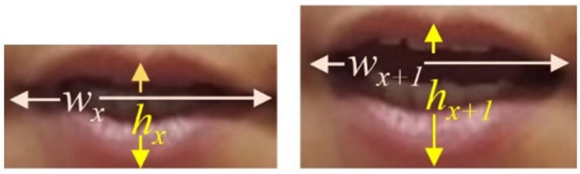

Fig. 2. Lip width (w) and lip height (h) for the current and the next frame.

Lip dimensions were calculated by (3) on each frame. With the help of the increasing function, decreasing function, relative maximum, and relative minimum of a graph related to lip dimensions, all keyframes were extracted and selected. The relative maximum and relative minimum are lying on the slope at the x interval. If the c is some x in a graph, where c is the relative maximum or relative minimum, then the slope at c would be zero, represented in (4).

$$f(x) = w_x + h_x \quad (3)$$

$$f'(c) = 0 \quad (4)$$

$f'(c)$ is the function that has a relative maximum or minimum at c, $f(x) \leq f(c)$ for the relative maximum, and $f(x) \geq f(c)$ for the relative minimum.

Fig. 3. Image frame sequence from a sample video of the dataset.

Each video is possible to produce several relative maxima and relative minima as keyframes. In this experiment, there are three keyframes assigned as a single syllable. The first keyframe will represent the frame where the mouth starts to pronounce a syllable. This is the SLI. The second keyframe (MLI) is the frame where the mouth is at the maximum opening limit or movement of the syllable. The last keyframe (ELI) is at the end of the pronunciation of that syllable. For example, the word 'one' in English is pronounced as 'nueng' (Ñnụ̀ng) in Thai language. Fifty one frames were split from a video in the dataset and the area of the lip was cropped as shown in Figure 3. The frame sequence numbers 9, 18, and 32 are representing the three keyframes based on relative maximum and relative minimum, and are shown in Figure 4.

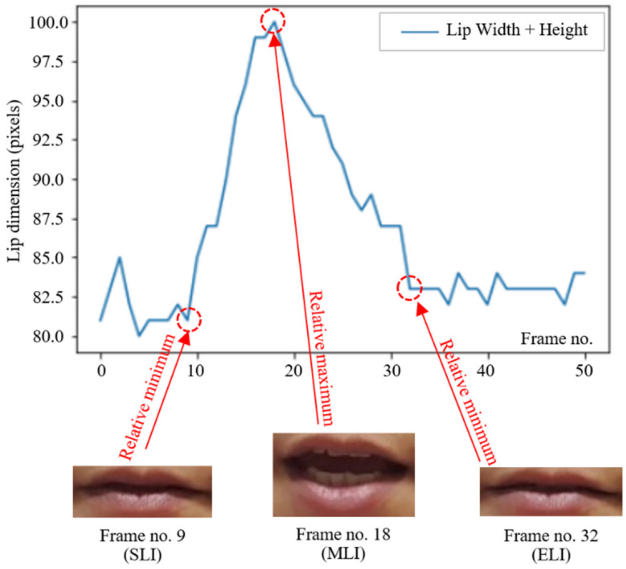

Fig. 4. The three keyframes representing a single syllable.

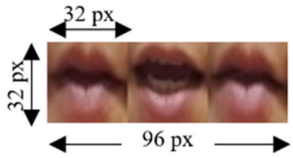

Fig. 5. The image dimensions for training in the current study.

After the three keyframes were selected, each keyframe image was scaled down to 32×32 pixels and merged into a single RGB color image with a dimension of 96×32 pixels, as shown in Figure 5, while other studies use image dimensions of 80×60 or 224×224 pixels ([7] and [6] respectively). Thus, each input color image frame as the dataset for building the model has a pixel density ratio of 3,072 pixels, which is less than the one used in other studies [6, 7].

D. Model Development

The model was designed based on the CNNs with a total of 13 layers. It consists of 1 normalization layer, 6 convolutional (Conv) layers with or Rectified Linear Unit (ReLU) activation function for each convolutional layer, 3 max-pooling layers, and 3 Fully-Connected (FC) layers, which include a flatten layer and two dense layers. There are a total of 846,890 parameters for training. The model architecture is shown in Figure 6.

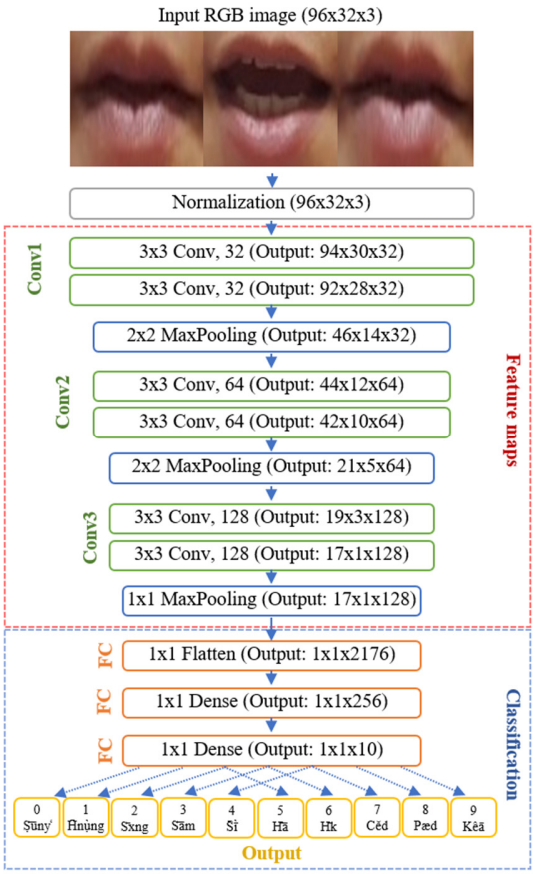

Fig. 6. Model architecture for lip reading applied to C3-SKI.

All images of the Thai digit dataset were used in the training and validation of the designed model. The training and validation images were the 80% and 20% of the dataset respectively. A total of 200 epochs of training was set with a minibatch gradient descent size of 32. The model was built using an Intel Core i7-7700HQ PC at 2.80GHz, 16GB of memory, 512GB of Samsung Solid State Drive, and a 3GB NVIDIA GeForce GTX 1060. This system was running through Python version 3.8.3 on Windows 10 x64 architecture.

E. Model's Effectiveness Evaluation

The developed model has validated by the accuracy rate and loss value. The accuracy was calculated by (5), and the loss value or loss function, also known as cross-entropy which is the favorite function for classification is defined in (6) [32-38]:

$$Accuracy = \frac{C}{N} \quad (5)$$

where C refers to the total number of samples recognized correctly, N refers to the total number of all samples.

$$Cross-entropy = -\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j} \log\left(p_{i,j}\right) \quad (6)$$

where N refers to the total number of all samples, M refers to the total number of classes, $y_{i,j}$ refers to the true value when the sample i belongs to the class j, $p_{i,j}$ refers to the probability predicted by the sample model of i to belong to the class j.

The model will stop training automatically using the 'EarlyStopping' feature supported by Keras library with the callback function. The patience value was 5, and monitoring to the validation loss value. The model stopped training when there was no improvement in the validation loss for 5 consecutive epochs. The minimum validation loss was reached at 37 epochs of training.

III. RESULTS

According to the experiments on the developed model, the accuracy and loss value were determined while building the model. After the 37 epochs of training, the model gave 95.06% accuracy and 4.61% loss value. The model stopped training after 47 epochs. Considering the model's validation, the model gave validation accuracy value of 86.03% and validation loss of 9.04%. The accuracy and loss for both training and validation are shown in Figures 7-9.

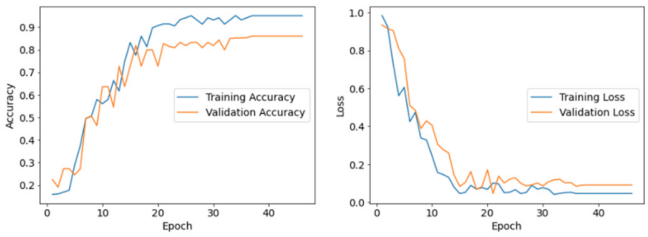

Fig. 7. The accuracy value of training and validation (left) and the loss value of training and validation (right).

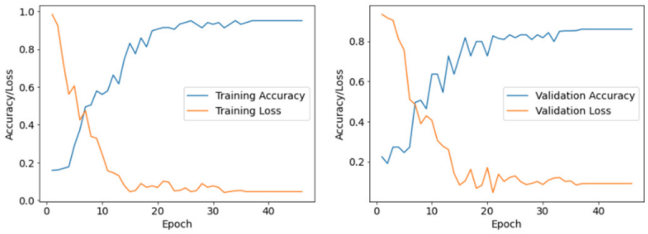

Fig. 8. The training accuracy and loss (left), and the validation accuracy and loss (right).

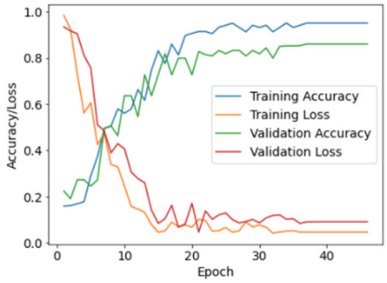

Fig. 9. The accuracy and loss of training and validation of the model.

Moreover, this experiment compares the accuracy of the technique for input image reduction with other studies using the AVDigits dataset. It was found that the lip reading models built by MDBN [39], MDAE [40], RTMRBM [26], am-LSTM

[41], and alm-GRU [20] gave accuracy of 55.00%, 66.74%, 71.77%, 85.23%, and 85.53%, respectively while this work gave 85.62%. The model performances are shown in Table II and Figure 10.

TABLE II. THE MODEL'S EFFECTIVENESS EVALUATION

Dataset	Method	Modality	Accuracy (%)
AVDigits	MDBN [39]	Visual-only	55.00
AVDigits	MDAE [40]	Audiovisual	66.74
AVDigits	RTMRBM [26]	Audiovisual	71.77
AVDigits	am-LSTM [41]	Audiovisual	85.23
AVDigits	alm-GRU [20]	Visual-only	85.53
AVDigits	C3-SKI	Visual-only	85.62
THDigits	C3-SKI	Visual-only	86.03

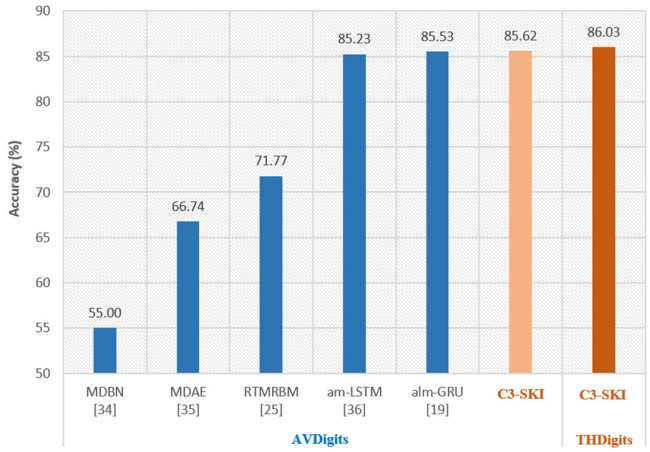

Fig. 10. Comparison with the state-of-the-art methods.

IV. CONCLUSION AND DISCUSSION

This paper proposes the application of the C3-SKI to a CNN for lip reading. The C3-SKI consisting of SLI, MLI, and ELI was tested in lip reading recognition on THDigits and AVDigits datasets. Its primary input is images applied to machine learning according to deep learning techniques. Reducing the number of images can be used by finding the image keyframes based on the relative maximum and relative minimum values. In this paper, 32×32 pixels of 3 sequence keyframes were assigned to represent any syllable of a digit between zero and nine, and these keyframes were concatenated to produce a new 96×32 px image as an input to the neural network. These new input images make the resulting image dimensions and density of pixels less than most state-of-the-art methods. Thus, there were 3,000 keyframe input images with 10 classes that were divided to 80% for training and 20% for validation. The developed CNN model included 1 normalization layer, 6 convolutional layers, 3 max-pooling layers, and 3 fully-connected layers. Training took a total of 47 epochs and was finalized when the validation loss value reached its minimum and the model did not improve any further. As a result, the model had accuracy, validation accuracy, loss, and validation loss values of 95.06%, 86.03%, 4.61%, and 9.04% respectively.

The model's accuracy was 85.62% when it was applied to the AVDigits dataset. Therefore, the reduced number of keyframes with relative maximum and relative minimum can be applied in conjunction with CNN for lip reading. The results of this study conform to the conclusions of [6, 7] in which the CNN method with reduced concatenated frame images was applied to lip reading with a high level of effectiveness. These concatenated frame images have smaller dimension than the traditional images which are usually used to image classification in CNNs (224×224 px).

In future work, the researchers plan to create a dataset for Thai sentences and test the C3-SKI technique. New keyframes could be scaled down and compared to the number of keyframes used to represent each syllable for continuous speech or sentences, such as 5, 7, or 9 keyframe images sequentially. This technique will develop a function or equation to find the lip shift's common point from one syllable to the following syllable. Besides, word prediction methods from the corpus would be used for prediction combined with image processing to increase lip reading accuracy on deep learning in real-time processing.

ACKNOWLEDGMENT

The authors are grateful to the Graduate College, Faculty of Information Technology and Digital Innovation at King Mongkut's University of Technology North Bangkok, and the Faculty of Science and Technology, Suan Sunandha Rajabhat University, for supporting this research.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings IEEE Computer Visualization and Pattern Recognit*, Las Vegas, NV, USA, 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.

[2] S. Fenghour, D. Chen, and P. Xiao, "Decoder-encoder LSTM for lip reading," in *Proceedings of the 2019 8th International Conference on Software and Information Engineering*, Cairo, Egypt, Apr. 9-12, 2019, pp. 162–166, http://doi.org/10.1145/3328833.3328845.

[3] S. Petridis, Z. Li, and M. Pantic, "End-to-end visual speech recognition with LSTMS," in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech, and Signal Processing*, New Orleans, LA, USA, Mar. 5-9, 2017, pp. 2592–2596, https://doi.org/10.1109/ICASSP.2017.7952625.

[4] S. Chung, J. S. Chung, and H. Kang, "Perfect match: Improved cross-modal embeddings for audio-visual synchronisation," in *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Brighton, UK, May 12-17, 2019, pp. 3965–3969, https://doi.org/10.1109/ICASSP.2019.8682524.

[5] R. Bi and M. Swerts, "A perceptual study of how rapidly and accurately audiovisual cues to utterance-final boundaries can be interpreted in Chinese and English," *Speech Communication*, vol. 95, pp. 68–77, 2017, https://doi.org/10.1016/j.specom.2017.07.002.

[6] D. Jang, H. Kim, C. Je, R. Park, and H. Park, "Lip reading using committee networks with two different types of concatenated frame images," *IEEE Access*, vol. 7, pp. 90125–90131, 2019, https://doi.org/10.1109/ACCESS.2019.2927166.

[7] A. Mesbah, A. Berrahou, H. Hammouchi, H. Berbia, H. Qjidaa, and M. Daoudi, "Lip reading with Hahn convolutional neural networks," *Image and Vision Computing*, vol. 88, pp. 76–83, 2019, http://doi.org/10.1016/j.imavis.2019.04.010.

[8] J. S. Chung and A. Zisserman, "Learning to lip read words by watching videos," *Computer Vision and Image Understanding*, vol. 173, pp. 76–85, 2018, http://doi.org/10.1016/j.cviu.2018.02.001.

[9] Z. Thabet, A. Nabih, K. Azmi, Y. Samy, G. Khoriba, and M. Elshehaly, "Lipreading using a comparative machine learning approach," in *Proceedings of the 2018 First International Workshop on Deep and*

Representation Learning, Cairo, Egypt, 2018, pp. 19–25, https://doi.org/10.1109/IWDRL.2018.8358210.

[10] S. Petridis, J. Shen, D. Cetin, and M. Pantic, "Visual-only recognition of normal, whispered and silent speech," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Calgary, AB, Canada, 2018, pp. 6219–6223, https://doi.org/10.1109/ICASSP.2018.8461596.

[11] A. Koumparoulis and G. Potamianos, "Deep View2View mapping for view-invariant lipreading," in *Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT)*, Athens, Greece, December 18-21, 2018, pp. 588–594, https://doi.org/10.1109/SLT.2018.8639698.

[12] J. Wei, F. Yang, J. Zhang, R. Yu, M. Yu, and J. Wang, "Three-dimensional joint geometric-physiologic feature for lip-reading," in *Proceedings of the 2018 IEEE 30th International Conference on Tools with Artificial Intelligence*, Greece, 2018, pp. 1007–1012, https://doi.org/10.1109/ICTAI.2018.00155.

[13] I. Fung and B. K. Mak, "End-to-end low-resource lip-reading with Maxout CNN and LSTM," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Calgary, AB, Canada, 2018, pp. 2511–2515, https://doi.org/10.1109/ICASSP.2018.8462280.

[14] T. Thein and K. M. San, "Lip localization technique towards an automatic lip reading approach for Myanmar consonants recognition," in *Proceedings of the 2018 International Conference on Information and Computer Technologies*, IL, USA, 2018, pp. 123–127, https://doi.org/10.1109/INFOCT.2018.8356854.

[15] S. Yang *et al.*, "LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild," in *Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition*, Lille, France, 2019, pp. 1–8, https://doi.org/10.1109/FG.2019.8756582.

[16] J. S. Chung and A. Zisserman, "Lip reading in profile," in *Proceedings of the 28th British Machine Vision Conference*, London, UK, 2017, https://doi.org/10.5244/C.31.155.

[17] P. P. Filntisis, A. Katsamanis, P. Tsiakoulis, and P. Maragos, "Video-realistic expressive audio-visual speech synthesis for the Greek language," *Speech Communication*, vol. 95, pp. 137–152, 2017, https://doi.org/10.1016/j.specom.2017.08.011.

[18] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–11, 2018, https://doi.org/10.1109/TPAMI.2018.2889052.

[19] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2018, pp. 6548–6552, https://doi.org/10.1109/ICASSP.2018.8461326.

[20] Y. Yuan, C. Tian, and X. Lu, "Auxiliary loss multimodal GRU model in audio-visual speech recognition," *IEEE Access*, vol. 6, pp. 5573–5583, 2018, https://doi.org/10.1109/ACCESS.2018.2796118.

[21] S. Petridis, T. Stafylakis, P. Ma, G. Tzimiropoulos, and M. Pantic, "Audio-visual speech recognition with a Hybrid CTC/Attention architecture," in *Proceedings of the 2018 IEEE Spoken Language Technology Workshop*, Athens, Greece, 2018, pp. 513–520, https://doi.org/10.1109/SLT.2018.8639643.

[22] W. J. Ma, X. Zhou, L. A. Ross, J. J. Foxe, and L. C. Parra, "Lip-reading aids word recognition most in moderate noise: A bayesian explanation using high-dimensional feature space," *PLoS ONE*, vol. 4, no. 3, 2009, Art. no. e4638, https://doi.org/10.1371/journal.pone.0004638.

[23] M. Wand, J. Koutník, and J. Schmidhuber, "Lipreading with long short-term memory," in *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Shanghai, China, 2016, pp. 6115–6119, https://doi.org/10.1109/ICASSP.2016.7472852.

[24] A. Gabbay, A. Shamir, and S. Peleg, "Visual speech enhancement," in *Proceedings of Interspeech 2018, 19th Annual Conference of the International Speech Communication Association*, India, Sep. 2-6, 2018, pp. 1170–1174.

[25] M. Wand, J. Schmidhuber, and N. T. Vu, "Investigations on end-to-end audiovisual fusion," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Calgary, AB, Canada, 2018, pp. 3041–3045, https://doi.org/10.1109/ICASSP.2018.8461900.

[26] D. Hu, X. Li, and X. Lu, "Temporal multimodal learning in audiovisual speech recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 3574–3582, https://doi.org/10.1109/CVPR.2016.389.

[27] A. Fernandez-Lopez and F. M. Sukno, "Automatic viseme vocabulary construction to enhance continuous lip-reading," in *Proceedings of the 12th International Conference on Computer Vision Theory and Applications*, Porto, Portugal, Feb. 27- Mar. 1, 2017, pp. 52–63.

[28] K. Paleček, "Experimenting with lipreading for large vocabulary continuous speech recognition," *Journal on Multimodal User Interfaces*, vol. 12, no. 4, pp. 309–318, 2018, https://doi.org/10.1007/s12193-018-0266-2.

[29] P. Viola and M. J. Jones, "Robust real-time face detection," International *Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004, https://doi.org/10.1023/B:VISI.0000013087.49260.fb.

[30] Y.-Q. Wang, "An analysis of the Viola-Jones face detection algorithm," *Image Processing On Line*, vol. 4, pp. 128–148, 2014, https://doi.org/10.5201/ipol.2014.104.

[31] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," *International Journal of Computer Vision*, vol. 91, pp. 200–215, 2011, https://doi.org/10.1007/s11263-010-0380-4.

[32] K. Janocha and W. M. Czarnecki, "On loss functions for deep neural networks in classification," *Schedae Informaticae*, vol. 25, pp. 49–59, 2016, https://doi.org/10.4467/20838476SI.16.004.6185.

[33] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proceedings of the 32nd Conference on Neural Information Processing Systems*, Montréal, Canada, Dec. 2-8, 2018.

[34] Q. Zhu, Z. He, T. Zhang, and W. Cui, "Improving classification performance of softmax loss function based on scalable batch-normalization," *Applied Sciences*, vol. 10, no. 8, pp. 29–50, 2020, https://doi.org/10.3390/app10082950.

[35] N. Srivastava and R. Salakhutdinov, "Learning representations for multimodal data with deep belief nets," presented at *the 29th International Conference on Machine Learning Workshop*, Edinburgh, UK, Jun. 26-Jul. 1, 2012.

[36] M. B. Ayed, "Balanced communication-avoiding support vector machine when detecting epilepsy based on EGG signals," *Engineering, Technology & Applied Science Research*, vol. 10, no. 6, pp. 6462–6468, 2020, https://doi.org/10.48084/etasr.3878.

[37] S. Nuanmeesri, "Mobile application for the purpose of marketing, product distribution and location-based logistics for elderly farmers," *Applied Computing and Informatics*, 2019, https://doi.org/10.1016/j.aci.2019.11.001.

[38] A. N. Saeed, "A machine learning based approach for segmenting retinal nerve images using artificial neural networks," *Engineering, Technology & Applied Science Research*, vol. 10, no. 4, pp. 5986–5991, 2020, https://doi.org/10.48084/etasr.3666.

[39] A. U. Ruby, P. Theerthagiri, I. J. Jacob, and Y. Vamsidhar, "Binary cross entropy with deep learning technique for image classification," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 4, pp. 5393–5397, 2020, https://doi.org/10.30534/ijatcse/2020/175942020.

[40] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. "Multimodal deep learning," in *Proceedings of the 28th International Conference on Machine Learning*, Washington, USA, 2011, pp. 689–696.

[41] C. Tian, and W. Ji, "Auxiliary multimodal LSTM for audio-visual speech recognition and lipreading," 2017, arXiv preprint arXiv:1701.04224v2.

AUTHORS PROFILE

Lap Poomhiran is currently a Ph.D. student in the Department of Information Technology, Faculty of Information Technology and Digital Innovation, King Mongkut's University of Technology North Bangkok (KMUTNB), Thailand. His research interests include web and mobile programming, augmented reality (AR) and virtual reality (VR) development, image processing, data mining, machine learning, deep learning, and the internet of things (IoT).

Phayung Meesad is an associate professor at the King Mongkut's University of Technology North Bangkok (KMUTNB), Thailand. His research interests include computational intelligence, artificial intelligence, machine learning, deep learning, data analytics, big data analysis, data science, data mining, digital signal processing, image processing, business intelligence, time series analysis, and natural language processing.

Sumitra Nuanmeesri is an assistant professor in the Department of Information Technology, Faculty of Science and Technology at Suan Sunandha Rajabhat University, Thailand. Her research interests include speech recognition, data mining, deep learning, image processing, mobile application, supply chain management systems, internet of things, robotics, augmented reality, and virtual reality.