Improving the Representative Concatenated Frame Images Base on Convolutional Neural Network for Thai Lip Reading Recognition

1st Lap Poomhiran
Department of Information Technology
Faculty of Information Technology and Digital Innovation
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand
lap.p@email.kmutnb.ac.th

2nd Phayung Meesad
Department of Information Technology
Faculty of Information Technology and Digital Innovation
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand
pym@kmutnb.ac.th

3rd Sumitra Nuanmeesri
Department of Information Technology
Faculty of Science and Technology
Suan Sunandha Rajabhat University
Bangkok, Thailand
sumitra.nu@ssru.ac.th

***Abstract*—Lip reading could be improved by providing more training data with deep learning techniques. This work aims to improve the concatenated frame images for use as input in deep learning processes by reducing the number of frames and image size for visual speech recognition in Thai. The developed model could be used the Convolutional Neural Network to detect and classify the lip motion from speakers on videos in Thai. The experimental result showed that the developed model with concatenated five frame images gave a training accuracy at 95.67% and training loss was 4.23%, where the validation accuracy was 87.12%, and validation loss was 8.79%. It could be said that the concatenated five frame images were represented as input data and improved Thai lip-reading recognition using the convolutional neural network.**

Keywords—concatenated frame images, convolutional neural network, deep learning, lip reading, visual speech recognition

I. INTRODUCTION

Lip reading is a fascinating skill to understand words using visual-only data without audio [1]. For example, the movies where there is no vocal sound or some voice scenes are lost. Lip reading is sometimes called the third ear for deciphering each word by observing the lip movement and expression on the face of the speaker [2]. The introduction of deep learning in conjunction with lip reading has become more widespread, such as Convolutional Neural Network (CNN). CNN is one of the deep learning approaches that could be used in lip-reading recognition. However, most lip-reading recognition depends on the patterns of input data and the model architecture that are introduced into deep learning. The learning dataset can be categorized into three groups: 1) visual-only (VO), 2) audio-only (AO), and 3) audio-visual (AV). In the case of a visual-only dataset, some researchers required a technique for selecting frame images for use in learning to match lip movement patterns or even facial expressions to words or sentences.

Nowadays, several datasets were created for lip reading recognition which included digits, letters, words, phrases, and sentences, such as AVDigits [3]. Most of the datasets are in English, and there are also datasets being developed in other languages such as Czech [4], Japanese [5], Greek [6], Spanish [7]. However, the Thai language dataset has not yet been created for lip reading. Hence, this research aims to create a dataset for numerical lip reading recognition that can be read aloud in Thai. Moreover, this work improved the input image dimensions by selecting the representative images, rescaling and concatenating them based on CNN learning.

II. RELATED WORK

A. Convolutional Neural Network

The Convolutional Neural Network is an artificial neural network [8] in deep learning or machine learning technique by relying on mathematical principles to assist in learning from examples [9]. The CNN includes the filter or kernel used to calculate the feature map that uses simulating small areas to extract feature and classify in sub-area. It is then put back together to decide what the results will ultimately be classified according to probability. CNN has two main parts: the hidden layers part and the classification part. The hidden layers part is responsible for detecting and extracting the desired feature extraction from the input data. It consists of a convolution layer and a pooling layer, which may be assigned several layers for the machine to learn until it gets the obvious features repeatedly. The classification part consists of fully connected layers that will classify the input data according to the feature extracted from the hidden layers. The fully connected layer and a Softmax function were included in the classification part. It helps classify the features of the data again until the final result came out.

B. Concatenated Frame Images

The concatenated frame images (CFIs) is a technique for put the sequence images of video frames together at the end of the previous frame in a single row or single column or mixed between row and column. The method of collating images might be different depending on the individual research design. The first presented technique was applied for visual speech recognition using the CNN method [10]. Some studies applied the concatenated frame images for visual speech recognition with the final concatenated image dimension of 80x60 pixels [11] and 224x224 pixels [12][13]. Thus, the CFIs were composed of multiple frame images combined into a single image that could be used as CNN training input data.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

C. Lip Reading

It is to be noted that not much work is reported on techniques for recognizing Thai lip reading. However, some studies were conducted to compare the techniques for lip reading recognition. For example, the CNN models were conducted for visual speech recognition [14][15]. Similar to Burton et al. [16], the CNN technique is the most robust classifier for lip reading. In some studies, the CNN methods have been performed to predict phonemes in spoken Japanese for visual speech recognition [5]. Hashmi et al. [13] created 224x224 pixels of input CFIs with 7x7 frames and built the CNN model with batch normalization, resulting in the validation accuracy was increased up to 56%. Jang et al. [12] applied two types of CFIs with 5x4 frames from the OuluVS2 dataset using CNN based on the Visual Geometry Group (VGG) model. They found that the model gave accuracy at 90.9%. Fung and Mak [17] combined Bidirectional Long Short Term Memory (BLSTM) and CNN with rectified linear unit activation function for the OuluVS2 dataset. Their technique showed that the word accuracy was 87.6%.

Based on the importance and previous studies about recognizing lip reading techniques, the researchers decided to improve the Thai lip reading recognition dataset based on CNN.

III. Methodology

There are four processing methods for improving Thai lip reading recognition, including frame image preprocessing, building the concatenated frame images, CNN modeling for lip reading, and the model evaluation.

A. Frame Image Preprocessing

This section describes the pipeline for generating a visual speech recognition dataset in Thai. The authors have created videos that speakers have spoken a digit between zero and nine in Thai number pronunciation. These speakers were aged between eighteen and twenty-one years, comprising seventy-eight males and twenty-two females. All speakers have spoken each digit three times at different speeds, allowing the same numeric speech to result in different video lengths. The total videos of three thousand files were recorded by the camera on any smartphone in front view. There are various video frame resolutions between 720x404 pixels and 1920x1080 pixels.

The prepared dataset has split the frame images from each video file in this process. The number of resulting frame images might be different depending on the video length and video frame rate. For example, a video file with a length of 1.5 seconds and a frame rate of thirty frames per second (fps) will produce forty-five frame images. During the split frame image process, face detection is also applied at this stage. The Haar-like feature technique was conducted to detect the face on a frame image. This technique was presented by Viola-Jones [18][19].

This work focuses on the lip position that shifted or changed between the first and last frames that found the face. Thus, the mouth region of the speaker was located on the face based on the sixty-eight facial landmarks approach [20]. After that, a specific mouth region image was cropped according to the lip's width and height that appeared on each frame image. The resulting cropped image of only the lips that are sequenced from the beginning of opening the mouth to speak and then ending the closing of the mouth to end one syllable.

B. Building the Concatenated Frame Images

The number of cropped images in each video was different depending on the speaking speed mentioned previously. Therefore, the selection of five suitable representative images was applied based on slope values to find increasing and decreasing function locations on cropped frame images sequentially. The slope of each cropped frame image is calculated from the height and the width of the outer lip concerning adjacent frames as Fig. 1.

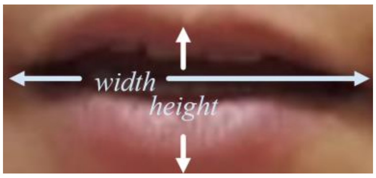

Fig. 1. The outer lip width and height.

A cropped frame image with a zero slope is the relative maximum or relative minimum, based on the slope of the previous and next cropped frames. Thus, five cropped frame image sequences were selected as illustrated in Fig. 2 and described as follows.

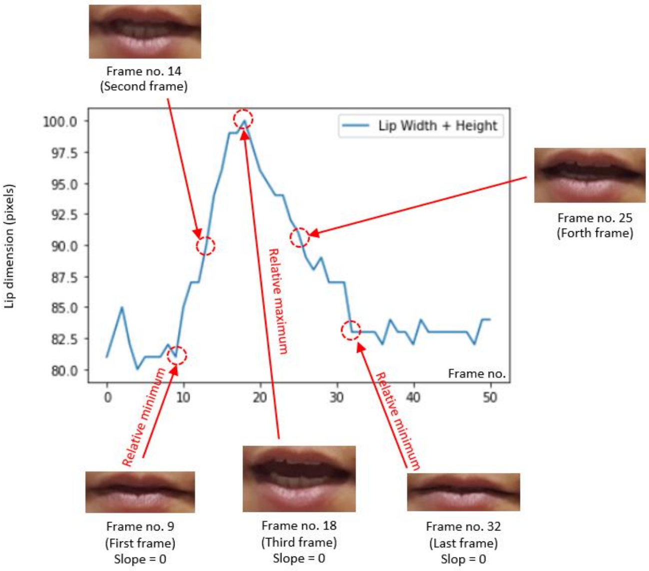

Fig. 2. The concatenated five frame images were selected.

- *The 1st frame image* is represented to the cropped frame image with the relative minimum and mouth opening at the speaking begin.
- *The 2nd frame image* is represented to the cropped frame centered on the positive slope line between the 1st frame image and the 3rd frame image.
- *The 3rd frame image* is represented to the cropped frame image with the maximum relative and maximum mouth opening of each syllable.
- *The 4th frame image* is represented to the cropped frame centered on the negative slope line between the 3rd frame image and the 5th frame image.
- *The 5th frame image* is represented to the cropped frame image with the relative minimum and mouth closing at the end of speaking.

Finally, all representative cropped frame images were rescaled the image dimension to 32x32 pixels, then concatenated all five cropped frame images as a single image for each video file resulting in new image dimensions at 32x160 pixels.

C. *CNN Modeling for Lip Reading*

All representative concatenated frame images were processed to create a Thai lip reading recognition model using CNN that was written in Python version 3.8.1 with Tensorflow and Keras library. The developed model including seven convolutional layers with rectified linear unit activation function, four max-pooling layers, one flatten layer, and two dense layers. The final output was ten classes with 4,651,050 trainable parameters in total. For the CNN modeling, some configurations were set as follows: gradient descent size=32, each trail to learn=500 epochs, optimizer algorithm=adam, and early stopping=true (with monitor=val_loss and patience=5). The model was built based on the Windows platform with 64bits architecture, including 2.8 gigahertz of a 4-cores processor and 16 gigabytes of memory.

D. *The Model Evaluation*

In this study, the authors evaluated the developed model in terms of multi-class cross-entropy loss and accuracy. The categorical cross-entropy was set as a loss function where the metric was accurate by using the Keras compile function. Sometimes, the categorical cross-entropy loss was called softmax loss, which combines softmax activation function and cross-entropy loss. The categorical cross-entropy loss (*L*) was calculated as in (1) [21], and the conventional accuracy (*Acc*) was formulae in (2) [22].

$$L(y,\hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\log(\hat{y}_{ig}) \qquad (1)$$

where

N refers to the total number of sample;

i refers to one-hot encoded sample;

$\hat{y}$ refers to the parameter of network for training data;

$\hat{y}_{ig}$ refers to the probability of predicting the class with ground truth by the sample *i*.

$$Acc = \frac{P_c}{N} \qquad (2)$$

where

P_C refers to the total correct number of prediction;

N refers to the total number of samples.

The concatenated five frame images dataset was split into 80% of the training set and 20% of the validation set for the model evaluation.

IV. Results

The experimental result has shown that the CNN model for Thai lip reading was built with the dataset that concatenated five frame images gave the training accuracy at 95.67%, the training loss at 4.23%, the validation accuracy at 87.12%, and the validation loss was 8.79% at the end of sixty-two epochs. The training accuracy and validate accuracy value of the CNN model for Thai lip reading were illustrated in Fig. 3, and the training loss and validate loss were shown in Fig. 4.

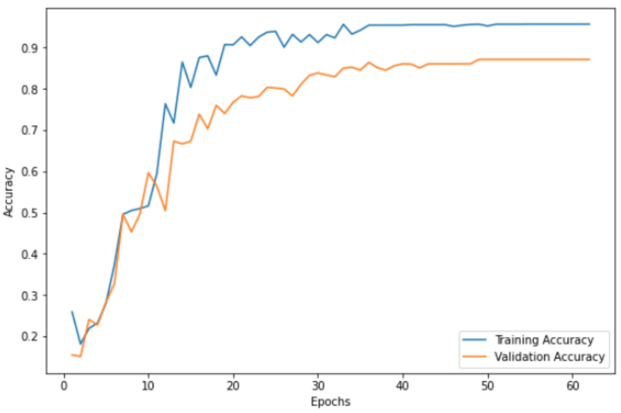

Fig. 3. The training accuracy and loss of the Thai lip reading model.

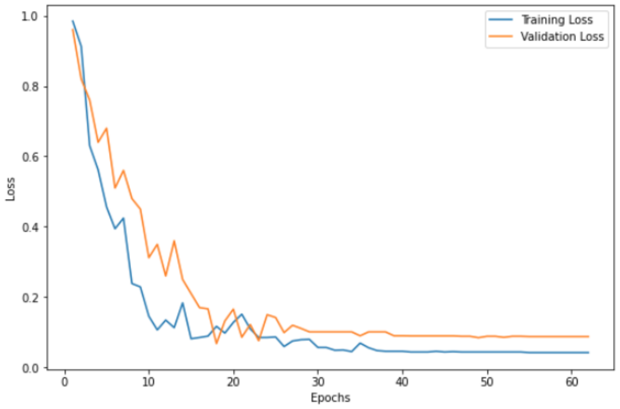

Fig. 4. The training loss and validation loss of the Thai lip reading model.

Moreover, this work gave the highest accuracy compared to other studies for visual-only lip reading by applied concatenated frame images at utterance level in Table I.

TABLE I. THE ACCURACY COMPARED TO OTHER STUDIES BY APPLIED CONCATENATED FRAME IMAGES

Dataset	Method	Accuracy (%)
AVLetters	HCNN model [11]	59.23
AVDigits	Alm-GRU [24]	85.53
AVDigits	C3-SKI [25]	85.62
THDigits	C3-SKI [25]	86.03
THDigits	This work	87.12

V. Conclusions

This work proposed improved Thai lip-reading recognition at utterance level by applying the concatenated five frame images which were trained and validated in the developed CNN model based on accuracy and loss function. The result showed that the model gave 95.67% of training accuracy and 4.23% of training loss, while the validation accuracy was 87.12%, and the validation loss was 8.79% when the model trained end for sixty-two epochs. It could be said that the concatenated five frame images were representative of a syllable of digit and could be used in Thai lip reading recognition based on CNN. In this works, the new image dimension is less than the typical image of 224x224 pixels [12][13] for classical image classification in the pre-trained model such as VGG16 [23]. The learning resources and cost were reduced as well when decreased the image dimension. Nevertheless, decreasing the image dimension until it is less than a certain extent might also decrease learning efficiency. However, these new images also gave accuracy at the highest level of Lip reading recognition.

For further work, the authors would be building a Thai dataset in phrases and sentences for continuous speech recognition and designing the CNN model architecture suitable for lip reading in Thai sentences. Moreover, the model would compare the effectiveness when applied to the dataset in different views, such as front view, side view, and thirty to sixty degrees of view.

Acknowledgment

The authors are grateful to the Faculty of Information Technology and Digital Innovation at King Mongkut's University of Technology North Bangkok and the Faculty of Science and Technology at Suan Sunandha Rajabhat University, who supported and gave an opportunity in this research.

References

[1] 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. doi: 10.1016/j.cviu.2018.02.001

[2] "A beginner's guide to lipreading," 2019. Accessed on: Jul. 1, 2020. [Online]. Available: https://www.lipreading.org/beginners-guide-to-lipreading

[3] 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

[4] 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

[5] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Lipreading using convolutional neural network," in *Proc. 15th Annual Conference of the International Speech Communication Association, INTERSPEECH 2014*, Singapore, Sep 2014, pp. 1149–1153.

[6] 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

[7] 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. arXiv preprint arXiv:1704.08035

[8] T. Ozcan and A. Basturk, "Lip reading using convolutional neural networks with and without pre-trained models," *Balkan Journal of Electrical and Computer Engineering*, vol. 7, no. 2, pp. 195–201, 2019. doi: 10.17694/bajece.479891

[9] V. Franc and J. Čech, "Learning CNNs from weakly annotated facial images," *Image and Vision Computing*, vol. 77, pp. 10–20, 2018. doi: 10.1016/j.imavis.2018.06.011

[10] T. Saitoh, Z. Zhou, G. Zhao, and M. Pietikäinen, "Concatenated frame image based CNN for visual speech recognition," in *Computer Vision – ACCV 2016 International Workshops*, Taipei, Taiwan, Nov 2016, pp. 277–289. doi: 10.1007/978-3-319-54427-4_21

[11] 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

[12] 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

[13] S. N. Hashmi, H. Gupta, D. Mittal, K. Kumar, A. Nanda, and S. Gupta, "A lip reading model using CNN with batch normalization," in *Proc. of the 2018 Eleventh International Conference on Contemporary Computing (IC3)*, Noida, India, Aug 2018, pp. 1–6. doi: 10.1109/IC3.2018.8530509

[14] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proc. of the Computer Vision - ACCV 2016*, Cham, 2017, pp. 87–103.

[15] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 3444–3453. doi: 10.1109/CVPR.2017.367

[16] J. Burton, D. Frank, M. Saleh, N. Navab, and H. L. Bear, "The speaker-independent lipreading play-off; a survey of lipreading machines," in *Proc. of the 2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS)*, Dec 2018, pp. 125–130. doi: 10.1109/IPAS.2018.8708874

[17] I. Fung and B. Mak, "End-to-end low-resource lip-reading with Maxout CNN and LSTM," in *Proc. of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr 2018, pp. 2511–2515. doi: 10.1109/ICASSP.2018.8462280

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

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

[20] 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. doi: 10.1007/s11263-010-0380-4

[21] 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. doi: 10.3390/app10082950

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

[23] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in 3rd *International Conference on Learning Representations*, San Diego, CA, USA, May 2015. arXiv preprint arXiv: 1409.1556

[24] 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

[25] L. Poomhiran, P. Meesad, and S. Nuanmeesri, "Improving the recognition performance of lip reading using the concatenated three sequence keyframe image technique," Engineering, Technology & Applied Science Research, vol. 11, no. 2, pp. 6986–6992, 2021. doi: 10.48084/etasr.4102