ISBN: 978-981-18-7950-0
2023 the 13th International Workshop on Computer Science and Engineering (WCSE 2023)
doi: 10.18178/wcse.2023.06.013

Lip Shape Classification of Sounds for Speech Therapy using SlowFast Networks

Penpicha Boonsri [1], Salita Eiamboonsert [2] and Punnarai Siricharoen [1, +]

[1] Perceptual Intelligent Computing Lab, Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Thailand

[2] Media Technology, King Mongkut's University of Technology Thonburi, Thailand

Abstract. It is important for patients with Aphasia and Dysarthria to have speech and language therapy to practice breathing exercises, tongue strengthening exercises, and especially speech sounds such as short vowel sounds. To ensure the clarity of the pronunciation sound and the correct position of mouth shape, it is required to be monitored by a therapist. We proposed an automated method using convolutional networks to identify the motion of pronunciation of 9 short vowel sounds which is required for speech exercises in the Thai language. Firstly, videos of vowel sound pronunciation are captured, then preprocessed to crop only the mouth area using Dlib library. The cropped image sequence is then fed into audiovisual SlowFast Networks based on convolutional networks which have Slow and Fast visual pathways to capture spatial and temporal information of a video. We compared our selected model with the transformer-based state-of-the-art model, such as TimeSformer. Our proposed framework using SlowFast networks achieved average accuracy at 97.3% for 9-class video classification of Thai vowel sounds. It shows our proposed framework has a potential for use as a tool for speech sound self-exercises and therapy.

Keywords: Deep Learning, Speech Therapy, Stroke, SlowFast, Timesformer, Video Classification

1. Introduction

Stroke is one of the major causes of death and even though patients are able to survive, most of them end up being long-term or permanent disabilities due to blockage or rupture of blood vessels. This can affect several parts of the body, for example, hemiplegia, and especially, loss of ability to convey meaning which causes communication obstacles as a result, participation in social activities decreased [1]. To improve the quality of life in these communication-struggling ischemic stroke patients, speech and language therapy is therefore crucial. In this speech therapy, the patient needs to be assessed by a speech therapist in order to diagnose and customize the symptoms-specific therapy for them. The initiation of speech and language therapy is basic for the patient and family can practice by themselves, starting from the management of the breathing process and pronunciation training with tongue exercises repeatedly, gradually increasing the difficulty of pronunciation to practice correct sound and word formation [2,3].

The current trend is to use technology in medicine and speech therapy extensively. This includes technologies such as Virtual Reality and Artificial Intelligence that have been developed to analyze the patient's voice to assist in therapy and reduce the workload of rehabilitation therapists [4]. There are also speech therapy apps for stroke patients that use iPads with various tests to help patients train themselves [5]. Additionally, there is the use of Deep learning technology, specifically Convolutional Neural Network architecture, is used to classify organs and detect various diseases, such as optic disc and cup classification for glaucoma detection [6] or automatic classification of fetal facial ultrasound images [7]. Furthermore, there are also programs developed for physical rehabilitation of the hand, which can recognize hand gestures to promote exercise and hand and wrist conditioning [8].

This research presents the development of pronunciation classification, especially single vowel sounds in Thai from videos of ordinary people, to be adapted for those who want to practice correct pronunciation, particularly for stroke patients who have problems with the loss of the ability to communicate.

[+] Corresponding author. Tel.: +665-238-0322
E-mail address: 6370209921@student.chula.ac.th.

2. Related Work

Stroke is the syndrome of acute, focal neurological deficit attributed to vascular injury of the central nervous system. Possible complications of this syndrome include brain damage, cerebral palsy, physical paralysis, and impaired functioning of the organs of speech or hearing resulting in the loss of the ability to convey meaning. In speech therapy, the patient needs to be diagnosed by a physician and rehabilitated under the supervision of a speech therapist in order to best suit each patient. One of the speech therapy processes that the patient and family can do at home by themselves is breathing exercises, tongue strengthening exercises, and practicing speech sounds, for example, the practice of 5 English vowel sounds “a”, “e”, “i”, “o”, and “u” [2,3]. However, the kind of this therapy process can be different depending on the patient’s own language. In Thailand, the speech and language therapy method starts with breathing and tongue exercises, which are mostly the same as in other countries, but the difference is that the practice is based on Thai sounds which have 9 single pronunciation sounds, furthermore, Thai vowels are categorized into 3 groups by the shape of the mouth: group 1 is “ee”, “a”, and “ae” group 2 is “eu”, “er”, and “ar” group 3 is “oo”, “o”, and “or”. The patient needs to practice this process repeatedly, and the difficulty level of pronunciation should be gradually increased so that the patient can correctly regain their ability to pronounce sounds and words.

Machine learning technology has been developed for a long time and has been applied to various fields. For example, Babikier et al. (2017) have classified and detected human behaviors from closed-circuit television images using Artificial neural network techniques [9]. Similarly, Mekruksavanich et al. (2022) utilized deep learning networks to classify closely related sports activities [10]. Additionally, another field that has greatly benefited from the use of Deep learning technology is medicine and rehabilitation. Namely Fu H et al. team developed the U-shape convolutional network technique to classify the Optic Disc and Cup, which are important nerve structures used in diagnosing glaucoma [6]. Moreover, Maleewan Rungruanganukul & Thitirat Siriborvornratanakul (2020) used convolutional neural network techniques to classify hand images in order to develop a program for hand rehabilitation [8]. Another interesting study involved the use of robots and the C4.5 model to assist patients with Apraxia of Speech by recognizing mouth images that produce the sounds “a”, “u”, or is a closed mouth [13]. Previous work [6, 8] used video classification for rehabilitation and therapy. Recently, video classification is advanced; it shows the performance of classifying large-scale datasets, for example, TimeSformer [11] and SlowFast [12] models. Bertasius et al. (2021) utilized the TimeSformer [11], which is a deep learning model based on the state-of-the-art transformer, to classify an activity from a video. The model is trained with large-scale datasets of Kinetics-400 and Kinetics-600 datasets with accuracy of 80.7% and 82.2%. SlowFast presented by Fiechtenhofer et al. (2019) [12] processing with video recognition benchmarks such as Kinetics-400 and Kinetics-600 and achieved accuracy of 79.8% and 81.8% respectively.

Many researchers have experimented with video classification for various topics. Thus our work presents an automated model framework for identifying 9 different vowel sounds which is required for speech therapy in Thai language.

3. Dataset

We collected data from 100 participants who speak normally and in an age range of 20 to 70 years old. The data consists of videos of each participant's entire face. Each participant was required to record 9 video clips, one for each of the 9 Thai vowel sounds. Each video focuses on the pronunciation of a single vowel sound, resulting in 9 separate videos.

Fig. 1: single Thai vowel pronunciation

4. Method

After obtaining a video, it will be put into the pre-processing process to crop the video to only show the mouth area. Then, the resulting cropped video will be used in the model training and subsequently put into model validation.

Dataset Preprocessing

The Dlib technique [14] is a modern C++ toolkit containing a machine learning algorithm initially applied to each video frame to detect and crop a person's face. This technique can identify 68 facial landmarks, e.g., face structure, eyes, nose, etc. The 49th - 68th landmarks represent the mouth shape as shown in Fig. 2. Once the mouth area is detected, it is then cropped to remove any unnecessary information such as the eyes and nose. Each cropped image in the video sequence is resized to 224×224.

SlowFast Method

The pre-processed image sequence is then feed into the model which is the SlowFast model as architecture shown in Fig. 2. The model processes information using a single-stream architecture using two pathways, namely the slow path and the fast path. The slow path captures the semantics of the object and processes it with a large temporal stride using a convolutional network. The fast path captures the motion of the object and consists of a convolutional network with three key properties. Advantages include a higher frame rate, avoiding the transient oversampling to maintain good temporal resolution, and a lower channel capacity making it lightweight. After processing through the slow and fast paths, the results are entered into a fully connected layer to find the probability and predict the final class of the video.

5. Experiments

5.1. Experimental overview

We compare our work using the SlowFast method with the transformer-based state-of-the-art TimeSformer method [12]. TimeSformer Method Self-attention is a video processing technique that analyzes frames by dividing them into patches and comparing them to nearby frames to understand their relationships.The video dataset of 100 participants are divided into 70:30 for training and testing.

5.2. Experimental setting

The SlowFast model is utilized using ResNet-50 backbone and model pre-trained with ImageNet dataset. The parameters are configures as follows: (i) dropout ratio set to 0.4, (ii) total epochs is 100, (iii) number of clips as 25 clips, (iv) clip length as 1 frame, (v) frame interval of 1, (vi) resize scale as 224 ×224, and (vii) the stride as 16 frames in the Slow Pathway and 2 frames in the Fast Pathway. We compared model with TimeSformer which uses a divided space-time attention approach, with pre-trained model on ImageNet with the additional settings: (i) patch size of 16 patches, (ii) 15 epochs, (iii) clip length of 8 frames, (iv) frame interval of 32, (v) num clip of 1 clip, and (vi) a resize scale of 224×224.

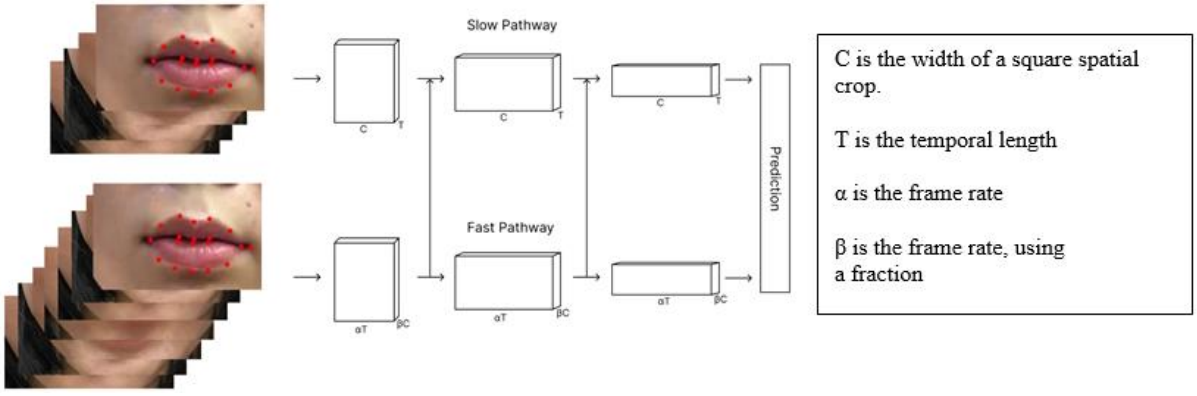

Fig. 2: SlowFast architecture

5.3. Experimental result

In our experiment, we use two types of videos, a full-face view (No Crop) and a mouth-only view (Crop) for processing with TimeSformer and SlowFast methods. The results are shown in Table 1. The best model

for classifying 9 vowel sounds from a pronunciation video is the SlowFast method with mouth-cropping processing with top-1 accuracy of training dataset is 85.4% when processing videos in mouth-only view and the highest top-1 accuracy achieved in testing the model is 97.3% when processing videos in mouth-only view. The result shows the difficulty of TimeSformer model in classifying 9-vowel sounds from pronunciation. Due to SlowFast divides the processing into slow and fast paths to capture semantic and motion information, making it suitable for videos with slight differences and movements only in certain areas. While the TimeSformer method performs temporal processing, it divides the frame's spatial area into patches to compare with similar frames, resulting in less accurate results.

Table 1: Vowel sounds classification performance

Method	Data	Training Top 1	Training Top 5	Test Top 1	Test Top 5
TimeSformer	No Crop	0.1185	0.563	0.1	0.5556
	Crop	0.1111	0.5643	0.1111	0.5643
SlowFast	No Crop	0.8095	0.9762	0.9444	0.9984
	Crop	**0.854**	**0.9778**	**0.973**	**1**

The model predictions across different classes are presented in a confusion matrix as shown in Fig. 3 and Fig. 4 for no-crop and crop versions of the SlowFast model. The most accurate model was SlowFast, which utilized mouth-only view image sequence (Crop), as shown in Fig. 4. The second-most accurate model was SlowFast, which processed data from the full-face view as shown in Fig. 3. The class that was predicted most accurately was group 3, which includes “oo”, “o”, and “or” sounds. However, the model made the most mistakes when predicting the wrong lip, but still within the same group, e.g., “eu” and “er” are mis-classified to “er” and “eu” in the same group 2.

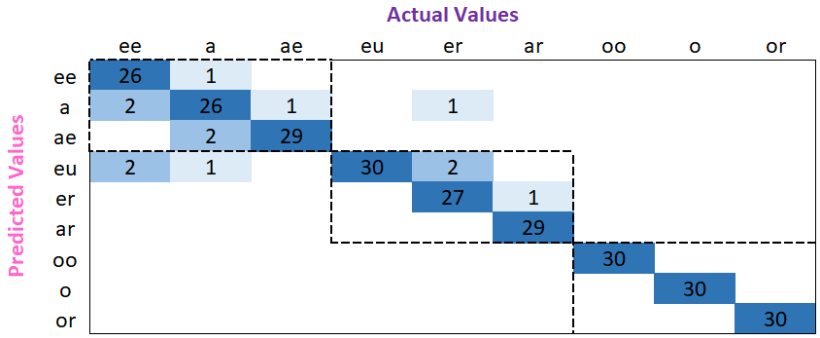

Fig. 3: Confusion matrix of testing data using SlowFast model with a full-face view (No Crop).

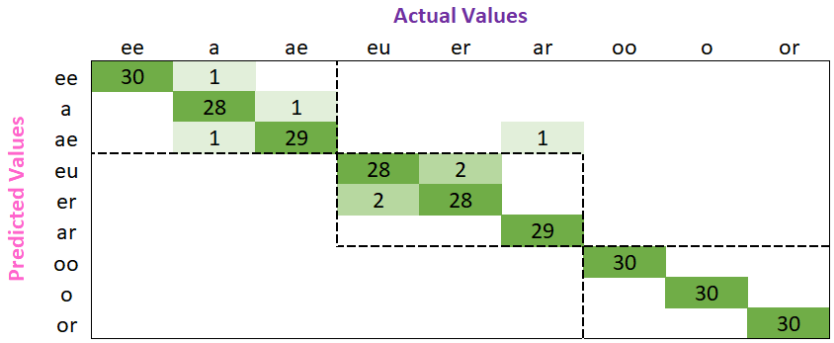

Fig. 4: Confusion matrix of testing data using SlowFast model with a mouth-only view (Crop).

6. Conclusion

A model framework is proposed for classifying videos of a single vowel sound pronunciation in the Thai language, which consists of 9 sounds. The two-step framework comprises person’s face detection using Dlib followed by a model for image sequence classification. We compared our select model, convolutional-based SlowFast with TimeSformer model, and compare results. The SlowFast outperforms TimeSformer with

accuracy of 97.3%. The small error aligns within the same group with similar pronunciation. Our research has a potential to be used for automatically identifying correct vowel-sound pronunciation for speech and language therapy for patients with aphasia resulting from stroke.

7. Acknowledgment

I would like to express my sincere gratitude to Dr. Channaya Thanaklang a speech therapist, for her invaluable guidance and support throughout this research study. Furthermore, we would like to express our gratitude to all the participants who took the time to participate in the data collection for this research.

8. References

[1] The Top 10 Causes of Death. Translated by WHO Global Health Estimates. 9 Dec. 2020; Available from: www.who.int/news-room/factsheets/detail/the-top-10-causes-of-death.

[2] Otr/L, Elizabeth Denslow. “The Best Speech Therapy Exercises to Regain the Ability to Speak.” Flint Rehab, January 17, 2023. https://www.flintrehab.com/Speech-Therapy-Exercises.

[3] Ben-Aharon, A. Top 5 Speech Therapy Exercises for Stroke Patients. 8 Aug. 2020.; Available from: https://greatspeech.com/Speech-Therapy-for-Stroke-Patients.

[4] Egaji, Oche A., Ikram Asghar, Mark Griffiths, and William Warren. "Digital speech therapy for the aphasia patients: Challenges, opportunities and solutions." In Proceedings of the 9th International Conference on Information Communication and Management, pp. 85-88. 2019.

[5] Stark, Brielle C., and Elizabeth A. Warburton. "Improved language in chronic aphasia after self-delivered iPad speech therapy." Neuropsychological rehabilitation 28, no. 5 (2018): 818-831.

[6] Fu, Huazhu, Jun Cheng, Yanwu Xu, Damon Wing Kee Wong, Jiang Liu, and Xiaochun Cao. "Joint optic disc and cup segmentation based on multi-label deep network and polar transformation." IEEE transactions on medical imaging 37, no. 7 (2018): 1597-1605.

[7] Shalev-Shwartz, Shai, and Tong Zhang. "Stochastic dual coordinate ascent methods for regularized loss minimization." Journal of Machine Learning Research 14, no. 1 (2013).

[8] Rungruanganukul, Maleewan, and Thitirat Siriborvornratanakul. "Deep learning based gesture classification for hand physical therapy interactive program." In Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Posture, Motion and Health: 11th International Conference, DHM 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I 22, pp. 349-358. Springer International Publishing, 2020.

[9] Babiker, Mohanad, Othman O. Khalifa, Kyaw Kyaw Htike, Aisha Hassan, and Muhamed Zaharadeen. "Automated daily human activity recognition for video surveillance using neural network." In 2017 IEEE 4th international conference on smart instrumentation, measurement and application (ICSIMA), pp. 1-5. IEEE, 2017.

[10] Mekruksavanich, Sakorn, and Anuchit Jitpattanakul. "Multimodal wearable sensing for sport-related activity recognition using deep learning networks." Journal of Advances in Information Technology (2022).

[11] Bertasius, Gedas, Heng Wang, and Lorenzo Torresani. "Is space-time attention all you need for video understanding?." In ICML, vol. 2, no. 3, p. 4. 2021.

[12] Feichtenhofer, Christoph, Haoqi Fan, Jitendra Malik, and Kaiming He. "Slowfast networks for video recognition." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 6202-6211. 2019.

[13] Castillo, José Carlos, Diego Alvarez-Fernandez, Fernando Alonso-Martin, Sara Marques-Villarroya, and Miguel A. Salichs. "Social robotics in therapy of apraxia of speech." Journal of Healthcare Engineering 2018 (2018).

[14] “Dlib C++ Library,” n.d., http://dlib.net/.