Classification of Myanmar Syllables Sound Using Image Classification Approach

Ou Ou Khin*[1], Ye Kyaw Thu[2], Tadashi Sakata[3], Yuichi Ueda[3]
[1]Graduate School of Science and Technology, Kumamoto University, Japan
[2]Language and Speech Science Research Lab., Waseda University, Japan
[3]Faculty of Advanced Science and Technology, Kumamoto University, Japan
*Corresponding author: ououkhin1994@gmail.com

Abstract- In recent years, both audio classification and image classification have been increasingly developed in research stage after the inception of artificial intelligence and deep learning. However, deep learning techniques usually require much more data and also computationally expensive than traditional algorithms. In order to reduce the development costs, we applied transfer learning using Google's pre-trained CNN model (Inception-v3) that is ready to be fine-tuned. The method is tested on our recorded database of 45 sound classes containing spectrograms of similar Myanmar syllable sounds and achieved recognition (validation test-set) accuracy of 90.70%. In the closed testing, our model obtained the classification accuracy with 75.78%, whereas, 38.00% accuracy in open testing. Our experiments show that the method is applicable to the classification of Myanmar syllables sound.

Index Terms-Audio classification; Image classification; Myanmar syllable; Transfer learning; Deep learning

I. Introduction

IN audio classification, it takes a sound sample as input and gives the respective class label as output. Audio can be represented in many ways, such as zero crossing statistics, fundamental frequency, spectral centroid, harmonicity, temporal envelope descriptions, chromagrams and spectrogram [1]. Among them, spectrogram, a visual representation of frequencies of sound, can be used for audio classification. In recent years, spectrogram-based audio classification using neural networks models has become the research interest in audio classification area. For example, Lu Lu et.al used Convolutional Neural Network (CNN) for acoustic scene classification [2]. Moreover, Boddapati et.al considered the classification accuracy and used image recognition networks for different image representations (Spectrogram, MFCC, and CRP) of environmental sounds [3]. Also, Venkatesh Boddapati et.al classified environmental sounds with Image network. Based on the literature study, classification of audio based on spectrogram images using deep network yields the best accuracy rates [4]. In this study, the Myanmar syllable sounds are treated with the CNN-model (inception-v3) to fit spectrogram images, performing the transfer learning from pre-trained weights on ImageNet to syllable classification. This paper is organized as follows: Section II describes the related works of the syllable-based speech recognition system. In section III, we present the brief introduction about Myanmar language, Inception-v3 and transfer learning. In section IV, we explain details of experimental setup. In section V, we discuss about the results in details, and finally, section VI concludes the paper.

II. Related Work

There are some works that have been done for speech recognition system based on syllable in other languages such as Turkish, Polish and Mandarin. On the other hand, in Myanmar language, there are a few researches done based on the syllables. Piotr Majewski et.al described that syllables are fruitful sub-word units in language modeling of Polish language and also the syllable-based model is adequate for language modelling in many cases like small available corpora or highly inflectional language [5]. Wunna Soe and Yadana Thein presented syllable-based speech recognition system for Myanmar language with HMM [6]. Moreover, Hay Mar Soe Naing et.al investigated the automatic speech recognition performance differences between the word-based and syllable-based language models for Myanmar language. In this paper, the architecture of a Myanmar large vocabulary continuous speech recognition system was presented in details [7]. Still, there is no research done for classifying and recognizing Myanmar syllable sounds based on image classification. Our experiment is the first such system for Myanmar language. We explored the Myanmar syllables audio classification system from learning the nature and features of spectrograms of each syllable sound using pre-trained Inception-v3 (CNN) model.

III. Background information

A. Myanmar syllable

Myanmar language, also known as Burmese language, is the Sino-Tibetan language spoken in Myanmar as an official language by 33 million people and as second language by 10 million people. The Myanmar script consists of 33 basic consonants, 12 vowels, 4 basic medials, other symbols and special characters. However, there are only 23 distinct pronunciation for consonants since some consonants have the same pronunciation in Myanmar language. For example, "ဒ", "ဓ", "ဍ" and "ဎ", share the same pronunciation. Myanmar syllables are generally composed of consonants and (zero or more) vowel combinations starting with a consonant. For example, in this word မိန်းမ (min: ma), there are two syllables. The first syllable is formed with the combination of consonant မ (ma) with dependent vowel ◌ိ (i), consonant န (na) and killer ◌် (asat). The second one is just consonant မ (ma).

B. Inception-v3

In Google, there are many neural network models that have been made publicly available for use in TensorFlow [8]. In our experiment, Inception-v3 was used for transfer learning. Inception-v3, a convolutional neural network, is trained on more than a million images from the ImageNet database. The Inception-v3 model has achieved 78.00% top-1 and 93.90% top-5 accuracy on the ImageNet test dataset [9]. Moreover, the network is 48 layers deep and consists of two parts: (1) feature extraction part with a convolutional neural network, and (2) classification part with fully-connected and softmax layers. In the first part, the model extracts general features from input images and classifies them based on those features in the second part. The architecture of inception-v3 was explained in [10].

C. Transfer learning

Transfer learning is a machine learning technique where the knowledge gained during training in one problem is used to train in other similar type of problem. In transfer learning, the base network and task are trained on a base dataset, and then repurpose the learned features, or transfer them, to a second target network to be trained on a target dataset and task. For deep learning, the first few layers are trained to identify features of the problem. During transfer learning, it can replace the last layer with the desired dataset. For our experiment where the problem is to automatically classify the Myanmar syllables, we need to collect a large amount of labeled data due to train the syllable-classification models for each syllable. However, it is computationally expensive and require much time to get the train model. In such case, transfer learning can help in training neural networks with considerably less amount of time. In Fig.1., the architecture of transfer learning for Myanmar syllables classification was explained.

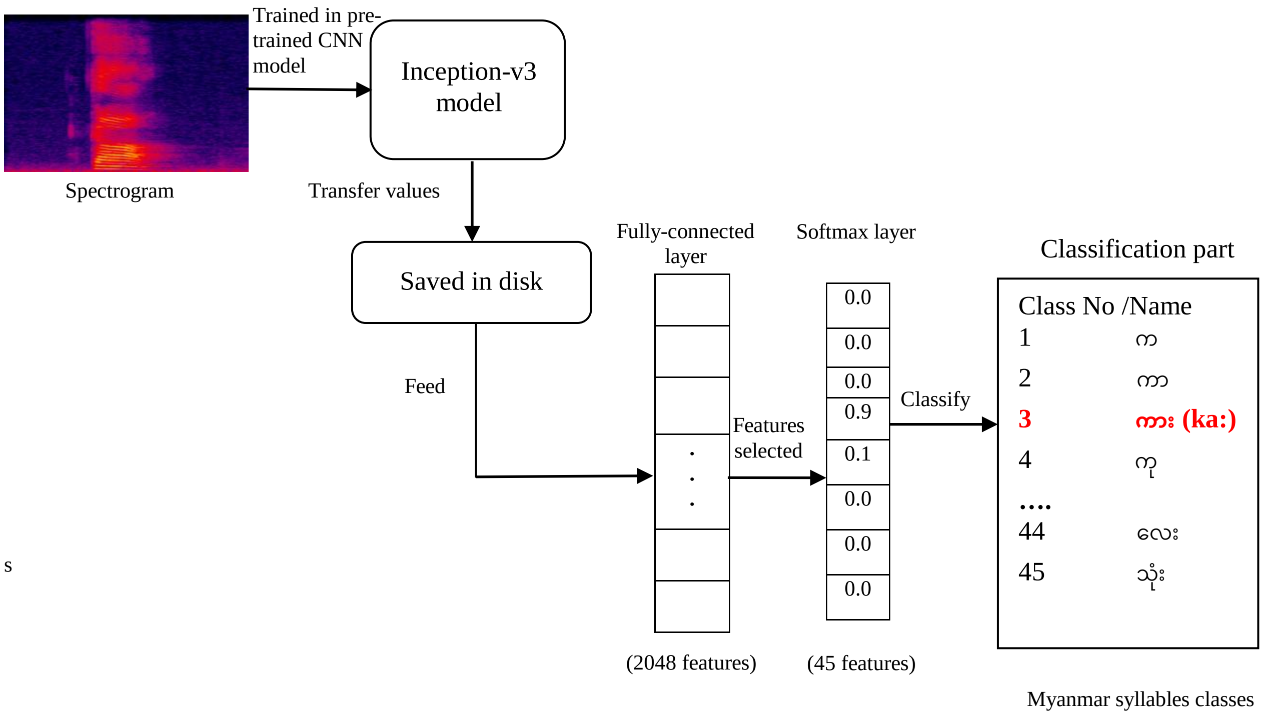


Fig. 1. Architecture of transfer learning for Myanmar syllables classification

IV. Experimental Setup

A. Data pre-processing

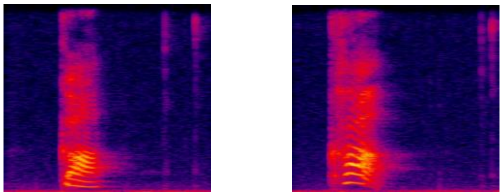
Among 3,000 Myanmar syllables, we basically selected 45 syllables, which have similar sounds, as our experimental dataset. For instance, စ (s) and ဆ (hs) share the very similar phonemes. Firstly, these 45 syllables were recorded by female speaker using the airpods (wireless Bluetooth earbuds produced by Apple). For each syllable, there were 60 audio files and each file has duration of one second. Then, the recorded audio data was downsampled from the sampling rate of 44kHz to 16kHz with mono channel.

B. Audio Featuring

The data representation is a decisive step in any learning process. In our experiment, the input audio files were represented in the form of visual images (spectrograms). Although there are other visual forms for audio representation, we used the spectrogram since it can be used to identify spoken words phonetically. The examples of spectrograms of

Myanmar syllables are shown in Fig.2.

The spectrogram was extracted by using the Sox (Sound eXchange, Swiss Army knife of sound processing programs) command line utility [11]. In each syllable class, there are 60 spectrograms obtained. Among these spectrograms, we selected all images as training data. For the testing, 10 spectrograms were randomly picked out from each syllable class. For the closed testing, 10 spectrograms were randomly chosen from the training data, whereas, 10 spectrograms obtained from another audio files that were recorded again were took as open-testing data.



(a) Spectrogram of "က" (k) (b) Spectrogram of "ခ" (kh)

Fig. 2. Spectrograms of Myanmar syllables

C. Training the model

In the training stage, we used Google's pretrained CNN model (Inception-v3). The pre-trained model was loaded and trained a new classifier on top for the syllable spectrograms. The first step was to analyze all the images, and calculated and then saved the bottleneck values for each image to disk. In this stage, this penultimate layer was trained in order to output a set of values that is good enough for the classifier. In our experiment, we run 20,000 training steps. In each step, images were selected randomly from the training set, we found their bottlenecks from the cache, and fed them into the final layer to get the predictions. Then, we compared those predictions against the actual labels to update the final layer's weights through the back-propagation process. Training steps are based on the AudioNet, opensource speaker recognition experiment using tensorflow framework and Google's Inception model [12].

D. Testing the model

In the testing stage, using the retrained model, we tested the syllable-classification of 45 Myanmar syllables for both closed testing and open-testing (with the other recorded audio files of the same speaker). As mentioned above, 10 spectrograms from the dataset used in training were randomly chosen for closed testing. Another 10 spectrograms were picked out from second audio dataset.

V. RESULTS AND DISCUSSION

For the 45 Myanmar syllable sound classes, we achieved the test classification accuracy of 90.70% with our trained model. In the closed test, the classification accuracy was about 75.78%. In contrast, according to the results, the accuracy was 38.00% in open-test. Moreover, we divided the syllables into 14 pairs in which each syllable shares the similar tone within each pair. The 14 pairs of Myanmar syllables were shown in TABLE 1. The accuracy of classification results for these pairs for both closed testing and open-testing were shown in Fig.3. According to the Fig.3., it can be assumed that our model could not classify the syllables in pair 5 and 12 in open-test.

TABLE 1:14 PAIRS OF SIMILAR MYANMAR SYLLABLES

Pair No.	Class No	Myanmar Syllables /IPA Format
1	1, 11, 21	က, ခ, ဂ /k, kʰ,g
2	26, 27, 28	စ, ဆ, ဇ / s,sʰ,z
3	29, 31, 32	တ, ထ, ဒ / t,tʰ,d
4	34, 38, 39	ပ, ဖ, ဘ /p,pʰ, b
5	1, 2, 3	က, ကာ, ကား /k, ka, kà
6	11, 12, 13	ခ , ခါ ,ခါး /kʰ, kʰa, kʰà
7	4, 5, 6	ကု, ကူ ,ကူး /kú ,ku, kù
8	41, 42, 43	လု , လူ ,လူး /lú , lu, lù
9	14, 16, 17	ခံ , ခန့် , ခန်း /kʰàɴ, kʰa̰ɴ , kʰáɴ
10	8, 15, 22	ကက်, ခက် ,ဂက် /kɛʔ, kʰɛʔ, gɛʔ
11	10, 19, 24	ကြက်, ချက်, ဂျက်/**tɕ**ɛʔ, **tɕ**ʰɛʔ, g**ɕ**ɛʔ
12	9, 18, 23	ကျ, ချ ,ဂျ /tɕə, **tɕ**ʰə, gy
13	35, 36, 37	ပန်, ပန့် , ပန်း /pã, pa̰ɴ , páɴ
14	30, 33, 40	တစ်, နှစ်, ရှစ် /tiʔ, n̥iʔ, ʃiʔ

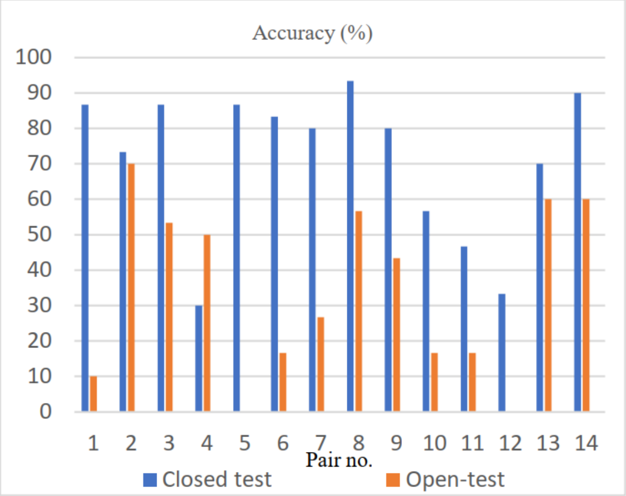
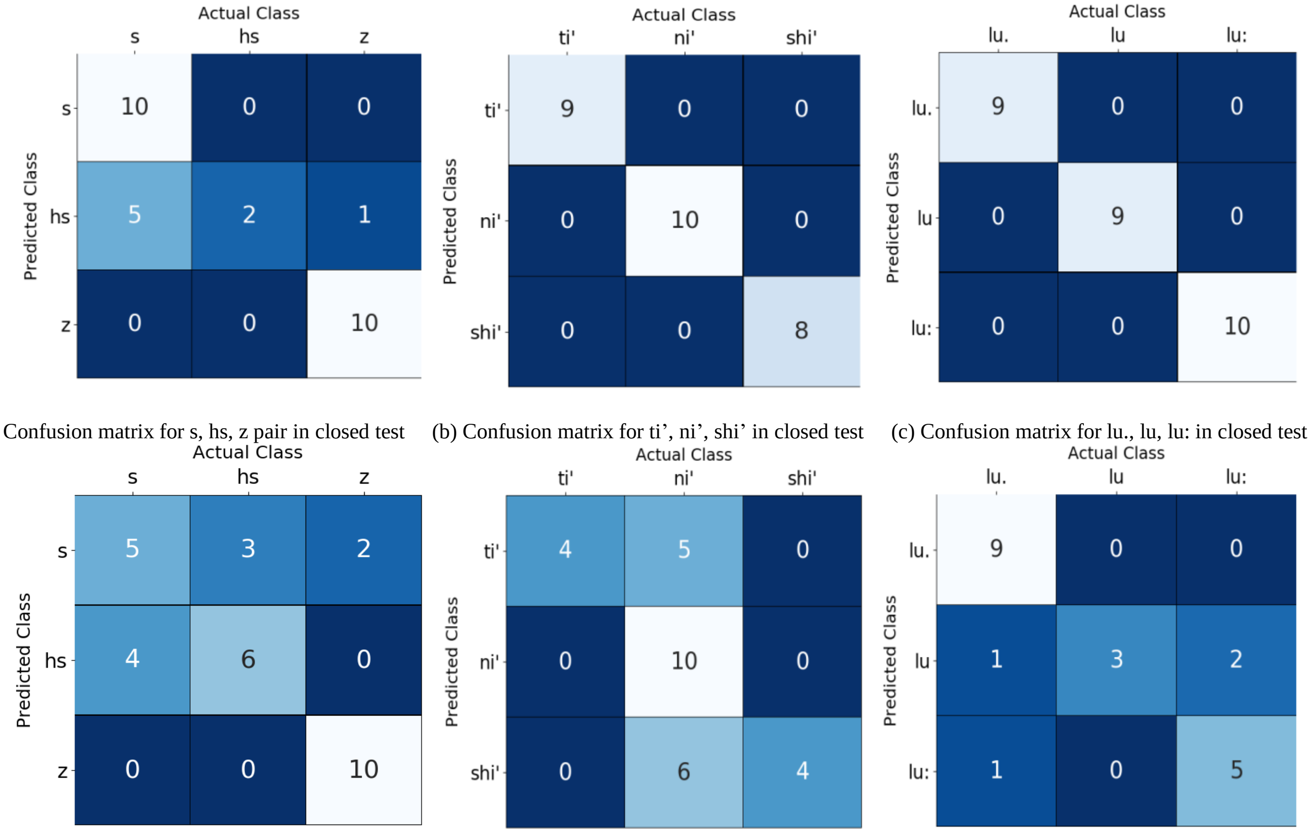


Fig. 3. Accuracy of classification results for 14 pairs in closed testing and open-testing

Moreover, for these 14 pairs of syllables, the classification results were shown as the confusion matrix as shown in Fig.4. According to the confusion matrix for each pair's results, in open-test, the optimizing pair that the model could recognize correctly was 2^{nd} pair (စ, ဆ, ဇ), with 5 times, 6 times, and 10 times for each class respectively, and it classified wrongly within its pair. In contrast, in closed test results, it could classify fully 10 times as စ and ဇ, but for ဆ, the correct classification time was only 2, with 5 times as စ, as ဇ, နှစ် and ဖ for 1 time exactly.

The second finest pair in both open and closed test was pair-14 (တစ်, နှစ်, ရှစ်). In open test, it could classify correctly for 4, 10, and 4 times for each class respectively and wrongly through pair and as ယ for 1 time only when it was recognizing class-30 (တစ်).

Third best classification pair in open-test was pair-8 (လု, လူ, လူး). While classifying လု, it wrongly classified as ဘ once, and as ငါး for four times when classifying လူ, and လူး. However, in closed test, our model could classify this pair best, with classifying wrongly as ဓ just a single time for လု and လူ.



(a) Confusion matrix for s, hs, z pair in closed test (b) Confusion matrix for ti', ni', shi' in closed test (c) Confusion matrix for lu., lu, lu: in closed test

(d) Confusion matrix for s, hs, z pair in open-test (e) Confusion matrix for ti', ni', shi' in open-test (f) Confusion matrix for lu., lu, lu: in open-test

Fig.4. Examples of confusion matrix for closed test and open-test results

VI. CONCLUSION

Through the experiment, the method of audio classification with image classification is relevant to the classification of Myanmar syllable sounds. In the future, this proposed system will be extended with more Myanmar syllables dataset and more than one speaker.

VII. REFERENCES

[1] Lonce Wyse, "Audio spectrogram representations for processing with Convolutional Neural Networks", arXiv:1706.0959v1 [cs.SD] 29 Jun 2017.

[2] Lu Lu, Jiang Yuzhi, Zhang Huiyu, Yang Yuhong, Hu Ruimin, Ai Haojun, Tu Weiping, Huang Weiyi, "Acoustic scence classification based on convolutional neural network", Detection and Classification of Acoustic Scenes and Events 2017.

[3] Venkatesh Boddapati, Andrej Petef, Jim Rasmusson, Lars Lundberg, "Classifying environmental sounds using image recognition networks", International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille, France.

[4] Venkatesh Boddapati, "Classifying Environmental Sounds with Image Networks", Master of Science in Computer science, February 2017, pp.8.

[5] Piotr Majewski (2008), "Syllable Based Language Model for Large Vocabulary Continuous Speech Recognition of Polish", University of Lodz, Faculty of Mathematics and Computer Science ul. Banacha 22, 90-238 Lodz, Poland, P. Sojka et al. (Eds): TSD 2008, LNAI 5246, pp. 397-401.

[6] Wunna Soe, Dr. Yadana Thein, "Syllable-based Speech Recognition System for Myanmar", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.5, No.2, April 2015.

[7] Hay Mar Soe Naing, Aye Mya Hlaing, Win Pa Pa, Xinhui Hu, Ye Kyaw Thu, Chiori Hori, Hisashi Kawai, "A Myanmar Large Vocabulary Continuous Speech Recognition System, 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA).

[8] https://www.tensorflow.org/

[9] https://mdeium.com/@williamkoehrsen/facil-recognition-using googlesconvolutional-neural-network-5aa752b4240e

[10] Christian Szegedy, Vincent Vanhoucke, Sergey Loffe, Jonathon Shlens, Zbigniew Wojna, "Rethinking the Inception Architecture for Computer Vision", arXiv:1512.00567v3 [cs.CV], 11 Dec 2015

[11] https://sox.sourceforge.net

[12] https://github.com/vishnu-ks/AudioNet/