Myanmar Speech Classification
Using Transfer Learning for Image Classification

Ou Ou Khin[1], Ye Kyaw Thu[2], Tadashi Sakata[3], Yoshinori Sagisaka[2], Yuichi Ueda[3]
[1]Graduate School of Science and Technology, Kumamoto University, Japan
[2]Language and Speech Science Research Laboratory, Waseda University, Japan
[3]Faculty of Advanced Science and Technology, Kumamoto University, Japan
ououkhin1994@gmail.com, wasedakuma@gmail.com, tadashi@cs.kumamoto-u.ac.jp,
ysagisaka@gmail.com, ueda@cs.kumamoto-u.ac.jp

Abstract

In this paper, our research on speech classification using an image classification approach is discussed for the Myanmar language. We tested the method for Myanmar consonants, vowels, and words, on our recorded database of 22-consonant, 12-vowel, and 54-word sound classes, containing spectrograms of Myanmar speech. Because Myanmar language is tonal, the sounds are very similar for precise classification based on audio features, while the visual representations differ. Therefore, it is important to consider the visual representations of audio in classifying the Myanmar language. In this study, we treated Myanmar speeches with a convolutional neural network model (Inception-v3) to fit spectrogram images, performing transfer learning from pre-trained weights on ImageNet. Validation accuracies of 60.70%, 73.20%, and 94.60% were achieved for the consonant, vowel, and word-level classifications, respectively. In order to determine the retrained model performance, both closed and open testing were conducted. Although our experiment was distinct from other traditional audio classification methods, promising results were obtained for the first exploration of Myanmar speech classification using transfer learning for image classification. In fact, these experimental results were attained using Google's Inception-v3 model, constructed with different image domains. Therefore, the research and results demonstrate that it is possible to perform Myanmar speech classification.

1. Introduction

In recent years, both audio and image classification have been increasingly developed in the research stage, following the inception of artificial intelligence and deep learning. Moreover, numerous classification methods, such as the hidden Markov model (HMM), Gaussian mixture model, artificial neural networks, and fuzzy logic, have been applied for speech classification in previous researches. For example, Soe and Thein presented a syllable-based speech recognition system for the Myanmar language with an HMM [1]. Furthermore, Khaing investigated a Myanmar continuous speech recognition system using dynamic time warping and the HMM [2]. However, deep learning techniques usually require substantially more data and are more computationally expensive than traditional algorithms.

Furthermore, several studies have been conducted on spectrogram-based audio classification using neural network models [3, 4, 5], and this has become a research interest in the audio classification area. In fact, there are various means of representing audio, such as zero crossing statistics, fundamental frequency, spectral centroid, harmonicity, temporal envelope descriptions, chromagrams, and spectrograms [6]. Based on the literature studies, it was found that audio classification based on spectrogram images and using deep networks yields the highest accuracy rates.

This paper presents the classification of Myanmar speech by using a convolutional neural network (CNN) model (Inception-v3) to fit spectrogram images, performing transfer learning for Myanmar speech classification in order to reduce development costs. No research has been conducted on classifying and recognising Myanmar sounds based on image classification. Our experiment is the first such system for the Myanmar language. We explored the Myanmar speech audio classification system by learning the nature and features of the spectrograms of each syllable and word sound, using the pre-trained Inception-v3 (CNN) model.

The remainder of this paper is organised as follows: Section 2 describes the nature of the

Myanmar language. In section 3, we describe the methodology that we used for classification of the Myanmar language. In section 4, we present details of the experimental setup. In section 5, we discuss the results obtained in detail; finally, section 6 concludes the paper.

2. Nature of Myanmar Language

2.1. Myanmar Language

Myanmar language, also known as the Burmese language, is the Sino-Tibetan language spoken as an official language by approximately 33 million people and as a second language by 10 million people in Myanmar. Moreover, Myanmar is a tonal language, which means that a syllable or word changes along the tone. The Myanmar script consists of 33 basic consonants, 12 vowels, 4 basic medials, and other symbols and special characters. However, only 23 distinct pronunciations exist for consonants, and certain consonants share the same pronunciation in the Myanmar language. For example, "ဓ", "ဍ", and "ဒ" have the same pronunciation, "da.". The Myanmar script is written from left to right. Conventionally, although sentences in the Myanmar script are delimited by sentence boundary markers, there are no white spaces between words, as in English. However, in model writing, spaces are used between words to provide readability.

2.2. Myanmar Syllables

The Myanmar script is generally syllabic in nature, consisting of sequences of syllables. In the Myanmar language, the syllable is the smallest linguistic unit, and it can generally be assumed that one word consists of one or more syllables. Moreover, in our experiment, because we aimed to perform speech classification, it was desirable to test the effectiveness of the results obtained from experiments using transfer learning for image classification intended for syllables and words, rather than at the sentence level. Myanmar syllables are composed of consonants and (zero or more) vowel combinations starting with a consonant. In general, at least one major syllable must exist in a Myanmar word. For example, in the word မိန်းမ (mein: ma.), there are two syllables. The first syllable is formed by the following combination: consonant မ (ma.) with dependent vowel ◌ိ (i), consonant န (na.), killer ◌် (asat), and ◌း (visarga). The second is a consonant မ (ma.) only.

3. Methodology

3.1. Inception-v3

In Google, numerous neural network models have been made publicly available for use in TensorFlow [7]. In our experiment, Inception-v3 was used for the transfer learning. It was released as the 2015 iteration of Google's Inception architecture for image recognition. Inception-v3, which is a CNN, was trained on more than one million images from the ImageNet database. The Inception-v3 model achieved 78.00% top-1 and 93.90% top-5 accuracy on the ImageNet test dataset [8]. Moreover, the network is 48 layers deep and consists of two parts: (1) feature extraction with a CNN, and (2) classification with fully connected and softmax layers.

In the first part, the model extracts general features from the input images; in the second part, it classifies these input images based on those features. Therefore, the first part involves pre-processing only, and it is only necessary to train the second part. The architecture of Inception-v3 is explained in [9] and illustrated in Figure 1.

3.2 Transfer Learning

Transfer learning is a machine learning technique whereby the knowledge gained during training in one problem is used for training in another, similar type of problem. In transfer learning, the base network and task are trained on a base dataset, following which the learned features are repurposed on a target dataset and task. In deep learning, the first several layers are trained to identify problem features. During transfer learning, the final layer can be replaced with the desired dataset. For our experiment, in which the problem was to classify Myanmar speech automatically, we needed to collect a large amount of labelled data for training the sound classification models for each consonant, vowel, and word. However, it is expensive and requires substantial time to obtain a trained model. In such cases, transfer learning can aid in training neural networks in considerably less time. In Figure 2, the architecture of the transfer learning for the Myanmar speech classification is explained. According to Figure 2, the Myanmar word "သရက်သီး" was recognised using transfer learning for image classification.

4. Experimental Setup

In this section, the details of the experimental setup for the Myanmar speech classification are described. The experiment consisted of four main parts: data pre-processing, audio featuring, training, and testing.

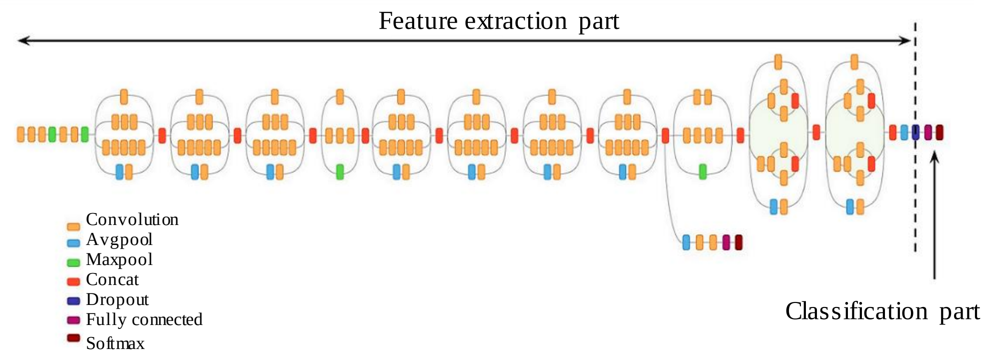


Figure 1. Inception-v3 architecture [9]

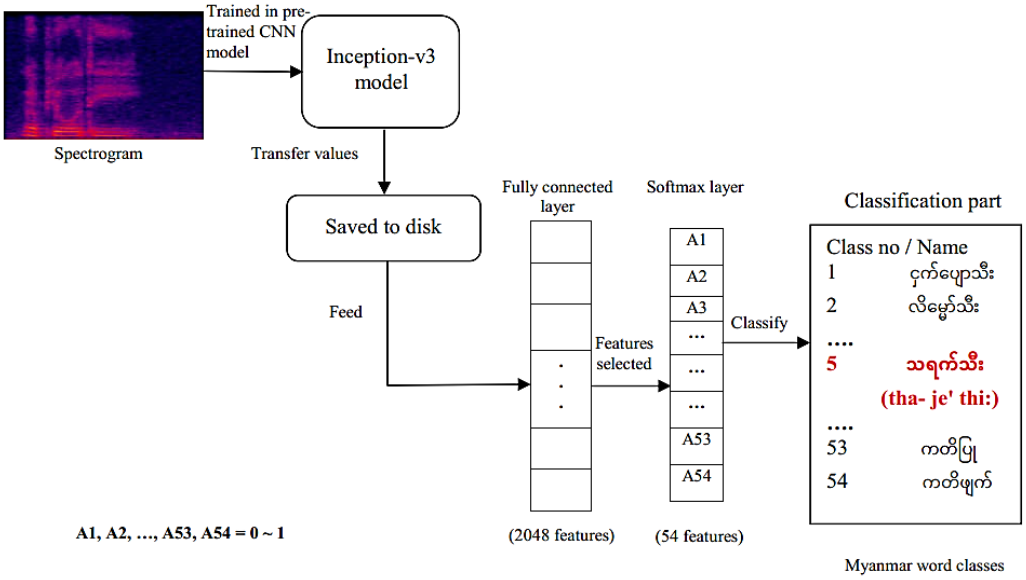


Figure 2. Architecture of transfer learning for Myanmar speech classification

4.1. Data Pre-processing

For the 22 consonants and 12 vowels, we prepared audio files by recording 45 female and 37 male speakers, including native speakers as well as other national ethnic races such as Kachin, Shan, and Rakhine, using a MacBook Air built-in microphone. For each consonant, there were approximately 200 audio files, and 332 files for each vowel class. Moreover, it was observed that certain races cannot pronounce certain consonants correctly. For example, some "Shan" speakers cannot pronounce the consonant "သ" (tha.) correctly. This could cause certain effects when classifying the consonants.

For the 54 words, audio files were prepared by recording four speakers, including one half-Chinese and three native speakers, using AirPods (wireless Bluetooth earbuds produced by Apple). There were 60 audio files for each word class.

The duration of each audio file was approximately one second. All recorded audio files were down-sampled from a sampling rate of 44 kHz to 16 kHz with a mono channel.

4.2. Audio Featuring

Data representation is a crucial step in any learning process. In our experiment, the audio files were represented in the form of visual images (spectrograms). Although other visual forms of audio representation are available, we used the spectrogram because it can be used to identify spoken words phonetically. In general, a spectrogram is a visual representation of sound frequencies that is used in music, sonar, radar, and speech processing. Spectrograms can be created by an optical spectrometer, band-pass filters, and Fourier transform.

Here, the spectrograms of the audio files were extracted using the Sound eXchange (SoX, Swiss Army knife of sound processing programs) command line utility [10]. Several examples of the word spectrograms are presented in Figure 3. In the figure, the spectrograms of ဝက်သား (we' tha:) and ကြက်သား (kye' tha:) appear alike as they have a similar tone, and may be difficult to distinguish.

4.3. Training

In the training stage, we used Google's pre-trained CNN model (Inception-v3). The pre-trained model was loaded and a new classifier was trained on top for the sound spectrograms. The first step was to analyse all of the images, following which the bottleneck values for each image were calculated and saved to disk. During this stage, this penultimate layer was trained to output a set of values that was sufficient for the classifier. In our experiment, we ran 20,000 training steps for the word-level classes' and 25,000 training steps for both the consonants and vowels. In each step, images were selected randomly from the training set, their bottlenecks were identified from the cache, and they were fed into the final layer to obtain the predictions. Thereafter, we compared the predictions against the actual labels to update the final layer weights through the back-propagation process. The training steps were based on the AudioNet open-source speaker-recognition experiment using the TensorFlow framework and Google's Inception model [11].

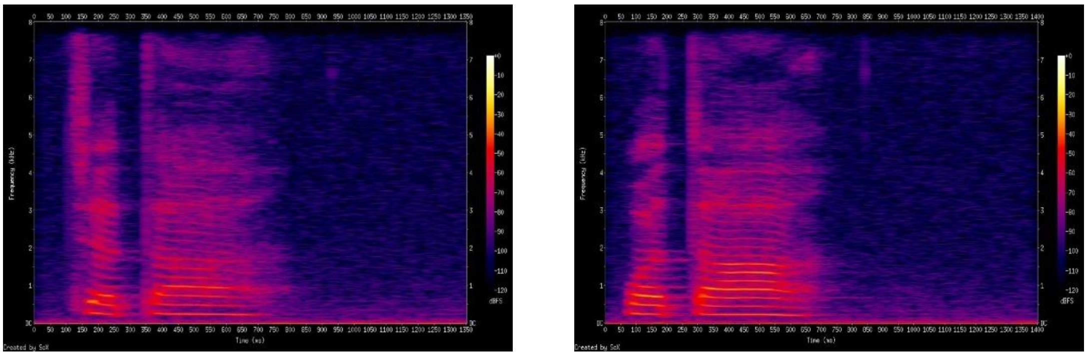
4.4. Testing

4.4.1. Experimental Setting for Consonants and Vowels

In the testing stage, using the retrained model, we tested the classification of 22 Myanmar consonants and 12 vowels using both closed and open tests. In the closed testing, spectrograms from the dataset, recorded by multiple speakers used in the training, were randomly selected. In the open testing, other spectrograms of different audio files were selected at random. We performed 20 classifications for each consonant class and 10 for each vowel class.

4.4.2. Experimental Setting for Words

For the word level, although the closed-test setup was similar to that of the consonants and vowels, the open testing was different. In open test, we selected spectrograms of other speaker audio files in order to confirm the manner in which our experimental approach could perform correct classification in the case of using one speaker's audio files for training and different speakers' audio files for testing. For both tests, 10% of the data was used for testing. Therefore, we performed classification six times for each word class.



(a) Spectrogram of "ကြက်သား" (kye' tha:) **(b) Spectrogram of "ဝက်သား" (we' tha:)**

Figure 3. Spectrograms of Myanmar words

5. Results and Discussion

We achieved validation test classification accuracies of 60.70%, 73.20%, and 94.60% for the 22-consonant, 12-vowel, and 54-word sound classes, respectively, when testing Myanmar speech classification using transfer learning for image classification.

5.1. Consonant Classification

For the consonant classification, the validation test accuracy was 60.70%. In the closed test, the classification accuracy was approximately 38.18%. In contrast, according to the results, the accuracy was 34.55% in the open test.

In the closed testing, the class most correctly classified by our model was class 8, "ည/ဉ" (nya), at 20 times, and the second was class 13, "ပ" (pa.), at 18 times. Unfortunately, our model could not classify class 10, "ထ/ ဌ" (hta.).

In the open test, class 13, "ပ" (pa.) was classified the most correctly in our experiment, with 19 out of 20 times, while the second was class 8 "ည/ဉ" (nya), with 17 times. Furthermore, the classes that our model could not classify were class 4 "င", class 10 "ထ/ ဌ" (hta.), and class 11 "ဍ/ဎ/ဒ/ဓ" (da.).

Furthermore, we divided the consonants into five pairs, where syllables shared a similar tone within each pair. The five pairs of Myanmar consonants are displayed in Table 1.

A comparison of the classification accuracies for these five pairs in the closed and open testing is presented in Figure 4. According to Figure 4, it can be assumed that our model could classify pair 5 most correctly and pair 3 least correctly in both the closed and open testing.

Table 1. Five pairs of similar Myanmar consonants

Pair no.	Class no.	Myanmar consonants / IPA format
1	1, 2, 3	က, ခ, ဂ /ဃ / k, kʰ, g
2	5, 6, 7	စ, ဆ, ဇ /ဈ / s, sʰ, z
3	9, 10, 11	တ/ ဋ, ထ/ ဌ, ဍ/ဎ/ဒ/ဓ / t, tʰ, d
4	4, 8, 12	င, ည/ဉ, ဏ/န / ŋ, ɲ, n
5	13, 14, 15	ပ, ဖ, ဗ/ဘ / p, pʰ, b

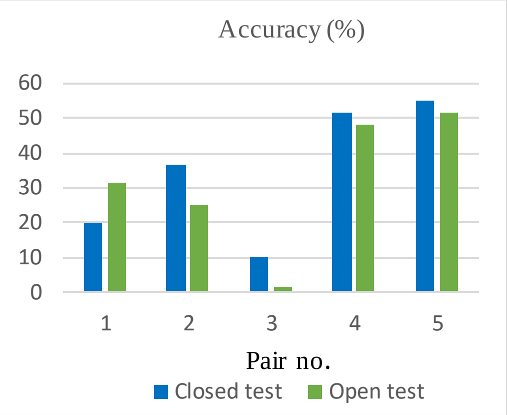
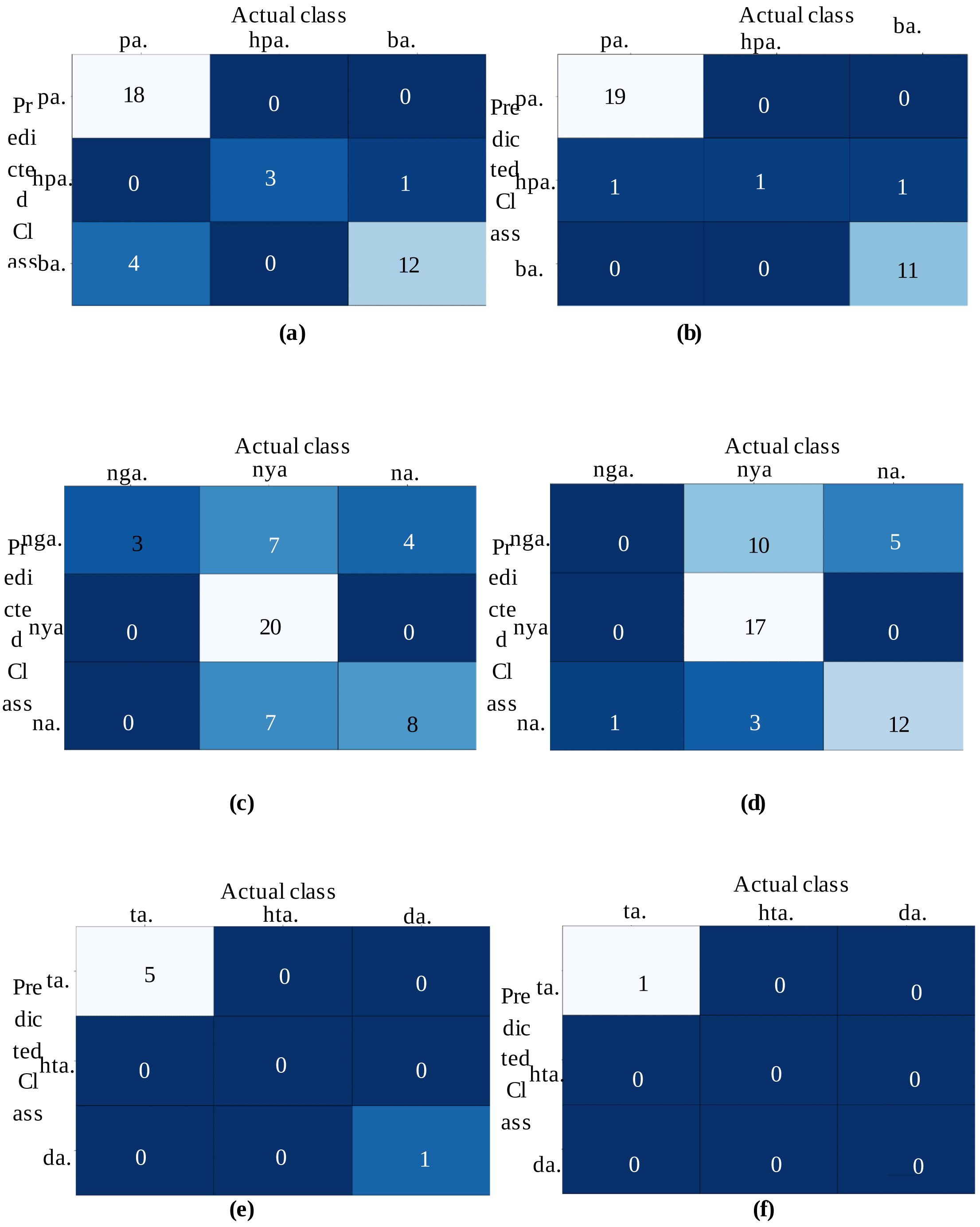


Figure 4. Comparison of consonant classification accuracies for five pairs in closed and open tests

Moreover, for these five consonant pairs, the classification results were presented as a confusion matrix in order to determine the performance, as illustrated in Figure 5. According to the confusion matrix for the results of each pair, the pair that the model could recognize most correctly was pair 5 (ပ, ဖ, ဗ/ဘ), with 19 times, one time, and 11 times in the open test and 18 times, 3 times, and 12 times in the closed test for each class, respectively. In contrast, the pair that our model classified incorrectly numerous times was pair 3 (တ/ ဋ, ထ/ ဌ, ဍ/ဎ/ဒ/ဓ). In the closed test, our model classified class 9 correctly only five times, and class 11 correctly once. However, our model could not classify other classes within the third pair correctly, with the only exception being class 9. The second-best pair in both the open and closed test was pair 4 (င, ည/ဉ, ဏ/န). However, in the open test, it was classified incorrectly as other classes when classifying class 4.



Predicted Class	pa.	hpa.	ba.
pa.	18	0	0
hpa.	0	3	1
ba.	4	0	12

(a)

Predicted Class	pa.	hpa.	ba.
pa.	19	0	0
hpa.	1	1	1
ba.	0	0	11

(b)

Predicted Class	nga.	nya	na.
nga.	3	7	4
nya	0	20	0
na.	0	7	8

(c)

Predicted Class	nga.	nya	na.
nga.	0	10	5
nya	0	17	0
na.	1	3	12

(d)

Predicted Class	ta.	hta.	da.
ta.	5	0	0
hta.	0	0	0
da.	0	0	1

(e)

Predicted Class	ta.	hta.	da.
ta.	1	0	0
hta.	0	0	0
da.	0	0	0

(f)

Figure 5. (a) Confusion matrix for pa., hpa., and ba. pair in closed test; (b) confusion matrix for pa., hpa., and ba. pair in open test; (c) confusion matrix for nga., nya, and na. pair in closed test; (d) confusion matrix for nga., nya, na. pair in open test; (e) confusion matrix for ta., hta., and da. pair in closed test; and (f) confusion matrix for ta., hta., and da. pair in open test

5.2. Vowel Classification

In the vowel classification, the validation test accuracy was 73.20%. The list of 12 vowels that we classified is displayed in Table 2, with the class numbers and IPA formats. When our model was tested in the closed test, the overall classification accuracy was approximately 61.67%, while in the open test, the accuracy was approximately 60.00%.

In the closed testing, our model could classify class 4 “အု” (u) and class 2 “အိ” (i) best, at 9 out of 10 times. Conversely, our model could classify class 9 “အော်” (o), and class 5 “အူ” (ū) only once, and twice with relevance.

Table 2. List of vowels

Class no.	Vowels / IPA format	Class no.	Vowels / IPA format
1	အာ / á	7	အဲ / ɛ̀
2	အိ / í	8	အော့ / ɔ̀
3	အီ / i	9	အော် / ɔ
4	အု / ú	10	အံ / àN
5	အူ / u	11	အား / a
6	အေ / e	12	အက် / ʔɛʔ

In the open test, the classes that our model could classify completely were class 4, "အု" (u), while class 11, "အား" (a:) was the second finest with 9 times correct classification. The classes that our model classified least correctly in both tests were classes 5 and 9.

In both tests, our model classified class 3 "အီ" (i) incorrectly many times when classifying the other vowels. Figure 6 presents a comparison of the classification accuracy results of the 12 vowels for both the closed and open testing.

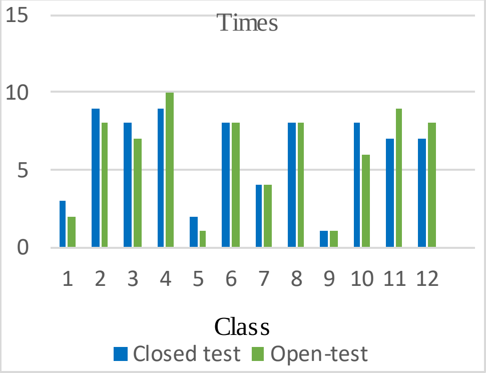


Figure 6. Comparison of vowel classification accuracies in closed and open test

5.3. Word-level Classification

In the word-level classification, we obtained a validation test accuracy of 94.60%. With our model, the test accuracy in the closed test was approximately 81.80%, while we achieved only 31.17% in the open test. In this case, it is significant that a large difference value was obtained in the closed and open test accuracies (81.80% and 31.17%). The reason for this is that we used the spectrograms of one speaker's audio files in the training, and selected spectrograms of different speakers' audio files in the open testing.

In both tests, our model classified classes 4, "စပျစ်သီး" and 7, "ကြက်သား" incorrectly many times while classifying other word classes.

According to the closed test results, the total number of classes that our model could classify perfectly was 22. However, our model classified class 1, "ငှက်ပျောသီး" correctly only once, and classes 5, "သရက်သီး" and 6, "သစ်တော်သီး" only twice.

In terms of the open testing, the optimally classified class was class 4, "စပျစ်သီး". The second-highest classified classes, with 4 times, were classes 2, 10, and 14. Unfortunately, classes 17, 18, 24, 28, 35, 41, 51, 53, and 54 could not be classified correctly by our trained model.

Moreover, words in which syllables shared a similar tone were divided into six pairs. Table 3 displays these six pairs of words that we classified. Moreover, Figure 7 presents a comparison of the classification accuracies for these six pairs in both the closed and open testing.

Table 3. Six pairs of similar Myanmar words

Pair no.	Class no.	Myanmar consonants / IPA format
1	7, 8, 11	ကြက်သား, ဝက်သား,ချက်ထား / tɕɛʔθá, wɛ́θá, tɕʰɛʔtʰá
2	9, 10, 12	ဆိတ်သား, ဘဲသား, အမဲသား / sʰ̀eiˀθá, b̀θá, am̀θá
3	16, 17, 18	တိုးတိုး, ထိုးထိုး, ဒိုးဒိုး / tótó, tʰótʰó, dódó
4	19, 20, 21	မီးခိုး, တီးတိုး, ထီထိုး / míkʰó, tító, tʰìtʰó
5	22, 23, 24	ဘီပိုး, စီပိုး, ဆီထိုး / bìpó, sìpó, sʰìtʰó
6	32, 34, 36	မမ, ဘဘ, ပပ / ma̰ma̰, ba̰ba̰, pa̰pa̰

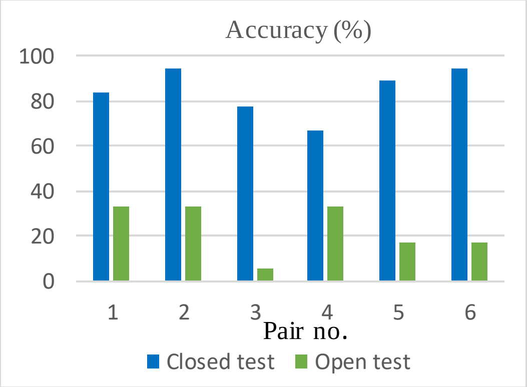


Figure 7. Comparison of word-level classification accuracy for six pairs in closed and open tests

Moreover, in order to examine the performance of our model, these six pairs were represented as confusion matrices, as illustrated in Figure 8. Based on the confusion matrix for each pair result, in the open test, the first (ကြက်သား, ဝက်သား, ချက်ထား) and second (ဆိတ်သား, ဘဲသား, အမဲသား) pairs were the most correctly classified pairs. The first pair was recognised correctly three times; twice and once as each word class, respectively, and classified incorrectly as other classes such as လိမ္မော်သီး, စပျစ်သီး, ဆိတ်သား, and within the pair. In pair 2, the numbers of times that our model classified each class correctly were 1, 4, and 1, respectively, and it classified these incorrectly as other classes within the pair and as စပျစ်သီး, ကြက်သား, and လိမ္မော်သီး.

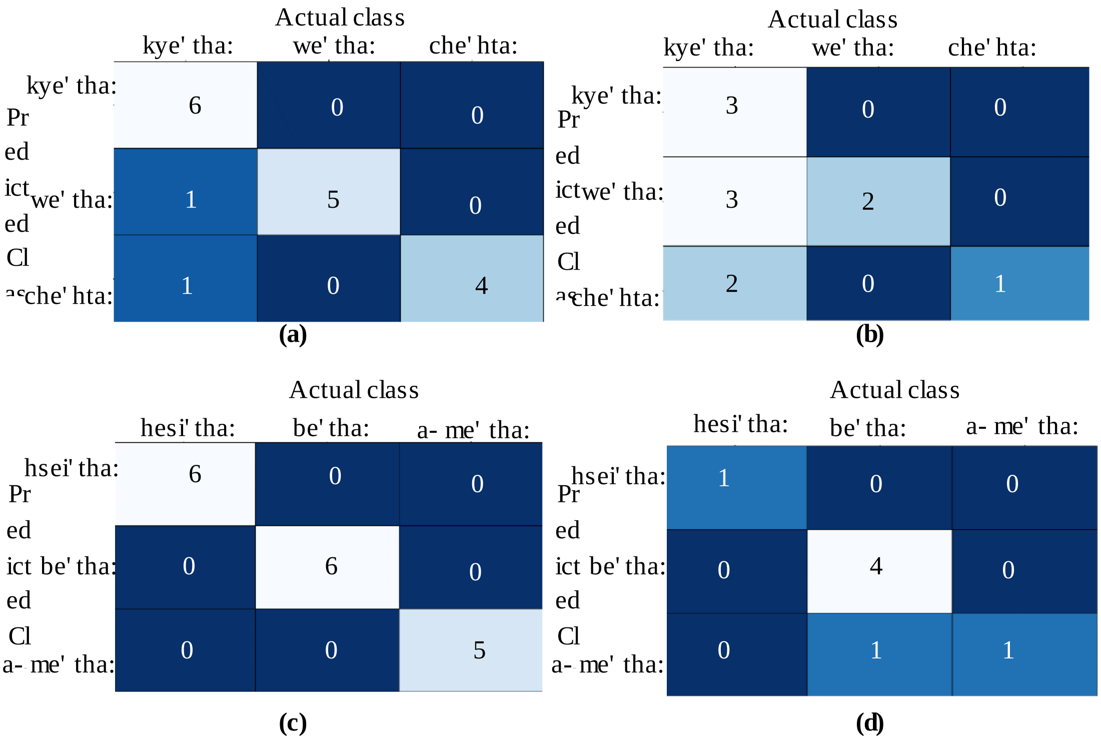
However, in the closed testing, the second pair was the best recognised, with 6, 6, and 5 times, respectively, and it was classified incorrectly as စိုးမိုး once when classifying အမဲသား. Furthermore, pair 1 was the third-highest classified pair, with 6, 5, and 4 times, respectively. Our model recognised incorrectly as another class (ဆိတ်သား), not included in this pair, once, while recognising ချက်ထား.

6. Conclusion

In this paper, the classification of Myanmar speech using transfer learning for image classification has been presented, achieving validation accuracies of 60.70%, 73.20%, and 94.60% for the consonant, vowel, and word-level classifications, respectively.

In the closed testing, we achieved classification accuracies of 38.18%, 61.67%, and 81.80% with 34.55%, 60%, and 31.17% in the open testing for the consonant, vowel, and word sound classes, respectively.

Based on the experiment, the method of audio classification with image classification is relevant to the classification of Myanmar speech, including consonants, vowels, and words. The results demonstrated that transfer learning can achieve classification when the number of classes is not high.



(a) (b) (c) (d)

Figure 8. (a) Confusion matrix for kye' tha:, we' tha:, and che' hta: pair in closed test; (b) confusion matrix for kye' tha:, we' tha:, and che' hta: pair in open test; (c) confusion matrix for hsei' tha:, be' tha:, and a- me' tha: pair in closed test; and (d) confusion matrix for hsei' tha:, be' tha:, and a- me' tha: pair in open test

In the future, we propose conducting further experiments using transfer learning with other freely available models, and comparing the results. Moreover, in order to use this approach in practical applications, it is intended to conduct a further study concerned with the retraining part of the transfer learning.

7. Acknowledgement

The authors gratefully acknowledge the teachers and students from the University of Computer Studies, Banmaw, who participated in recording the sounds for the Myanmar consonants and vowels, and other speakers from Kumamoto University, who aided in recording the words.

References

[1] 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.

[2] Ingyin Khaing, "Myanmar Continuous Speech Recognition System Based on DTW and HMM", Department of Information and Technology, University of Technology (Yatanarpon Cyber City, near Pyin Oo Lwin, Myanmar, International Journal of Innovations in Engineering and Technology (IGIET), Vol. 2 Issue 1 February 2013

[3] 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.

[4] 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.

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

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

[7] An open source machine learning framework: https://www.tensorflow.org/

[8] Inception-v3 performance: https://medium.com/@ williamkoehrsen/facial-recognition-using googlgesconvolutional-neural-network-5aa752b4240e

[9] 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

[10] Sox cross-platform command line: https://sox. sourceforge.net

[11] AudioNet open-source speaker-recognition: https://github.com/vishnu-ks/AudioNet/