I.J. Image, Graphics and Signal Processing, **2021, 2, 44-52**
Published Online April 2021 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2021.02.04

Myanmar Continuous Speech Recognition System Using Convolutional Neural Network

Yin Win Chit
University of Computer Studies (Lashio)
Email: yinwin.chit@gmail.com

Win Ei Hlaing
University of Technology (Yatanarpon Cyber City)
Email: wineihlaing@utycc.edu.mm

Myo Myo Khaing
University of Computer Studies (Lashio)
Email: myomyokhaing973@gmail.com

Received: 21 August 2020; Revised: 14 October 2020; Accepted: 27 November 2020; Published: 08 April 2021

Abstract: Translating the human speech signal into the text words is also known as Automatic Speech Recognition System (ASR) that is still many challenges in the processes of continuous speech recognition. Recognition System for Continuous speech develops with the four processes: segmentation, extraction the feature, classification and then recognition. Nowadays, because of the various changes of weather condition, the weather news becomes very important part for everybody. Mostly, the deaf people can't hear weather news when the weather news is broadcast by using radio and television channel but the deaf people also need to know about that news report. This system designed to classify and recognize the weather news words as the Myanmar texts on the sounds of Myanmar weather news reporting. In this system, two types of input features are used based on Mel Frequency Cepstral Coefficient (MFCC) feature extraction method such MFCC features and MFCC features images. Then these two types of features are trained to build the acoustic model and are classified these features using the Convolutional Neural Network (CNN) classifiers. As the experimental result, The Word Error Rate (WER) of this entire system is 18.75% on the MFCC features and 11.2% on the MFCC features images.

Index Terms: Automatic Speech Recognition, Convolutional Neural Network, Mel Frequency Cepstral Coefficient, Continuous Speech, Speech Segmentation.

1. Introduction

Automatic Speech Recognition system is one popular generation technology for interaction of computer and human. The ASR for the continuous speech is a very difficult task but it can help for the deaf people to get the easier their lives. In many research areas, ASR system are initially tested on TIMIT phone words training dataset with mono-phone HMM with MFCC features and then these are tested on the several huge amount of vocabulary speech recognition. ASR can be defined as the independent computer driven translation of spoken language into the readable text words. The main direction of the feature extraction is to calculate the feature vector of sequence providing a compact representation of the given input speech signal.

Over the last twenty years, the advantages of Neural Network based acoustic modelling for SR system has been performed as the main role because of the feasible Neural Network with many hidden units of speech data [1]. Before the convolutional neural network was used over the windows of the acoustic frames to achieve stable features for classes such as gender phone and speakers, CNNs have been used to create the acoustic model in the speech recognition system [2]. Deep architecture of Neural Network with the matrix enable to handle a SR system with various types of speech signal. The SR system with CNN classifier has reduced the WER value nearly 10 % than the WER values of DNN based SR system on the TIMIT dataset. To get the better performance of ASR system, this system proposed the SR with Convolutional Neural Network based acoustic model.

The most important part of the continuous speech recognition system for Myanmar Language is the segmentation in continuous speech. The major difficulty in the research area of Myanmar Speech Recognition System is the lack of Myanmar Speech Corpus. Generally, it is not easy to build the speech corpus because it requires a huge amount of

speech data and it is very difficult for correct segmentation of Myanmar Continuous Speech [3]. In our country, there are many people who cannot hear the sound deaf but they can read the text of mother language. Everybody wants to know about the important news of his/her country such as national news, weather news. Therefore, this system presents automatic continuous speech recognition for continuous speech in Myanmar Weather News.

2. Related Works

Many research on the previous literature have been implemented the developing system for speech recognition with the various models and various types of speech.

In 2002, H.Lukman and Thiang developed a SR system with the fuzzy logic matching techniques which was used on PC. In the step of feature extraction step, Fast Fourier Transform (FFT) is used to extract the feature from the speech signal. The main step of SR, recognition step in which Hidden Markov Model (HMM) is used to recognize with multiple mixtures and models, and then the recognition results nearly 97.1% have been achieved [4].

S. Kavita, et al. presented speech recognition is a broader solution which refers to a technology that can recognize a speech without being targeted at single speaker as the call system that can recognize arbitrary voice in 2012. The fundamental purpose of speech is communication, i.e., the transmission of messages. The problem in speech recognition is the speech pattern variability. The most challenging sources of variations in speech are speaker characteristics including accent, co-articulation and background noise [5].

In 2014, A. Stolcke and et al. implemented the techniques to achieve the higher accuracy of the ASR system with the phonetic segmentation based on the acoustic HMM phonetic model. In this system, the testing results are improved by using more useful and powerful statistical models for boundary correction. As their experimental result of system, segmentation error can be reduced based on the TIMIT corpus dataset [6].

In 2016, E. Chandra and S. Karpagavalli created the voice model with the most natural mode of the individual personal communication. The task of SR system is to translate the speech into a string of the words by a computer program. SR model supports the people with the speech or voice an alternative mode of interaction of machine or mobile devices [7].

According to the previous literatures, Speech Recognition system have many challenges to get the better accuracy for the various languages. To improve the accuracy of the recognition system, the high performance classifiers or recognizers are needed in the recognition of signal features. In this system, the Convolutional Neural Network classifier is used to get the high performance of the recognition system.

3. Proposed Methodology

In the previous literature, CNN classifier was described as the higher performance classifier in sound recognition system and other emotional recognition systems. But CNN classifier was mostly used for image recognition system because it can produce the higher accuracy. And then features converted images recognition systems are improved in many sound recognition system to get the optimal solution. So, MFCC signal features are used as two types as signal features type and features converted images types to analyze the performance of the speech recognition system for Myanmar Language Continuous Speech.

In the past, many speech recognition systems are proposed for various languages. Other languages have separated rules for continuous speech but Myanmar Language has many challenges for the continuous speech. In the previous Myanmar continuous speech recognition system created in the speech segmentation using one or two type features types. In this system, two time domain features and one frequency feature are applied to get the more exact speech segments.

The proposed system design and methods of this system are presented in this section. Dynamic thresholding method is used to segment the continuous speech of daily weather news. Features of speech are extracted from each segment by using MFCC feature extraction method and then these extracted features are converted into the feature images. These extracted features and feature images are trained to crate the speech dataset of Myanmar weather news and the related words of input speech segment are classified by using CNN classifier. This system is used to recognize the weather news for Myanmar deaf people. The methods used in this system, segmentation, feature extraction and recognition are described the step-by-step processes in detail.

A. Proposed System Design

The training step and the testing step of the proposed system design are shown in the Fig. 1. The main parts of the system are described step by step in this section. The continuous speech recognition system is implemented in windows environment and Matlab Tool Kit is used for developing this system.

The proposed continuous speech recognition system has three major steps as follows:

- Preprocessing and Word Level Segmentation
- Feature Extraction with MFCC

• Recognition with CNN

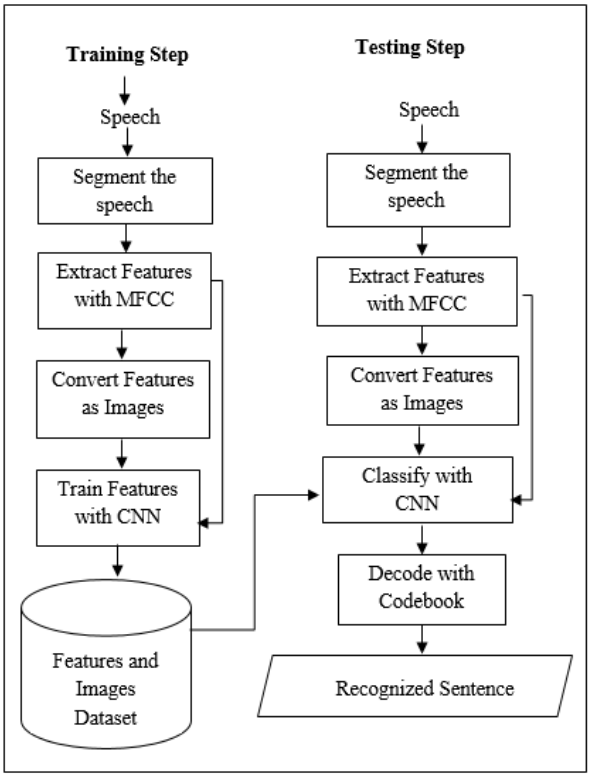

Fig. 1. Proposed System Design

The most important part of the continuous speech recognition system for Myanmar Language is the segmentation in continuous speech. The dynamic thresholding method using time and frequency domain features was effectively developed for the large vocabulary recognizers.

In the feature extraction step, two types of MFCC features are extracted. The first type is MFCC features from the segmented speech signal. As the seconds type, MFCC features with the various frame size are extracted from the each segmented speech signal and then these features are converted into the type of image (32*32*3, ".png" format). To get the better ASR, this proposed system presents the speech recognition system using CNN classifier.

B. Speech Acquisition

In this system, Myanmar Continuous Speech sentences from the video files are acquired to get the continuous speech. The Daily Myanmar Weather News video files are collected from the Department of Meteorology and Hydrology, Nay Pyi Taw, Myanmar. The collected weather news video files are converted into audio wave file.

And then the whole news audio files are cut into the single sentence using Easy MP3 Cutter software. The continuous speech audio files are collected to use as the training data set and testing dataset. The continuous speech of these audio with the multiple female speakers are used as the training data of the system. This system translates the weather news report speech into the long sentences of Myanmar text.

C. Speech Preprocessing

This step includes elimination of background noise, framing and windowing. Background noise is removed from the data so that only speech samples are the input to the further processing. Continuous speech signal has been separated into a number of segments called frames, also known as framing.

After the pre-emphasis, filtered samples have been converted into frames, having frame size of 40 ms. Each frame overlaps by half. Windowing is done to reduce the edge effect of each frame segment.

D. Segmentation

In popular research areas of ASR, only time domain features or the addition of one time domain feature and one frequency feature are used for the segmentation step. In this system, two time-domain signal features and one frequency-domain signal feature are extracted to define the threshold value of the dynamic threshold segmentation method. In segmentation step of this system, Short-Time Energy feature and Zero-Crossing Rate features are used as time domain signal features and Spectral Centroid features are used as the frequency domain feature.

After computing speech feature sequences, a simple dynamic threshold-based algorithm is applied in order to detect the speech word segments.

- Compute the Mean values of smoothed feature sequences.
- Find the local maxima of histogram.
- If at least two maxima M1 and M2 have been found
- Threshold value is calculated as follows:

Threshold,

$$T = \frac{W \times M_1 + M_2}{W + 1} \quad (1)$$

Otherwise,

$$T = \frac{Mean}{2} \quad (2)$$

Where W is a user-defined weight parameter, here, W=10. The above process is applied for both feature sequences and finding two thresholds: T1 based on the energy sequences, T2 based on the zero crossing rates and T3 based on the spectral centroid sequences. After computing two thresholds, the speech word segments are formed by successive frames for which the respective feature values are larger than the computed thresholds T1 and T3 but smaller than the T2.

The threshold value based segmentation is more effective and exact to detect the unvoiced sound. The detailed explanations of these features are described in following section. After windowing, compute the short-time energy features, zero-crossing rate and spectral centroid features of each frame of the speech signal. After computing speech feature sequences, a simple dynamic threshold-based algorithm is applied in order to detect the speech word segments. In this step, the small unit segments are segmented from the long sentences continuous speech.

The sample speech signal of Myanmar Weather News sentences is shown in Figure 2.

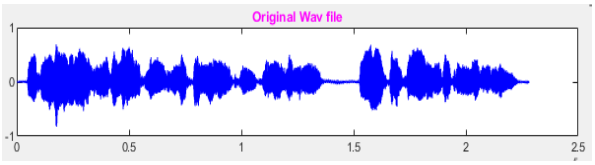

Fig.2. Continuous Speech Signal of Long Sentence Myanmar Weather News

Finally, segments of long speech are achieved using dynamic threshold method based on these extracted three features. In Figure 3, the segments of sentence are shown with green color for each segment. After segmentation, nine segments are divided from this sample sentence.

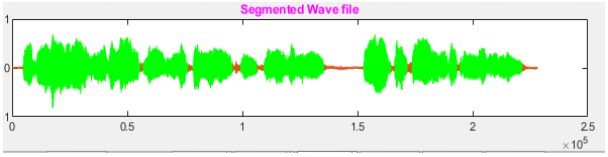

Fig.3. Segment Speech of Sentence

E. Feature Extraction

In the feature extraction step, two MFCC features are extracted to build acoustic model from each segmented speech. The sample MFCC features are extracted from the segmented speech signal and then acoustic model is created using these extracted features. This system calculated the MFCC feature values (20 points coefficient) and then these values are resized to get the (32*32 – 2D matrix). And then, MFCC coefficient values (mel frequency cepstral coefficient values) are extracted from three different frame sizes to create the second type of feature images. Then features from each frame size are assigned into the each layer of the images and then created the three layer image (32*32*3) format for each segmented word. The converted images are trained to construct the acoustic model of speech recognition system by using training step of Convolutional Neural Network classifier. The 22 segments and 49 Myanmar words of the sample sentence are described in Fig. 4. After segmentation and feature extraction step, the MFCC features (as images) are collected to train as the acoustic model in the training and these are tested in the recognition step. The images are used to train with convolutional neural network.

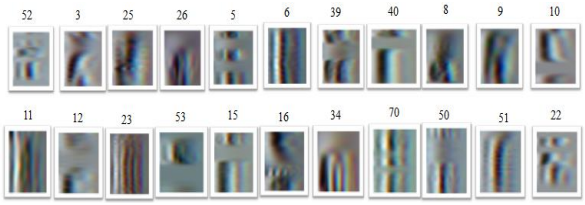

Fig. 4. Segment ID-1 from Three Frames Size for Myanmar Sentence

F. Recognition

In using the CNN for speech recognition, the input data need to be organized as a number of feature maps to be fed into the CNN. This is a term borrowed from image-processing applications, in which it is intuitive to organize the input as a two-dimensional (2-D) array, being the pixel values at the (horizontal and vertical) coordinate indices. For color images, RGB (red, green, blue) values can be viewed as three different 2-D feature maps. The input images 32*32*3 are used to train with Convolutional Neural Network. CNNs run a small window over the input image at both training and testing time, so that the weights of the network that looks through this window can learn from various features of the input data regardless of their absolute position within the input. In this system, the following 15 layers are used to train and to build as the acoustic model:

- Image Input Layer
- Convolution2DLayer
- MaxPooling2DLayer
- ReLULayer
- Convolution2DLayer
- ReLULayer
- AveragePooling2DLayer
- Convolution2DLayer
- ReLULayer
- AveragePooling2DLayer
- FullyConnectedLayer
- ReLULayer
- FullyConnectedLayer
- SoftmaxLayer
- ClassificationOutputLayer

In the training step, each segmented speech is trained with CNN to identify the class ID number. In the testing step, word segment ID numbers are identified by using CNN. Then the Myanmar words are recognized the equivalent identified ID number. Finally the Weather News sentences are recognized as Myanmar words.

4. Experimental Resuls

In this section, the experimental results of the system are discussed and the implementation steps of this work are also described. And then the various analyses are made on the different types of dataset and then the results of these analyses are expressed.

A. Data Collection

In this system, dataset (Acoustic Model) is created by using daily weather news reports of 6 years period (2012-2017). These daily weather news reports videos are collected from the Department of Meteorology and Hydrology, Nay Pyi Taw, Myanmar. Then these video files are converted to the audio files using format converter. The speech audio files for each sentence of weather news are spilt from the converted audio files by using Easy MP3 Cutter software. The split sentences of daily weather news audio files are used as the training dataset and testing dataset for this system.

The collected data of the system is described in Tables 1. The two acoustic models with these two feature types of segmented speech are built using DCNN classifier. The number of female speakers and time for training dataset are described in this table.

Table 1 Training Speech Dataset of Female Speakers

Number of news report sentences	8000-sentences
The number of trained MFCC various Frame Images	134,545- words
The number of trained mat file for MFCC feature values	134,545- words
Number of Female Speakers	32-speakers
Time taken for training dataset	56,522-seconds

The training speech dataset of the system with the total news report speech sentences of male speaker is shown in Table 2. And then the total number of male speech segments, number of male speakers and total seconds of training dataset are described in this table.

Table 2. Training Speech Dataset of Male Speakers

Number of news report sentences	1600-sentences
The number of trained MFCC various Frame Images	11,257-words
The number of trained mat file for MFCC matrix values	11,257-words
Number of speakers (Male)	8-speakers
Time taken for training dataset	8056-seconds

The number of words in each sentence is between 25 and 58 Myanmar words (words in sentence), and the number of segment is 7 to 31 segments (segmented word). All speech signal of this system are digitized at the sample rate of 44100 Hz using 16 bits (.wav file) format. In real time testing, any speaker speech can be recognized. The procedure for creating the dataset for this speech recognition system is shown in Fig. 5.

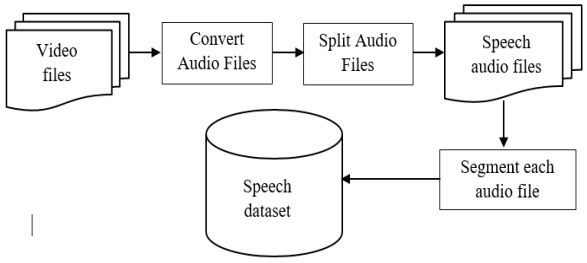

Fig. 5. Procedure for Creating Dataset

B. Performance Analysis

In this system two different types of features are used and the experimental results are calculated on these two features. Experimental result-1 shows the result on the traditional MFCC features with deep convolutional neural network and the experimental result-2 shows the result of proposed MFCC features images type. In this experiment, spoken weather news report sentences are recognized by the system. The system output was recognized each sentence of daily weather news with Myanmar Language words. Many speech recognition systems used WER to measure the performance of the system accuracy. The experimental result of this system can be calculated using WER to analysis the performance of the system and it can be computed as following equation 3:

$$WER = \frac{Substitution+Delection+Insertion}{Total\ number\ of\ words\ in\ the\ reference} \times 100\% \quad (3)$$

Table 3 shows the word error rate and number of segments for each sentence. In addition, the number of correct segments and number of missing segments are shown in this table. These sentences are selected from the open dataset and then these are tested using two types of feature in recognition step.

Table 3. Word Error Rate for Sample Sentences

Sentence ID	Features	No. of total words	No.of correct words	No.of missing words	Word Error Rate (%)
SID-1	MFCC Feature Image	22	22	0	0
	MFCC Features	22	21	1	4.55
SID-2	MFCC Feature Image	15	14	1	6.67
	MFCC Features	15	13	2	13.33
SID-3	MFCC Feature Image	21	20	1	4.76
	MFCC Features	21	18	3	14.29
SID-4	MFCC Feature Image	18	16	2	11.11
	MFCC Features	18	15	3	16.67
SID-5	MFCC Feature Image	15	15	0	0
	MFCC Features	15	14	1	6.67
SID-6	MFCC Feature Image	10	9	1	10
	MFCC Features	10	9	1	10
SID-7	MFCC Feature Image	14	14	0	0
	MFCC Features	14	13	1	7.14
SID-8	MFCC Feature Image	14	14	0	0
	MFCC Features	14	13	1	7.14
SID-9	MFCC Feature Image	14	14	0	0
	MFCC Features	14	14	0	0
SID-10	MFCC Feature Image	11	11	0	0
	MFCC Features	11	10	1	9.09

Table 4 shows the comparison result of MFCC feature image and MFCC feature on the female speech dataset. This table shows the average comparison results on 100 sentences for the whole female testing dataset according to this analysis. The analysis is made on the system training dataset of the weather news Myanmar. These open data are collected from the Myanmar daily weather news with female speakers.

Table 4. Analysis of WER on Two Features of Female Speech Dataset (average rate on 100 sentences)

Type of features	1-100 sentences	101-200 sentences	201-300 sentences	301-400 sentences
MFCC Images	10.8	11.5	12.2	11.9
MFCC Features	13.2	14.4	16.7	15.5

In Fig. 6, the blue color bar represents the WER value of recognition result on the MFCC feature images and the red color bar represents for MFCC features recognition results. According to the analysis, the word error rate of female testing data is nearly 11 % on the MFCC feature images and nearly15 % on the MFCC features.

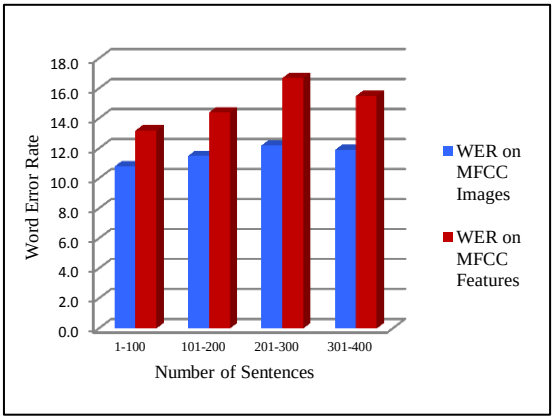

Fig. 6. WER of Female Testing Data on Whole Dataset

And the average comparison results on 100 sentences of male speaker for the total male speech testing dataset are shown in Table 3 and Figure 7. In this table, the analysis is made on the open male data on the MFCC feature images and MFCC features and then the word error rate of male speech data of the weather news are shown in Table 5.

Table 5. Analysis of WER on Two Features of Male Speech Dataset (average rate on 100 sentences)

Types of Features	1-100 sentences	101-200 sentences	201-300 sentences	301-400 sentences
MFCC Images	15.2	14.6	15.1	14.4
MFCC Features	17.7	16.9	18.2	17.1

In Fig. 7, the analysis result of Table 3 are shown with bar chart. The analysis can be seen that the WER on MFCC feature image is better than the WER on the MFCC features corresponding to the experimental results. The WER of male data is slightly increased than the female data. The word rate on the MFCC feature images is nearly 15 % and 17% on the MFCC features with male speakers.

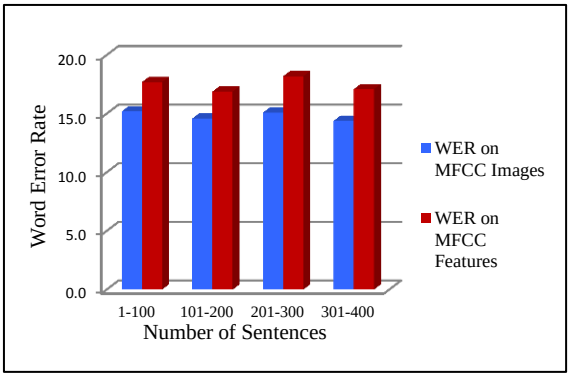

Fig. 7. WER of Male Testing Data on Whole Dataset

The accuracy results of weather news sentences that include three states are shown in Table 6. Three states are region, condition and temperature of the weather news. The accuracy of weather news sentences is analyzed on CNN classifier. The accuracy of temperature words is lowest because the dataset of weather news words has large amounts of temperature words.

Table 6. Accuracy of Weather News Sentences with Three States

Features	Accuracy of Region words	Accuracy of Condition words	Accuracy of Temperature words
MFCC features	87.5%	85%	83.5%
MFCC Images	90%	88%	86%

5. Conclusions

The performance of continuous speech recognition system for Myanmar Daily Weather News report is presented in this research. The acoustic features are extracted using MFCC features in two types and then the comparison results are analyzed on these features. The WER of open dataset from the weather news is 11.2% on feature images dataset and 18.75% on the MFCC features dataset. According to the experimental results, it can be seen that the use of MFCC feature images achieves the smaller error rate than traditional MFCC features. The extracted features from the various frame sizes get more effective features and these features support the recognition step to reduce the Word Error Rate. The performance analysis of the system is depended on the accuracy of the recognized segmented words. In this system, 415 segmented words are collected and the Word Error Rate is calculated on these segmented words. Therefore, the speech recognition system can provide the better accuracy using the proposed architecture and the created Myanmar speech dataset that can also be easily updated for daily weather news. The accuracy results of the system based on feature images recognition is better than the other Myanmar Continuous speech recognition system based on signal features.

References

[1] G. Hinton, “Training products of experts by minimizing contrastive divergence,” Neural Comput., vol. 14, pp. 1771–1800, 2002.
[2] D. Hau and K. Chen, “Exploring hierarchical speech representations using a deep convolutional neural network”. 11th UK. (UKCI ‟11), Manchester, U.K., 2011.
[3] İ. G. Khaing, K. Z. Linn, “Myanmar Continuous Speech Recognition System based on DTW and HMM”, IJIET., Vol.2, Issue 1, February, 2013.
[4] H. Lukman and Thiang, “Limited Word Recognition Using Fuzzy Matching ". ICOLA. Jakarta, 2002.
[5] H. Singh and A. K. Bathla "A Survey on Speech Recognition ". 9th ICMT, 2005.
[6] A. Stolcke, and et al., “Highly accurate phonetic segmentation using boundary correction models and system fusion,” ICASSP, 2014, 5552–5556.
[7] S. Karpagavalli and E. Chandra, “A Review on Automatic Speech Recognition Architecture and Approaches”. IJSP, Image Processing and Pattern Recognition, 9(4), 2016, 393-404.

Author’s Profile

Dr. Yin Win Chit, Assistant Lecturer**,** She is the assistance lecturer of University of Computer Studies (Lashio). She got the Doctor of Philosophy Degree from the University of Technology (Yatanarpon Cyber City) in 2019. Now, she have been developed many research area related work of speech.

How to cite this paper: Yin Win Chit, Win Ei Hlaing, Myo Myo Khaing, " Myanmar Continuous Speech Recognition System Using Convolutional Neural Network ", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.13, No.2, pp. 44-52, 2021.DOI: 10.5815/ijigsp.2021.02.04