Continuous Speech Recognition System Based on Deep Convolutional Neural Network for Myanmar

Yin Win Chit[#1], Soe Soe Khaing[*2], Yi Yi Myint[#1]

[#1]Faculty of Information and Communication Technology

University of Technology (Yatanarpon Cyber City)

Pyin Oo Lwin

Republic of the Union of Myanmar,

[*2]University of Computer Studies (Monwya)

Republic of the Union of Myanmar,

yinwin.chit@gmail.com

khaingss@gmail.com

yiyimyint.utycc@gmail.com

Abstract **- Automatic Speech Recognition (ASR) system, that translates the speech signal into text words, is still a challenge in the continuous speech signal. Continuous speech recognition systems develop with four separated steps: segmentation of the speech signal, feature extraction, classification and recognizing the words. These steps can be modeled with the various methods. Among them, the combination model of the dynamic threshold based segmentation, Mel-Frequency Cepstral Coefficient (MFCC) feature extraction method and Deep Convolutional Neural Network (DCNN) is proposed in this paper. Especially, DCNN-AlexNet has been applied in image processing because it can perform as a highly accurate, effective and powerful classifier. In the training and classification step of this system, the advantages of DCNN in image processing are applied using the MFCC feature images. The main purpose of this system is to transform the MFCC features of the speech signal to MFCC features images with various frame size for three layers of input images of DCNN. The three layers 32*32*3 images are used for the input images of DCNN-AlexNet to support the recognition step of the system. The experiments shows that the DCNN speech recognition system achieves the average Word Error Rate (WER) of 11.5 % on the proposed MFCC images training dataset and WER of 13.75% on the MFCC features value matrix training dataset.**

Keywords **– Automatic Speech Recognition, Mel-Frequency Cepstral Coefficient, Deep Convolutional Neural Network, Word Error Rate**

I. Introduction

Automatic Speech Recognition system is one generation technology for human and computer interaction. This technology is a very difficult task but it can support to get the easier life of deaf people. Many researches of ASR initially are tested on TIMIT phone training data set with mono-phone HMM for MFCC features [1] and then are tested on several large vocabulary speech recognition with tri-phone HMM model [3]. Over the last twenty years, the advance of ANN based approaches to acoustic modelling for speech recognition has been played as a significant role because of feasible training ANNs with many hidden units on many hours of speech data [4]. Before the convolution was applied over windows of acoustic frames to provide more stable acoustic features for classes such as phone, speaker, and gender, CNNs have been applied to the acoustic modelling for speech recognition [5]. Deep architectures with merit enable a speech recognition model to handle many types of variability in the speech signal [6].

The speech recognition applied CNN has reduced the nearly WER of 10% than the experimental result of the DNN-based speech recognition on the TIMIT dataset [6]. The reasonably good model for the effective speech feature extraction can be performed by the DCNN pertained model for the image application [7]. To get the better ASR, this paper presents the speech recognition system using deep Convolutional Neural Network.

This paper presents the one speech recognition model to get the better continuous speech recognition system for Myanmar. The most important part of the continuous speech recognition system for Myanmar Language is the segmentation in continuous speech. According to the experiment on continuous speech segmentation techniques, the dynamic thresholding method using time and frequency domain features was effectively developed for the large vocabulary recognizers. In the feature extraction step, 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).

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 [2]. 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.

This paper is organized with the remaining parts as follows: section II discusses about the proposed model with system design of training and testing step of the system and each of step-by-step process of the system. Then section III shows the experimental results of the system as WER of the two extracted MFCC features images classification, and the section IV describes the discussion. The last one describes the conclusion of this paper.

II. PROPOSED MODEL

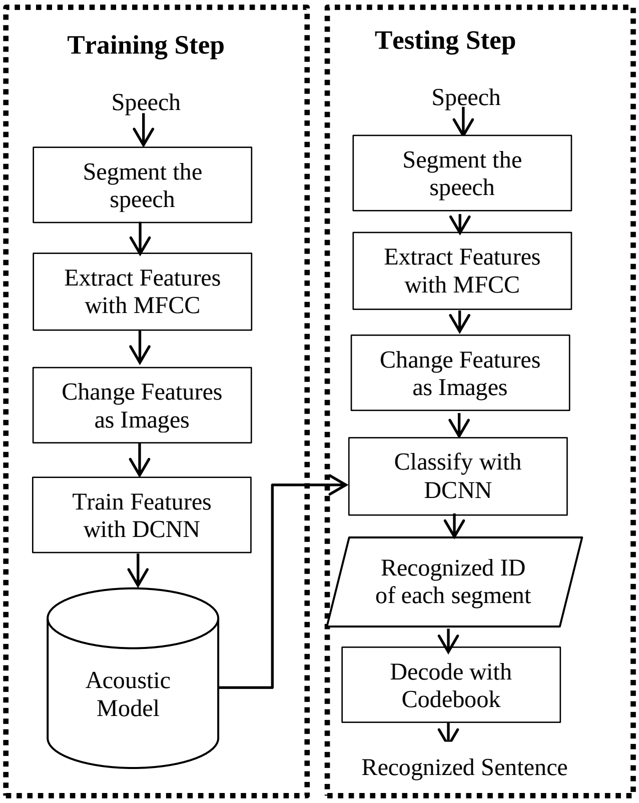

Fig. 1 The proposed system design

The training step (left side) and the testing step (right side) of the proposed system design are shown in the figure 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 five major steps as follows:

- ➢ Speech Acquisition and Pre-processing
- ➢ Speech Segmentation
- ➢ Feature Extraction with MFCC
- ➢ Recognition with DCNN

A. Speech Acquisitions and Pre-processing

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 and these video files are converted into the audio file using audio converter. The continuous speeches 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.

B. Speech Segmentation

Speech segmentation is a process of decomposing the speech signal into smaller units. It is used to detect the proper start and end point of segment boundaries of continuous speech by using dynamic thresholding method. Time-domain signal features are mostly used in many research area of speech recognition system for speech segment extraction. These features are useful when the system needs to have an algorithm that is easy to implement and efficient [8].

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 thresholding 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. The threshold value based segmentation is more effective and exact to detect the unvoiced sound. The detail explanations of these features are described following in this section.

1) *Short-Time Energy (STE)*: Short-time energy is the most important natural feature in speech segmentation. The following equation can calculate the short-time energy:

$$E_n = \sum_{m = n - N + 1}^{n} (x\,[m])^{2} \qquad \text{---- (1)}$$

where N is the length of frame and x (m) is the discrete-time audio signal.

2) *Zero-Crossing Rate (ZCR)*: The average zero-crossing rate is the rate of times speech samples changes algebraic sign in a given frame. Unvoiced speech components usually have ZCR values much higher than voiced speech components [9]. The zero-crossing rate can be calculated by using the following equation:

$$Zn = 1/2 \sum_{m=1}^{N} \left|\text{sgn}[x\,(m)] - \text{sgn}[x\,(m\text{-}1)]\right| \, w(n\text{-}m) \qquad \text{---- (2)}$$

where x(m) is the discrete time audio signal and w(n) is rectangle window function.

3) *Spectral Centroid:* Spectral Centroid feature is the dimension of the spectrum position with a high value equivalent to a brighter sound [10]. The spectral centroid , SC_i , of the ith frame is defined as the centre of “gravity” of its spectrum and it is given by the following equation:

$$SC_i = \frac{\sum_{m=0}^{N-1} f\,(m)\, X_i(m)}{\sum_{m=0}^{N-1} X_i\,(m)} \qquad \text{--- (3)}$$

Here, f (m) represents the centre frequency of i^{th} bin with length N and $X_i\,(m)$ is the amplitude corresponding to that bin in DFT spectrum. After calculating speech feature sequences, a simple dynamic threshold-based algorithm is applied in order to detect the speech word segments.

C. MFCC Feature Extraction

Nowadays, Automatic Speech Recognition system is based on several types of Mel-Frequency Cepstral Coefficient (MFCC) that are proved to be effective and robust in a variety of situation. Speech signals are normally pre-processed before features are extracted to improve the accuracy and efficiency of the extraction process. The MFCC is based on the known variation in the critical bandwidth frequencies of human ear with the filter spaced linearly at low frequency, logarithmically at high frequency, which is used to capture the important characteristics of speech [11]. The following block diagram illustrates the process for obtaining the MFCC features.

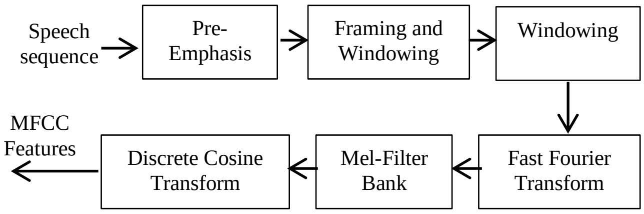

Fig. 2 The block diagram of MFCC feature extraction method

The digital speech waveform generally has a high dynamic range and extremely sensitive noise. The first step of MFCC, pre-emphasis is applied to reduce this range. Pre-emphasis used a first-order FIR high-pass filter to boost the amount of energy in the high frequency [13]. More information are obtained from these higher formants available to the acoustic model by boosting the high frequency energy.

The framing step of MFCC is used for analysing each frame in short time instead of analysing the entire signal at once and then overlapping is applied to frames. Hamming Window is also applied to get rid of some of the information at the beginning and end of each frame. Overlapping reincorporates this information back into the extracted features [13].

In most speech recognition, Hamming window is commonly used as window shape to avoid unnatural discontinuities in the speech segment and distortion in the underlying spectrum [14]. The Fast Fourier Transform is used to convert the convolution of the glottal pulse and the vocal tract impulse response in the time domain.

The bandwidths and spacing of the filter-bank with Band-pass filter are roughly equal to the critical and range of the centre frequencies covers the most important frequencies for speech perception. In the final step of MFCC, Discrete Cosine Transform (DCT) is used to convert the log Mel spectrum into the time domain. The result of this conversion is called Mel Frequency Cepstrum Coefficient.

In the feature extraction step, this system calculated the MFCC feature values (32*32) 2D matrix. And the extracted MFCC coefficient values from the various frame sizes are converted into the image (32*32*3) format. The converted images are used to construct the acoustic model of speech recognition system by using training step of DCNN classifier.

D. Recognition with Deep Convolutional Neural Network

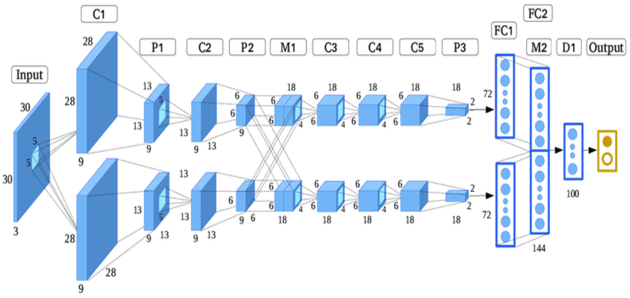

Fig. 3 Architecture of DCNN-AlexNet

Deep Neural Network (DNN) acoustic model has been used to improve the performance of Automatic Speech Recognition system. Among various Deep Learning methods, Convolutional Neural Network (CNN) was mostly and successfully used in image classification. CNN with three key properties (locality, weight sharing and pooling) has the potential to improve the performance of speech recognition. In addition, since Convolutional Neural Network can utilize local information on time and frequency in CNN, it has many advantages in ASR building acoustic model. Deep structure based CNN acoustic models significantly reduced the Word Error Rate (WER) of the Speech Recognition system [13]. The DCNN-AlexNet architecture network is described in figure3 that contains eight learned layers as five convolutional and three fully connected layers.

DCNN has many advantages in the classification system. The recent research described the main advantage of the DCNN that has the high accuracy in the image recognition and speech emotional recognition system [12]. DCNN-AlexNet has been applied in image processing because it can perform as a highly accurate, effective and powerful classifier. The main step of the proposed system is the transforming from the MFCC features of the speech signal to MFCC features images with variable framing and then the system shows the comparison accuracy results on the MFCC images training dataset and MFCC values matrix training dataset. This proposed system used DCNN-AlexNet as the training classifier to support the recognition step of the system.

III. EXPERIMENTAL RESULTS

The proposed system is tested on the long sentences of Myanmar Daily Weather News report as case-study with large number of segmented words. Daily Weather News Reports of 4 years period (2014 to 2017) are used as the training and testing data set of the system. Every sentence of weather daily report has 18-27 segments and has a few differences due to the different speakers' speech rate. This system is implemented on the large Myanmar Language Vocabulary Acoustic Model with many segments of the weather news report to get the better accuracy. One of the samples long sentences is shown as follow:

"နေပြည်တော်နဲ့အနီး - တဝိုက် - အတွက် - မနက်ဖြန် - ညနေ - ထိ - ခန့်မှန်း - ချက်မှာ - နေရာ - ကွက်ကျား - မိုးထစ် - ချုန်းရွာ - နိုင်ပြီး - နေ့အပူ - ချိန်မှာ - သုံးဆယ့် - နှစ် - ဒီဂရီ - စင်တီ - ဂရိတ် - ရှိနိုင်ပါတယ်"

This sample sentence has 21 segments and 49 Myanmar words. The mean of this report sentence in English is "Forecast for Naypitaw and neighbouring area until tomorrow evening, the weather can be available scatter rain and thundershowers. The day temperature may be thirty-two degree Celsius."

The datasets of the system are described in the table 1. Each sentence speech of daily report can be segmented in the various forms of segments by using thresholding method. Then the MFCC feature extraction method is used to extract the features from each segmented speech. The simple MFCC features matrixes are used in many speech recognition systems. The proposed system converted the MFCC features matrixes (32*32 matrixes) into the images (32*32, '.png' file format). Then this image is assigned into the three layers images (32*32*3) of the DCNN as the input images of the AlexNet framework.

TABLE 1

COLLECTED DATASET OF THE SYSTEM

Number of news report sentences	5200
Total Number of speech segments	109,500
No. of speakers (female)	32

The amount of MFCC single frame features images and MFCC feature matrixes from the segmented speech are used as the training dataset of DCNN. The two acoustic models with these two feature types of segmented speech are built using DCNN classifier. The training dataset of the DCNN is shown in table 2 and testing dataset of the system is shown in the table 3.

TABLE 2

TRAINING DATASET FOR THE DCNN

The number of trained MFCC various Frame Images	109,500
The number of trained mat file for MFCC matrix values	109,500
The number of training report sentences	5200
Total time of training dataset	36,400 seconds

TABLE 3

TESTING DATASET OF THE SYSTEM

The number of tested MFCC various Frame Images	16,800
The number of tested mat file for MFCC matrix values	16,800
The number of testing report sentences	800
Total time of training dataset	5,600 seconds

The performance accuracy of the system can be described by using Word Error Rate (WER) based on the number of segments. Word Error rate can be computed using following equation:

$$\text{WER} = \frac{\text{No. of miss Recognized words}}{\text{Total No. of Segmented words}} \qquad \text{---- (4)}$$

The table 4 shows the first experimental results that are analysed on the WER of the simple MFCC feature matrix values and the single frame MFCC feature images.

TABLE 4

FIRST EXPERIMENTAL RESULTS

Type of image dataset	WER
MFCC feature value matrix	13.75%
MFCC feature images with single frame size	14.6 %

In the second testing of the system, MFCC features are extracted with the various frame sizes such (0.1%, 0.15% and 0.2% of the frame). The features of each frame size are converted into the image and then assigned into the three layers input of the DCNN (32*32*3, '.png' format). These images are trained as the acoustic model using DCNN-AlexNet classifier. The classification result of these images dataset gets the higher accuracy than the single frame and matrix value of training dataset on the same amount of trained data. Table 5 shows the second experimental results of the system. The experimental results of the system are described with bar chart as shown in figure 4.

TABLE 5

SECOND EXPERIMENTAL RESULTS

Type of image dataset	WER
MFCC feature value matrix	13.75%
MFCC feature images with **single** frame size	14.6 %
MFCC feature images with **various** frame size	11.5%

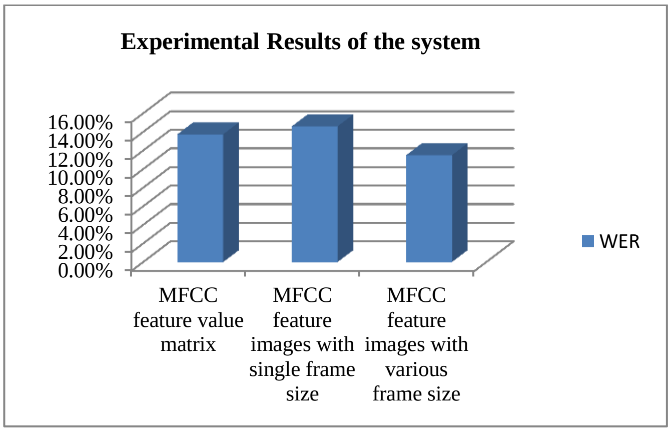

Fig. 4 Chart for experimental results of the system

According to the results of the analysis, the accuracy rate of MFCC feature images with various frame size is higher than another one. The features obtained from three types of frame sizes are put in the three layers of training images that can provide the better accuracy for the recognition step. ASR based DCNN with MFCC feature images training set can reduce the Word Error Rate than the DCNN with MFCC features value matrix training set.

IV. Discussion

In this system, the extracted MFCC features from the three types of frame size are used and then these features are converted into the images. Firstly the input speech signal are framed with various seizes such as 0.1%, 0.15% and 0.2% of frame size, then the features points of MFCC are extracted from each frame size. These feature extraction method with three 2D matrix values is more exact than the feature extraction method based on the single frame size with one 2D matrix value. The features from each types of frame size are assigned into each layer of image (32*32*3 format). The better performance of the system is achieved by using the features from the different frame and the converted features images. To verify the effectiveness of our proposed system, we compare it with commonly used MFCC feature values extraction and converted MFCC images, shown in table 5. The reason why the AlexNet helps speech recognition might be because we convert the audio signal into the image like representation as well as the strong features learning ability of the AlexNet. Besides, MFCC features from three types of frame size describe the segmented speech as certain shapes and structures, which are thus able to be effectively perceived by the AlexNet pre-trained on the image domain. The proposed method is based on the AlexNet similar to ImageNet large scale classification. It is capable of learning on million-scale training data, thus it is also interesting to retrain deep models on larger segmented speech data set. In our future work, the large vocabulary segmented speech data set based on DCNN- AlexNet for all Myanmar news are constructed and then these dataset are used to get better speech recognition system for Myanmar People.

V. Conclusions

This system presented the Continuous Speech Recognition System in which testing and training samples are extracted from Daily Weather New Report to get the better accuracy. Many advantages of DCNN in image processing are applied in this system and MFCC feature images of segmented speech are used to improve the performance of the speech recognition system. In this speech recognition system, dynamic thresholding method based segmentation, MFCC feature extraction method and DCNN classification method are implemented to create large amount of Myanmar sentences dataset. This system shows the experimental results of DCNN based ASR. Finally, this system provides the smaller WER rate of speech recognition because the advantages of DCNN of image processing are applied in the training step of ASR. In conclusion, we can say that DCNN can be suitable for Speech Recognition system according to the experimental results. The MFCC features should be used the various frame sizes to get the better Speech Recognition System.

Acknowledgments

In the preparation of this paper, I would like to special thanks to my supervisor Dr. Soe Soe Khaing, Rector (University of Computer Studies, Monwya) and my co-supervisor Dr. Yi Yi Myint for their fruitful advices. I especially thank the Rector, University of Technology (Yatanarpon Cyper City) and my colleagues in Myanmar.

References

[1] A. Mohamed, D. Yu, and L. Deng, “Investigation of full-sequence training of deep belief networks for speech recognition,” in Proc. Inter speech, 2010, pp. 2846–2849.

[2] I. G. Khaing, K. Z. Linn, “Myanmar Continuous Speech Recognition System based on DTW and HMM”, IJIET., Vol.2, Issue 1, February, 2013.

[3] F. Seide, G. Li, and D. Yu, “Conversational speech transcription using context-dependent deep neural networks,” in Proc. Interspeech, 2011, pp. 437–440

[4] G. Hinton, “Training products of experts by minimizing contrastive divergence,” Neural Comput., vol. 14, pp. 1771–1800, 2002

[5] D. Hau and K. Chen, “Exploring hierarchical speech representations using a deep convolutional neural network,” in Proc. 11th UK Workshop Comput. Intell. (UKCI ’11), Manchester, U.K., 2011.

[6] O. A. Hamid, A. R. Mohamed, H.Jiang, L. Deng, G. Penn and D. Yu, “Convolutional Neural Networks for Speech Recognition,” IEEE/ACM, Vol.22, No.10, October 2014.

[7] S. Zhang, S. Zhang, “Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyarmid Matching”, IEEE,1520-9210, 2017.

[8] M.M. Rahman and M. A. Bhuiyan, “Continuous Bangla Speech Segmentation using short time Speech Features Extraction Approaches”, IJACSA, Volume 3, No.11, 2012.

[9] L R Rabiner and M R Sambur, “An Algorithm for determining the endpoints of Isolated Utterances”, The Bell System Technical Journal, February 1975, pp 298-315.

[10] E. P.P. Soe and A. Thida, “Text to speech synthesis for Myanmar Language”, IJSER, Volume 4, Issue 6, June 2013.

[11] Vimala, C., Radha, V., “A review on speech recognition challenges and approaches”, World Computer. Sci. Inf. Technol., Vol. 2, (1), pp. 1-7, 2012.

[12] Y. Qian, P. C. Woodland., “Very Deep Convolutional Neural Networks for Noise Robust Speech Recognition,” IEEE/ ACM Transaction on Audio, Speech and Language Processing, pp. 2263-2276, December 2016.

[13] M. Kalamani, Dr. S. Valarmathy, C.Poonkuzhali and R. Karthiprakash, “Comparison of Cepstral and Mel Frequency Cepstral Coefficients for Various Clean and Noisy Speech Signl”, Proceedings of International Conference on Global Innovations In Computing Technology, ICGICT’14, Vol.2, Special Issue 1, March 2014.

[14] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen and M. Moonen, “ Sparse Linear Prediction and Its Applications to Speech Processing”, IEEE Transactions on Speech and Audio Processing, Vol. 20, No. 5, pp.1644, July 2012.