Isolated Myanmar Speech Recognition via ANN

Nan Phyu Phyu Hsan, Twe Ta Oo

University of Computer Studies, Yangon, Myanmar
phyuphyuhsan1994@gmail.com

Abstract

Automatic Speech Recognition (ASR) is a popular and challenging area of research in human computer interaction. This paper presents an isolated Myanmar speech recognition system that is speaker dependent as well as speaker independent and developed by using Artificial Neural Network techniques. In this system, the Mel Frequency Cepstral Coefficients extracted from the manually preprocessed words are considered as the features to acoustically identify the speeches. Those features are then used to train and test the Backpropagation neural network model. This system uses a database of 2800 utterances (names of the cities in Myanmar) by 10 talkers (4 males and 6 females), from which 2400 utterances are used for training and 400 are used for testing and recognition. As per the experimental results, the proposed system achieved the recognition rate of about 93.5% for known speakers (i.e., speaker dependent) and 76.5% for unknown speakers (i.e., speaker independent).

Keywords*: ASR, isolated speech, ANN, MFCC*

1. Introduction

Speech is a natural mode of communication in our daily lives. Speech recognition is the process in which the speaker's words are recognized based on the underlying information in the speech signal. Recognition techniques make us possible to use the speaker's voice for verifying his/her identity and controlling access to services like voice dialing, banking by telephone, telephone commerce, confidential information areas, and remote access to computers. The task of a speech recognizer is to automatically determine the spoken word, regardless of the variability introduced by speaker identity, manner of speaking, and environmental conditions.

Automatic Speech Recognition (ASR) is a popular and challenging area of research in human computer interactions. An ASR system involves two phases: training and testing. In the training phase, known speeches are recorded and then parametric representation of speeches are extracted and stored in the speech database. In the testing phase, parametric representation of the query speech is compared with the reference templates to recognize the utterance [3].

Speaking of an ASR system, feature extraction process that translates a speech signal into some acoustically identifiable digital form is the main part of the system. Obviously, a good feature may produce a good recognition result. In the literature, various kinds of feature extraction techniques such as Principal Component Analysis [11], Mel Frequency Cepstral Coefficients (MFCC) [9], Linear Predictive Coding (LPC) [4], and Wavelet Transform [8] have been proposed. In this system, the MFCC is used to extract the features from speech signals as it gives lower complexity and higher recognition accuracy in compared with other techniques [4].

In addition to feature extraction, recognition models also play an important role in ASR systems. Hidden Markov Model (HMM) [2], Artificial Neural Networks (ANNs) [7], Vector Quantization [5], and Dynamic Time Warping (DTW) [9] are some of the widely used recognition models. Among them, ANNs are utilized in many applications due to their parallel distributed memories, error stability, pattern learning, and distinguishing ability [7]. In the proposed system, backpropagation neural network is employed as the recognition model. Backpropagation is a supervised learning algorithm that uses gradient descent and it is also one of the most widely used ANN algorithms.

The following sections discuss the topics of related work of speech recognition in section 2, proposed system in section 3, system implementation in section 4, experimental results in section 5, and conclusion in section 6, respectively.

2. Related Work

Speech is the most commonly used form of communication among human beings. Speech can also be used as an interface to interact with machines. Speech recognition refers to a process by which an electronic device recognizes the speaker's voice and understands the meaning inferred by the speaker's voice. Researches for speech recognition have been started since 1930. Since then, there had been a lot of research experiments and achieved results in various

languages throughout the world [1]-[11] [13]. This section introduces some of those research works, especially developed for Myanmar language.

In 2003, Zaw Min Tun [13] proposed an ASR system for Myanmar language. In that system, MFCC was used as the front-end processing and then HMMs were constructed by Gaussian distribution function based on the MFCCs. That paper did not discuss anything about the recognition result as its main focus was to introduce how to develop an ASR system for Myanmar language.

In 2009, Aung Tun Tun Lwin [10] also proposed an ANN-based isolated Myanmar digits recognition system. That system discussed how to use the LPC for feature extraction and the feed forward multilayer perceptrons trained by the backpropagation for digit recognition. That system achieved an accuracy of more than 98% for speaker dependent isolated word recognition.

In 2015, Ei Mon Kyaw [6] proposed a speaker dependent Myanmar speech command recognition system by using MFCC and DTW techniques. That system was intended to recognize 10 commands from real-time microphone input. The accuracy was 90% in lower noisy speech condition, 80% in medium noisy speech condition, and 40% in high noisy speech condition.

In 2015, Su Myat Mon and Hla Myo Tun [5] also proposed a speech-to-text conversion system by using MFCC and HMM. That system used the MFCC feature vectors extracted from the original speech signals as the observation sequences of the HMM recognizer. That system only emphasized the speaker dependent recognition and achieved the recognition rate of 87.6%.

In 2016, Zin Zin Tun and Gun Srijuntongsiri [1] proposed a speech recognition system for Myanmar digits based on HMM using HTK Tools. That system recognized the speech utterances by converting the speech waveform into a set of feature vectors using the MFCC. It was stated that the system yielded the significant results for both context independent and context dependent models.

Although there had been a lot of researches done for Myanmar ASR, it is still far from a mature field. This paper focuses on an ASR system for isolated Myanmar words. Firstly in this system, the utterances are manually preprocessed by using the Audacity software [14]. As per the literature, the ZCR and STE techniques can be used to preprocess the speech. However, according to our experiments, the ZCR is not always promising for endpoint detection of words and using the STE alone is not good enough for feature extraction from speech. Thus in this system, the start and end points of each word are manually detected. The MFCC features are then extracted from the preprocessed words, which will then be used to train and test the neural network. According to the experimental results, it is found out that the proposed method is easy to implement and yields the satisfying recognition accuracy.

3. Proposed System

The proposed system is made up of three parts: preprocessing, feature extraction, and classification. The following subsections discuss those parts in detail.

3.1 Preprocessing

This paper presents an ASR system that recognizes the isolated Myanmar words, which are the names of the cities in Myanmar. The proposed system is developed by using the MFCC and ANN techniques. All the isolated words used in training and testing phases are recorded by using mobile phones in a normal room condition.

As human vocal tract system uses air flow to speak out a word, it is normal that each word is surrounded by silences. As those silences are not any help to speech recognition and also time consuming to process, it is important in ASR systems to remove those unwanted silence parts and to detect the exact start and end points of each word. In this system, the start and end points of each utterance is manually detected by using Audacity software.

3.2 Feature Extraction

When the input data to an algorithm is too large to be processed and it is suspected to be notoriously redundant, the input data should then be transformed into a reduced representation set of features (also named features vector). Transforming the input data into a set of features is called feature extraction [2]. The aim of feature extraction is to reduce the data size of a speech signal without losing acoustically identifiable components before pattern classification or recognition.

In this paper, Mel frequency cepstral coefficients are considered as the features of input speech. MFCC uses the Mel scale which is based on the human ear scale [3] and it is one of the most powerful feature extraction techniques used in ASR systems. The main steps of MFCC calculation are:

***Pre-emphasis and framing*:** Input speech is cropped to remove silence or acoustical interference that may be present in the beginning or end of the sound file and divided into frames (20-40ms long).

***Windowing*:** Reduces discontinuities of the speech by tapering the beginning and end of each frame to zero.

***Fast Fourier Transform (FFT)*:** Converts each frame from time domain to frequency domain.

***Mel filterbank and logarithmic transformation*:** Converts the frequency domain signal to Mel frequency scale that is more appropriate for human hearing and perceptions.

$$F(mel) = 2595 * \log_{10}\left(1 + \frac{f}{700}\right) \quad (1)$$

where *f* is the FFT signal from the previous step.

Discrete Cosine Transform (DCT): It converts the log Mel spectrum into the time domain to obtain the Mel frequency cepstral coefficients.

In general, the number of cepstral coefficients of MFCC depends on the number of filters used in the Mel filterbank (20 filters by default and used in this system), i.e. 20 feature vectors for 20 filters. If the signals contain a lot of closely spaced frequencies and want to resolve them, the number of filters might be increased. Among the resulting feature vectors, the first feature vector represents the average power in the speech signal and it is not often used in recognition applications because the average power varies considerably depending on the microphone placement and channel [12]. Thus in this system, the 19 feature vectors apart from the first one are used to represent the input speech and to train the neural net.

Before applying the MFCC features to the neural network, there is a point needed to be considered. The resulting MFCC features determine the number of input neurons of the neural net, *n* features means *n* input neurons. Traditional neural networks use the same network topology for all training patterns and thus the size of MFCC features needs to be the same for all input patterns. However, speech is a time-varying data. Even for the same word uttered by the same person, the size of the feature vector may be different from time to time and yields to time-varying input neurons. In order to solve that problem in this system, the feature vectors of all training data are padded with zero in such a way that they must be the same in length with the longest anticipated feature.

3.3 Classification

In this paper, a backpropagation neural network is utilized as the recognition model. This system consists of two phases: training and testing. During training, each input pattern has an associated target pattern. The whole objective of training is to find a set of network weights that provides a solution to the specific problem at hand. Before training converges, the weights are set to small random values. The input patterns are presented to the network one by one; then an error vector for each pattern is calculated and the weights are changed accordingly. This process is repeated until the error for all patterns during one epoch are within tolerance. Finally, the stable weights are stored and then used in testing.

The network topology of the proposed system is shown in Fig. 1. There are 1800 input nodes (each represents an MFCC feature), 100 hidden nodes, and 20 output nodes (represents 20 names of the cities in Myanmar).

General flow of the proposed speech recognition system is shown in Fig. 2. For training phase,

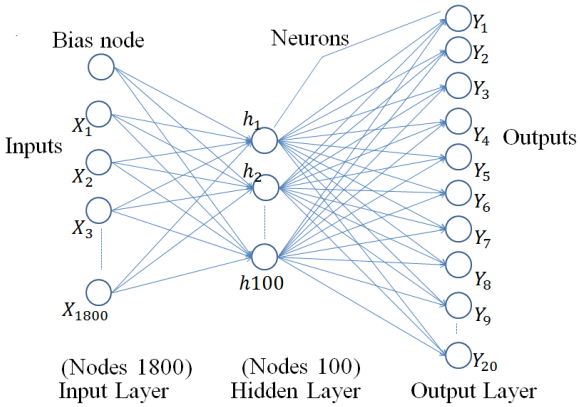

Figure 1. Neural network topology of the proposed system

Step 1: The recording speeches in *.amr* format are converted to *.wav* format.

Step 2: Audacity software is used to detect the start and end point of each word.

Step 3: The MFCC features are extracted from the preprocessed words and padded with zero if necessary so that the size is to be 1800 (the longest feature size in this system).

Step 4: The backpropagation algorithm is used to train the neural network and get the stable weight from the MFCC features.

For testing phase, the above steps 1-3 are applied to the query speech and the resulting MFCC

features are used to test the recognition model with the stable weights stored during the training phase.

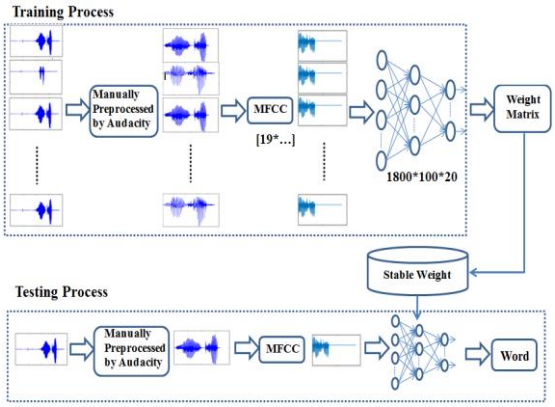

Figure 2. System flow of the proposed speech recognition system

4. System Implementation

This system, simulated by MATLAB, develops a speaker independent recognition system for Myanmar Language that tries to recognize the following 20 names of the cities in Rakhine state, Shan state, and Kachin state in Myanmar.

မိုးညှင်း၊ မိုးကောင်း၊ မိုးမောက်၊ မြောက်ဦး၊ မိုင်းဆတ်၊ မိုင်းဖြတ်၊ မိုင်းရယ်၊ မိုင်းမော၊ မိုးနဲ၊ မိုင်းပန်၊ မိုင်းကိုင်၊ မိုင်းခတ်၊ မောက်မယ်၊ မိုင်းယောင်း၊ မိုင်းလာ၊ မိုင်းယန်၊ မိုင်းနောင်၊ မိုင်းဇင်္ဂ၊ မိုင်းယု၊ မိုင်းခုတ်။

This system uses a database which is made up of training and testing datasets with 2400 and 400 utterances each. The 10 talkers (4 males and 6 females) participated in training and testing say each word thirteen times. The duration of each recorded speech is 1-2 seconds and the whole database is approximately 1.57 hours long. Recording format is WAV and sampling rate is 8000 Hz.

The performance of ASR systems is generally measured in terms of word recognition rate which is the ratio between correctly classified words and total number of tested words. In this system, the accuracy or recognition rate is computed by the following equation.

$$\text{Accuracy} = \frac{\text{No. of words correctly recognized}}{\text{Totalno.of words}} \times 100\% \quad (2)$$

5. Experimental Results

The proposed system is trained with the back-propagation algorithm by using the MFCC features of 2400 utterances, momentum value of 0.1 and learning rate of 0.4. During the training phase, the network becomes stable at 800-epoch with error of 0.00019 while the error tolerance is set to 0.0002.

Table 1 shows the recognition result for speaker dependent testing with 200 test words. From Table 1, it can be seen that the words မိုင်းဇင်္ဂ၊ မိုင်းပန်၊ မိုင်းရယ်၊ မိုင်းယန်၊ မိုင်းယု၊ မိုးကောင်း၊ မိုးမောက်၊ မိုးနဲ၊ မိုးညှင်း၊ မြောက်ဦး၊ and မိုင်းယောင်း are fully recognized with 100% accuracy. Some words such as မိုင်းခတ်၊ မိုင်းခုတ်၊ မိုင်းလာ၊ မိုင်းမော၊ and မိုင်းဖြတ် are recognized with 90% accuracy and the remaining words appear as the least recognized ones with 80% accuracy. Averagely, the proposed system achieved 93.5% accuracy.

Regarding speaker independent recognition, the proposed system is tested with 200 test words uttered by new speakers different from the ones participated in training. It can be seen from Table 2 that the words မိုင်းကိုင်၊ မိုင်းယု၊ မိုးမောက်၊ မြောက်ဦး၊ and မောက်မယ် are fully recognized with 100% accuracy, whereas the words မိုင်းခတ်၊ မိုင်းမော၊ မိုင်းနောင်၊ မိုင်းပန်၊ မိုင်းရယ်၊ မိုးကောင်း၊ မိုးနဲ၊ and မိုးညှင်း are recognized with 80% to 90% accuracy. Unfortunately, the word မိုင်းယန် is the least recognized one with 30% accuracy. It is because the voices of the new speakers who speak out မိုင်းယန် are too quiet and the way of speaking is also noticeably different from the trained speakers. As an average, the system achieved the speaker independent recognition rate of 76.5%.

From the results in Table 1 and 2, it can be seen that the recognition rate of the proposed system is not satisfying for speaker independent testing. During the experiment, it was found out that the performance of the recognizer depends on the speakers' way of speaking. If the new speakers sound alike the trained ones, the recognizer can recognize well; otherwise it fails. It is because the number of trained speakers in this system is too few and thus the recognizer fails to generalize new speakers.

In addition, the speakers participated in this experiment are just graduate students who have no knowledge and experience about how to pronounce the words that will be helpful for ASR systems. Thus, their way of speaking such as loudness and rate of speaking might also be the problems. Moreover, the speeches used in this system were recorded with mobile phones in a normal room condition. Thus, the characteristics of recording devices in mobiles and background noises might also hinder the performance of the system.

6. Conclusion

Generally, it is difficult for a speech recognition system to get 100% accuracy. The proposed system in this paper only achieved the accuracy of 93.5% for known speakers and 76.5% for unknown speakers. As discussed in the previous section, the accuracy of the system was below 100% probably due to some real life problem such as using the poor microphone, noisy environment, poor utterance of speakers, and choice of the neural network parameters as well.

In addition, speech database used in this system is too small, just 2800 utterances, and the number of trained speakers is also very few. If the size of the database and the number of speakers were increased, the recognition model could have achieved more generalization.

Thus in the future, more suitable training data set can be used to improve the performance of the recognizer. Additional to the above mentioned facts, we can also play various parameters of neural network such as the error threshold, learning rate, and the number of hidden node in order to get different, perhaps better result. In addition, other recognition tools like HMM, recurrent neural network (RNN), and long short term memory (LSTM) that can efficiently deal with time-series data like speech can also be tried out.

Table 1. Speaker dependent recognition result

Myanmar word	No. of samples for testing	Testing result	
		No. of properly recognized words	Recognition rate (%)
မိုင်းခတ်	10	9	90
မိုင်းကိုင်	10	8	80
မိုင်းခုတ်	10	9	90
မိုင်းလာ	10	9	90
မိုင်းမော	10	9	90
မိုင်းနောင်	10	8	80
မိုင်းဝေ့	10	10	100
မိုင်းပန်	10	10	100
မိုင်းဖြတ်	10	9	90
မိုင်းဆတ်	10	8	80
မိုင်းရယ်	10	10	100
မိုင်းယန်	10	10	100
မိုင်းယု	10	10	100
မိုးကောင်း	10	10	100
မိုးမောက်	10	10	100
မိုးနဲ့	10	10	100
မိုးညှင်း	10	10	100
မြောက်ဦး	10	10	100
မောက်မယ်	10	8	80
မိုင်းယောင်း	10	10	100
Total	**200**	**187**	**93.5**

Table 2. Speaker independent recognition result

Myanmar word	No. of samples for testing	Testing result	
		No. of properly recognized words	Recognition rate (%)
မိုင်းခတ်	10	9	90
မိုင်းကိုင်	10	10	100
မိုင်းခုတ်	10	6	60
မိုင်းလာ	10	5	50
မိုင်းမော	10	8	80
မိုင်းနောင်	10	8	80
မိုင်းဝေ့	10	7	70
မိုင်းပန်	10	8	80
မိုင်းဖြတ်	10	6	60
မိုင်းဆတ်	10	4	40
မိုင်းရယ်	10	9	90
မိုင်းယန်	10	3	30
မိုင်းယု	10	10	100
မိုးကောင်း	10	8	80
မိုးမောက်	10	10	100
မိုးနဲ့	10	8	80
မိုးညှင်း	10	9	90
မြောက်ဦး	10	10	100
မောက်မယ်	10	10	100
မိုင်းယောင်း	10	5	50
Total	**200**	**153**	**76.5**

References

[1] Zin Zin Tun and Gun Srijuntongsiri, "A speech recognition system for Myanmar digits," International Journal of Information and Electronics Engineering, vol. 6, no. 3, pp. 210-213, May 2016.

[2] M. D. Abdullah-al-MAMUN and F. Mahmud, "Performance analysis of isolated bangla speech recognition system using hidden markov model," International Journal of Scientific and Engineering Research, vol. 6, issue 1, January 2015.

[3] B. Medhi and P. H. Talukdar, "Isolated assamese speech recognition using artificial neural network," Indian Institute of Technology Guwahati (IITG), May 2015.

[4] M. R. Gamit and K. Dhameliya, "Isolated words recognition using MFCC, LPC and neural network," International Journal of Research in Engineering and Technology (IJRET), vol. 4, issue 6, June 2015.

[5] Su Myat Mon and Hla Myo Tun, "Speech-To-Text conversion (STT) system using hidden

markov model (HMM)," International journal of scientific and technology research, vol. 4, issue 6, June 2015.

[6] Ei Mon Kyaw, "Speech command recognition using Dynamic Time Warping," University of Computer Studies, Yangon, July 2015.

[7] S. Sunny, D. Peter, and K. P. Jacob, "Performance of different classifiers in speech recognition," International Journal of Research in Engineering and Technology (IJRET), vol. 2, issue 4, April 2013.

[8] E. R. Rady, A. H. Yahia, E. A. El-Dahshan, and H. El-Borey, "Speech recognition system based on wavelet transform and artificial neural network," Egyptian Computer Science Journal (ECS), vol. 37, no. 3, May 2013.

[9] W. Gevaert, G. Tsenov and V. Mladenov, "Neural Network used for speech recognition," Journal of Automatic Control, University of Belgrade, vol. 20, no. 1, June 2010.

[10] Aung Tun Tun Lwin, "Speech recognition for Myanmar digits using Neural Networks with LPC approach," University of Computer Studies, Yangon, May 2009.

[11] T. Takiguchi and Y. Ariki, "PCA-based speech enhancement for distorted speech recognition," Journal of Multimedia, vol. 2, no. 5, September 2007.

[12] L. Deng and D. O'Shaughnessy, "Speech processing," Marcel Dekker, New York, 2003.

[13] Zaw Min Tun, "Automatic speech recognition for Myanmar language," University of Computer Studies, Yangon, October 2003.

[14] https://www.audacityteam.org