Advances and Applications in Mathematical Sciences
Volume 21, Issue 9, July 2022, Pages 5457-5467

COMPARATIVE ANALYSIS OF MULTILINGUAL ISOLATED WORD RECOGNITION USING NEURAL NETWORK MODELS

BRAJEN KUMAR DEKA and PRANAB DAS

Department of Computer Applications
Assam Don Bosco University
Guwahati, India
E-mail: brajendeka@gmail.com
pranab.das@dbuniversity.ac.in

Abstract

The feature selection technique is essential in the overall performance of multiple language isolated word recognition systems. The main goal of this research is to see what kind of results can be obtained when using a neural network classifier to develop speech recognition systems in several languages. A neural network is a well-known method for classifying non-linear situations. On a recently developed multilingual speech database, the Mel Frequency Cepstral Coefficient (MFC) is used to compare the performance of the Artificial Neural Network (ANN) and the Recurrent Neural Network (RNN) as the function vector in this paper. The Multilanguage speech database contains data from 96 male and female speakers who have reported their speech. Speech samples were obtained in three different languages: English, Hindi, and Assamese. The Multilanguage speech recognition system achieved 94.4% (with ANN) and 97.5 % accuracy (with RNN).

I. Introduction

The speech recognition system is a type of natural communication that aims to give computer intelligence to communicate with people [1]. It connects with the computer by using a microphone as an input device and transforming the spoken word into an audio signal as an output signal. Speech recognition is the process of converting a speech signal into a sequence of words [2]. To develop techniques and frameworks for speech

2020 Mathematics Subject Classification: 68T07, 68T10.

Keywords: Artificial Neural Network, Keyword Spotting, Mel Frequency Cepstral Coefficient, Multilingual, Speech Recognition.

Received December 21, 2021; Accepted March 1, 2022

input to the machine through an algorithm implemented as a computer program.

Keyword Spotting (KWS) is an audio mining component that deals with the long-term detection of such keywords. People want to make human-machine communication as natural as possible in the current scenario so that the complexity of this subject has increased significantly. Human speech includes not only meaningless words but also unwanted noises like cough, exclamation, and noise. If we can extract only the embedded information, the computation can be much efficient and robust. For data processing tasks that process an increasingly large amount of expressions, such as real-time keyword monitoring and audio content indexing, keyword spotting is well-matched. Keyword Spotting is a technologically critical problem and plays a vital role in audio indexing and speech data mining applications [3]. A KWS gives a keyword or multiple instances of keywords and an utterance stream that is the search space where it has to find the keyword if present. A KWS that mainly detects relevant words of occurrences will be more efficient than a fully-fledged LVCSR engine [4] [5].

This paper compares a Multilanguage isolated word recognition based on a Recurrent Neural Network (RNN) and an Artificial Neural Network (ANN) methodology. The results of language-specific word recognition using RNN-based and ANN-based methods for English, Hindi, and Assamese, respectively. An RNN is a network of neurons with feedback connections that can be said replica of the human brain. RNN can understand a lot of behaviour and gain knowledge that many traditional machine learning approaches struggle to acquire. On the other hand, ANN is a well-known feed-forward neural network classifier to solve many problems, including classification and pattern recognition.

According to the literature review, there have been few studies in Indian regional languages for multilingual purposes, but a lot of effort has been done in foreign languages for multilingual word recognition systems. Multi-language speech information for six Indian dialects has been used in this study, including Hindi, Kannada, Malayalam, Marathi, Tamil, and Telugu, as well as Indian English as the seventh language [1]. Multilingual speech recognition and linguistic understanding are essential for the development of multilingual spoken communication systems [6][7]. Because of the various

languages spoken in a country like India, research on the state-of-the-art speech recognition is important.

The remaining paper is organized as follows: Section 2 describes the multilingual interfaces. The methodology of the system and database are discussed in Section 3. Section 4 presents the experimental results and discussion. The conclusion and future direction of this study are reported in Section 5.

II. Multilingual Interfaces

In the area of digital technology and information processing, creative development has been going on for ten years. With the rapid development of the very large scale integrated circuit (VLSI) technology, we were able to achieve very high data processing and computation speed. This has made real-life applications of increasingly complex algorithms a reality. Speech recognition algorithms are one of the most significant among them. In many multimedia applications, state-of-the-art speech recognition technology now finds its place in spoken language command interfaces, knowledge query systems, and many more such as biometric authentications, surveillance, etc. [9]. Most of today's speech interfaces essentially expect the user to speak in a given language, i.e. monolingual. But as the spread of the framework grows, we also have to come across the types of users whose expressions include multi-languages words [10]. With time, the need for systems or interfaces that can handle speech multi-languages behaviour will increase. One thing is quite clear about a multi-languages utterance dominating of one particular language compared to the other. It means that at unexpected intervals and some portion of the speech, the words that refer to the later occur. In the first step from a shared language, identifying such vocabulary will be the main phase of a desired multi-languages program. If the system has 'spotted' the words from a given language, it can take reasonable measures that can require adaptive encoding for the application style of a command interface or can be feed as an input to a sophisticated module to perform the role of more language-specific processing. In this way, it is possible to build a better and more powerful multi-languages system. A typical system block diagram is shown here in Figure 1.

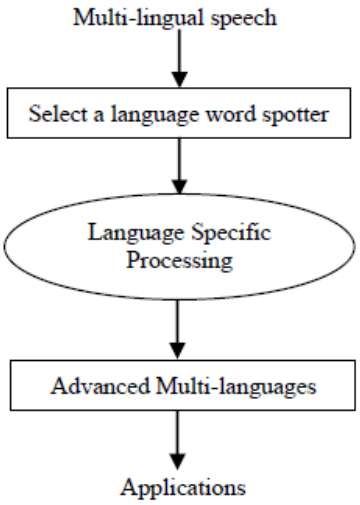

Figure 1. Multi-languages system recognizes specific words.

Here the word spotter includes all other essential voice recognizer modules such as the extractor feature, the trained decode etc. The system would integrate the 'Language-Specific word spotter of a multi-languages system, as well as this feature, will play a significant role in the overall success of the system as a whole.

III. Materials and Methods

3.1 Multilingual Speech Databases

The primary prerequisite for experimenting with speech processing is the speech database. A collection of utterances is required to study the speech recognition system. The speech database is a collection of words for three languages that correctly consider English, Hindi, and Assamese. The speech recognition program uses samples in the database for training and performance testing. Speech utterances from 48 male and 48 female native speakers are collected in this work to create the speech database. These utterances in three registered languages are of seven days, ten digits and twelve months. A speaker pronounces each isolated word five times. It contains a collection of 4640 vocabulary sizes for speech. Such samples will archive in the .wav file. Voice samples are recorded in a closed room, with high sound quality, in an almost noise-free environment. 16 kHz is the

designated sampling frequency. The recording has achieved at room temperature and natural humidity using mono channel mode. The 16-bit portable resolution microphone is positioned approximately 12-15 cm from the speaker's mouth. The above-constructed speech database includes a few speech samples for the unavailability of this kind of database for some isolated word recognition system languages. For evaluating such sample numbers, a speaker-dependent word recognition system technique was considered [8].

3.2 Multilingual Isolated Word Recognition Using ANN and RNN Methods

The area of keyword spotting is perhaps the most current area of investigation in human-computer interaction. ANN is a type of supervised learning algorithm that use in disciplines like data mining and pattern recognition. The ANN being fundamentally parallel, it can work in a comparable structure where all the neurons present in a layer perform figuring at their level of single machine activity [6]. RNN-based algorithms have also found a suitable place for problem speech processing and speech synthesis for their capability to capture time-varying properties. This classification technique can use recorded high-performance accuracy for pattern recognition. In our research, we have adopted algorithms based on ANN and RNN for classifying the isolated words of days, digits and months considering the language of English, Hindi and Assamese using MFC as the component vector. Since we have 7-days, 10-digits and 12-months, the target vector contains seven classes, ten classes and twelve classes.

3.2.1 Artificial Neural network

ANN is a computer having its architecture modelled after the brain. They mainly involve many simple processing units wired together during a complex communication network. Each simple processing unit represents a real neuron, and if it receives a robust signal from the opposite connection unit, it will send a replacement signal or trigger signal [11]. Artificial Neurons are the basic unit of an Artificial Neural Network that simulates the four main functions of a biological neuron. It is a mathematical process formulated as a natural neuron model.

Figure 2 shows the various inputs represented by the character, $i\ (n)$.

Each of these inputs has multiplied by connecting weights $w(n)$. Usually, the product is added and fed into the transfer function to generate the output result. The sigmoid function uses as a transfer function. Applications like speech recognition and keyword spotting are required to turn these real-world inputs into discrete values. These functions don't continually utilize networks composed of neurons that sum, and thereby smooth, inputs.

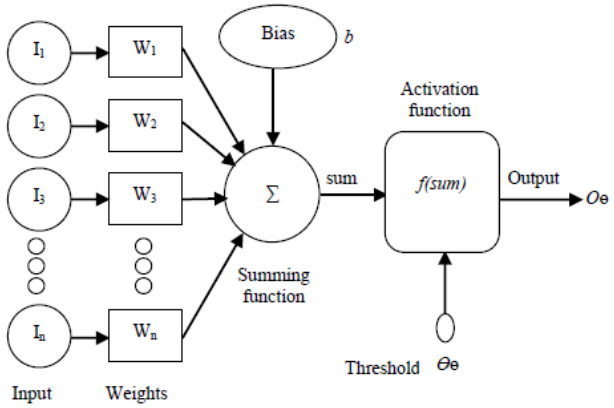

Figure 2. Basic Artificial Neuron.

3.2.2 Feed-Forward Network

The simplest and basic type of ANN is the feed-forward network. In this network, the hidden nodes only forward information from the input nodes to the output nodes. This network contains no loops or cycles. If 'b' is greater than 'a,' a neuron in layer 'a' can only accept input from layer 'b'. The embedded system's approach for modifying the free parameters of the neural network through continuous simulation is called learning. The back-propagation method was used to produce multi-layer perceptrons, the most modern type of layered Feed-Forward network. Perceptrons are frequently arranged in groups of two or more. One input layer, one or two hidden levels, and one output layer make up the structure. The hidden layer is essential because it serves as a feature extractor. To generate complex input functions, it uses a non-linear function like a sigmoid or a radial base. The output layers act as a logical net that chooses an index sent to the output depending on the hidden layer information it receives to minimize the classification error [12] [13]. Figure 3 shows the feed-forward fully connected with one hidden layer and one output layer.

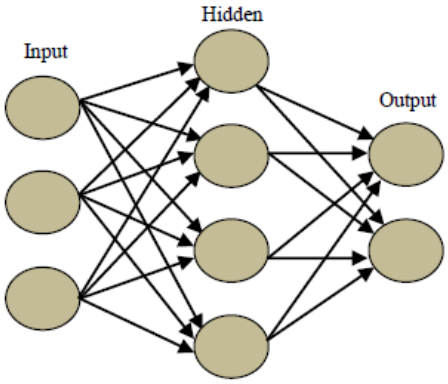

Figure 3. A Feed-Forward fully connected with one hidden layer and one output layer.

3.2.3 Recurrent Neural Network

A recurrent neural network (RNN) is a kind of neural network that runs in time. RNN accepts the input vector, updates its hidden state through an activation function, and uses it to predict the output. Neuron output gets multiplied by a weight and sent back to neuron inputs with a delay in this network. RNN has achieved better spotting accuracy levels for keywords than MLP, but again, the training algorithm is more complex and dynamically sensitive, which may cause problems [14]. The neural network shown in Figure 4 has weight connections from the input layer to the hidden layer and then to the output layer. The difference between this network and Feed-forward networks is that the output layer has a link that connects back to the input layer. With this network structure, unlike a feed-forward network that only learns from specific training data, it will learn and adapt every time data is processed through the network. This network structure seems to be very useful in applications involving the processing of arbitrary input sequences, such as keyword spotting detection.

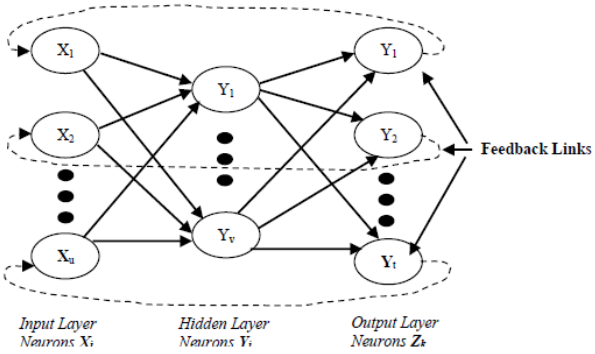

Figure 4. Recurrent Neural Network.

IV. Results and discussion

Consider the following steps:

• Concerning English, Hindi and Assamese with a 96-speaker consisting of 48 male and 48 female speakers who speak five times each, isolated words are reported predominantly in three languages: seven days, ten digits and twelve months.

• There are 4640 different samples of words recorded. The training set consisted of 3712 audio samples, with the remaining 928 audio samples possibly being used for testing.

• Creation of feature vectors for training and testing uses for the ANN and RNN algorithms.

• To test the algorithms using testing vectors.

• To use such samples not used for training, testing for isolated word recognition from multiple languages.

• Comparison of both algorithms results.

The recognition of the multiple languages isolated words considers the acoustic features for training and testing purposes. The Scaled Conjugate Gradient algorithm for the Artificial Neural Network gives an overall recognition rate of 94.4% [8] when the Recurrent Neural Network gives a recognition rate of 97.5% which is better than that of the ANN algorithm. Table 1 displays the accuracy rate using the classification based on ANN and

RNN for isolated words of days, digits, and months in multiple languages, respectively. Figures 5, 6 and 7 show the performance based on the ANN and RNN classifier for the multiple languages isolated words recognition.

Table 1. Accuracy rate for multiple languages isolated words based on ANN and RNN classifier.

		Days		Digits		Months	
Languages	Gender	ANN	RNN	ANN	RNN	ANN	RNN
English	Male	80.4	80.5	93.5	93.8	87.5	82.1
	Female	83.9	84.3	94.1	96.2	86.5	79.3
	Male and Female	82.15	82.4	93.8	95	87	80.7
Hindi	Male	80.4	81.3	93.7	93.8	80.2	82.1
	Female	83.9	79.3	92.5	95	84.4	83.5
	Male and Female	79.5	80.3	93.1	94.4	82.3	82.8
Assamese	Male	80.4	78.6	93.9	98.8	84.1	82.3
	Female	78.6	83.8	94.9	96.2	85.7	84.3
	Male and Female	79.5	81.2	94.4	97.5	84.9	83.3

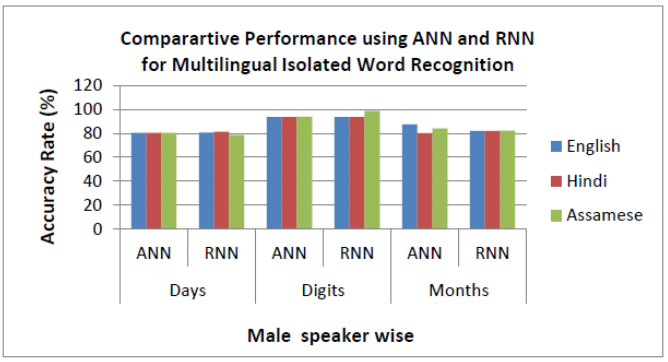

Figure 5. Performance-based on ANN and RNN by Male Speaker wise.

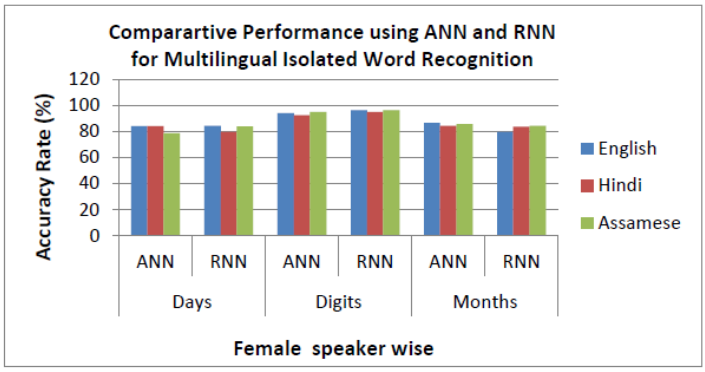

Figure 6. Performance-based on ANN and RNN by Female Speaker wise.

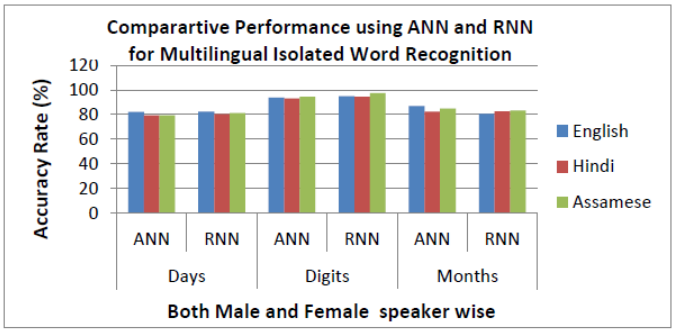

Figure 7. Performance-based on ANN and RNN by both Male and Female Speaker wise.

V. Conclusion

This paper presents a comparative analysis of ANN and RNN based algorithms for multiple languages isolated word recognition. Mel Frequency Cepstrum Coefficients is very reliable. The performance of the networks is heavily dependent on the standard pre-processing. Both the Backpropagation algorithm of the Multilayer Feed-forward Network and the Recurrent Neural Network produce satisfying results. The RNN classifier has a 97.5% accuracy rate, which is higher than the 94.4% achieved by the ANN-based classifier. Implementing several classifiers effectively and designing a fully functional system are some future directions to consider. The scope of this work ranges from extensive to continuous words.

References

[1] G. Hemakumar, P. Punitha, Speech recognition technology: A survey on Indian languages. International Journal of Information Science and Intelligent System 2(4) (2013), 1-38.

[2] M. K. Mand, D. Nagpal and Gunjan, An analytical approach for mining audio signals, International Journal of Advanced Research in Computer and Communication 2(9) (2013), 1346-1349.

[3] A. Jansen and P. Niyogi, Point process models for spotting keywords in continuous speech, IEEE transactions on Audio, Speech and Language Processing 17(8) (2009), 1457-1470.

[4] B. K. Deka and P. Das, A review of keyword spotting as an audio mining technique, International Journal of Computer Science and Engineering 7(1) (2019), 757-769.

[5] E. Chandra and K. A. Senthildevi, Keyword spotting: an audio mining technique in speech processing, A survey, IOSR Journal of VLSI and Signal Processing 5(4) (2015), 22-27.

[6] V. K. Jain, N. Tripathi, Speech features analysis and biometric person identification in multilingual environment, International Journal of Scientific Research in Network Security and Communication 6(1) (2018), 7-11.

[7] H. Bahi, N. Benati, A new keyword spotting approach, International Conference on Multimedia Computing and Systems, Proceedings (2014), 77-80.

[8] B. K. Deka, P. Das, Isolated keyword spotting in multilingual environment using ANN and MFCC, International Journal of Engineering and Advanced Technology 9(4) (2020), 5-8.

[9] H. Sing, Audio search of surveillance data using keyword spotting and dynamic models, M. Phil Thesis, the Chinese University of Hong Kong, (2001).

[10] P. Kumar and S. L. Lahudkar, Automatic speaker recognition using LPCC and MFCC. International Journal on Recent and Innovation Trends in Computing and Communication 3(4) (2015), 2106-2109.

[11] B. C. Kamble, Speech recognition using artificial neural network – A review, International Journal of Computing, Communications and Instrumental Engineering (IJCCIE) 3(1) (2016), 1-4.

[12] S. Shetty, K. K. Achary, Audio Data Mining Using Multi-perceptron Artificial Neural Network, IJCSNS International Journal of Computer Science and Network Security 8(10) (2008), 224-229.

[13] S. Patil, S. Ghorpade, R. Chaudhary, Evolution of artificial neural network along with comparative study and it's challenges, International Journal of Research and Analytical Reviews (IJRAR) 7(4) (2020), 597-605.

[14] R. L. K. Venkateswarlu, R. V. Kumari and G. V. JayaSri, Speech recognition by using recurrent neural networks, International Journal of Scientific and Engineering Research 2(6) (2011) 1-7.