Speech Recognition Using Artificial Neural Network – A Review

Bhushan C. Kamble[1]

***Abstract*--**Speech is the most efficient mode of communication between peoples. This, being the best way of communication, could also be a useful interface to communicate with machines. Therefore the popularity of automatic speech recognition system has been greatly increased. There are different approaches to speech recognition like Hidden Markov Model (HMM), Dynamic Time Warping (DTW), Vector Quantization (VQ), etc. This paper provides a comprehensive study of use of Artificial Neural Networks (ANN) in speech recognition. The paper focuses on the different neural network related methods that can be used for speech recognition and compares their advantages and disadvantages. The conclusion is given on the most suitable method.

***Keywords*--**Neural Networks, Training Algorithm, Speech Recognition, Artificial Intelligence, Feature Extraction, Pattern Recognition, LPC, MFCC, Perceptron, Feedforward Neural Networks, etc.

I. INTRODUCTION

SPEECH is probably the most efficient and natural way to communicate with each other. Humans learn all the relevant skills during early childhood, without any instruction, and they continue to rely on speech communication throughout their life. Humans also want to have a similar natural, easy and efficient mode of communication with machines. Therefore they prefer speech as an interface rather than using any other hectic interfaces like mouse and keyboards. But the speech is a complex phenomenon as the human vocal tract and articulators, being the biological organs, are not under our conscious control.

Speech is greatly affected by accents, articulation, pronunciation, roughness, emotional state, gender, pitch, speed, volume, background noise and echoes [1].

Speech Recognition or Automatic Speech Recognition (ASR) plays an important role in human computer interaction. Speech recognition uses the process and relevant technology to convert speech signals into the sequence of words by means of an algorithm implemented as a computer program. Theoretically, there should be the possibility of recognition of speech directly from the digitized waveform [2]. At present, speech recognition systems are capable of understanding of thousands of words under functional environment.

Speech signal provides two important types of information: (a) content of speech and (b) identity of speaker. Speaker recognition deals with the extraction of identity of speaker [3].

Speech recognition technology can be a useful tool for various applications. It is already used in live subtitling on television, as dictation tools in medical and legal profession and for off-line speech-to-text conversion or note-taking systems [4]. It has also many applications like telephone directory assistance, automatic voice translation into foreign languages, spoken database querying for new and unexperienced users and handy applications in field work, robotics and voice based commands [5].

II. SPEECH RECOGNITION PROCESS

The process of speech recognition is complex and a cumbersome job. The following figure 1 shows the steps involved in the process of speech recognition.

2.1 Speech

Speech is the vocalized form of human interactions. In this step, the speech of the speaker is received in waveform. There are many software available which are used to record the speech of humans. The acoustic environment and transduction equipment may have great effect on the speech generated. We can have background noise or room reverberation along with the speech signal which is completely undesirable.

2.2 Speech Pre-processing

Speech pre-processing is intended to solve such problems. This plays an important role in eliminating the irrelevant sources of variation. It ultimately improves the accuracy of speech recognition. The speech pre-processing generally involves noise filtering, smoothing, end point detection, framing, windowing, reverberation cancelling and echo removing [6].

[1]Student, Dept. of Mechanical Engineering, JDIET, Yavatmal, India

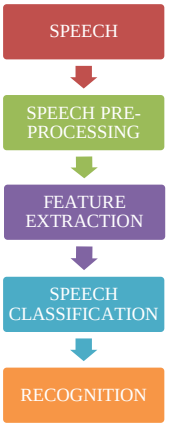

Fig. 1: Process of Speech Recognition

2.3 *Feature Extraction*

The speech varies from person-to-person. This is due to the fact that every person has different characteristics embedded in utterance. Theoretically, possibility should be there to recognize speech from the digitized waveform. But due to the large variation in speech signal, there arise a need to perform some feature extraction to reduce that variations. The following section summarizes some of the feature extraction technologies that are in use nowadays. These techniques are also useful in other areas of speech processing [7].

MFCC – Mel Frequency Cepstrum Coefficients (MFCC) is the most prominent method used in the process of feature extraction in speech recognition. It is based on the frequency domain which is based on Mel scale based on human ear scale. MFCCs, being frequency domain features, are more accurate than time domain features [8]. MFCC represents the real cepstral of windowed short time signal which is derived from Fast Fourier Transform (FFT). These coefficients are robust and reliable for variations of speaker and operation environment.

LPC – Linear Predictive Coding (LPC) is a tool most widely used for medium or low bit rate coder. Digital signal is compressed for efficient transmission and storage. Computation of parametric model based on least mean squared error theory is known as linear prediction (LP). The signal is expressed as a linear combination of previous samples. Format frequencies are the frequencies where resonance peak occurs [9].

2.4 *Speech Classification*

The most common techniques used for speech classification are discussed in short. These system involve complex mathematical functions and they take out hidden information from the input processed signal.

HMM – Hidden Markov Modelling (HMM) is the most successfully used pattern recognition technique for speech recognition. It is a mathematical model signalized on the Markov Model and a set of output distribution. This technique is more general and has a secure mathematical foundation as compared to knowledge based approach and template based approach. In this method, speech is split into smaller audible entities and these entities represent a state in the Markov Model. According to the probabilities of transition, there exists a transition from one state to another [10].

DTW – Dynamic Time Warping (DTW) technique compares words with reference words. It is an algorithm to measure the similarity between two sequences that can vary in time or speed [11]. In this technique, the time dimensions of the unknown words are changed until they match with that of the reference word.

VQ – Vector Quantization (VQ) is a technique in which the mapping of vector is performed from a large vector space to a finite number of region in that space. This technique is based on block coding principle. Each region is called as cluster and can be represented by its centre known as a code-word. Code book is the collection of all code-words [12].

III. Artificial Neural Network from the Viewpoint of Speech Recognition

3.1 *What is Artificial Neural Network?*

Artificial Neural Networks (ANN) are nothing but the crude electronic models based on neural structure of brain. The human brain basically learns from the experiences. It is a fact that some problems which are beyond the scope of current computers can be are easily solvable by energy efficient packages. Such type if brain modelling also provides a less technical path for the development of machine solution.

ANN are computer having their architecture modelled after the brain. They mainly involve hundreds of simple processing units wired together in complex communication network. Each simple processing unit represents a real neuron which sends off a new signal or fires if it receives a strong signal from the other connected unit [13].

3.2 *Artificial Neuron*

Artificial Neurons are the basic unit of Artificial Neural Network which simulates the four basic function of biological neuron. It is a mathematical function conceived as a model of natural neuron. The following figure shows the basic artificial neuron.

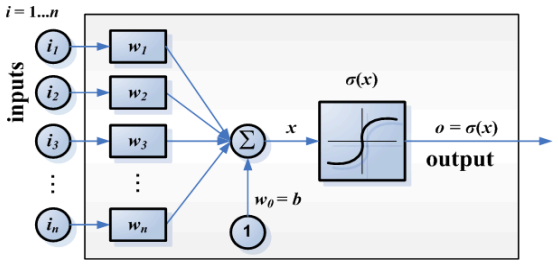

Fig. 2: Basic Artificial Neuron

In this figure, various inputs are shown by the mathematical symbol, i(n). Each of this inputs are multiplied by connecting

weights w(n). Generally, this products are simply summed and fed to the transfer function to generate the output results. The applications like text recognition and speech recognition are required to turn these real world inputs into discrete values. These applications don't always utilize networks composed of neurons that simply sum, and thereby smooth, inputs. In the software packages, these neurons are called as processing elements and have many more capabilities than the basic artificial neuron described above.

IV. Types of Artificial Neural Network

Researchers from the world have found out countless different structure of Artificial Neural Network. Short description of each is given below.

4.1 Feedforward Network

Feedforward network is the first and the simplest form of ANN. In this network, the information flows only in one i.e. forward direction from input node via hidden nodes to the output node. This network contains no loops or cycles. A neuron in layer 'a' can only send data to neuron in layer 'b' if b > a. Learning is the adaptation of free parameters of neural network through a continuous process of stimulation by the embedded environment. Learning with teacher is called as (a) supervised training; and learning without teacher is called as (b) unsupervised training. The back-propagation algorithm has emerged to design the new class of layered feedforward network called as Multi-Layer Perceptrons (MLP). It generally contains at least two layers of perceptrons. It has one input layer, one or more hidden layers and output layers. The hidden layer plays very important role and acts as a feature extractor. It uses a nonlinear function such as sigmoid or a radial-basis to generate complex functions of input. To minimize classification error, the output layer acts as a logical net which chooses an index to send to the output on the basis of input it receives from the hidden layer [14].

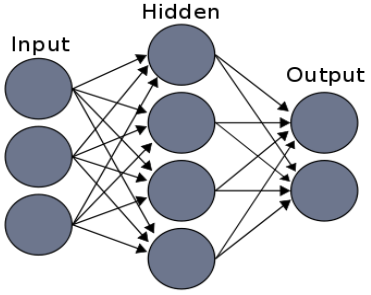

Fig. 3: A Fully Connected Feedforward With One Hidden Layer And One Output Layer.

4.2 Recurrent Neural Network

A Recurrent Neural Network (RNN) is a neural network that operates in time. RNN accepts an input vector, updates its hidden state via non-linear activation function and uses it to make prediction on output. In this network, the output of the neuron is multiplied by a weight and fed back to the inputs of neuron with delay. RNN have achieved better speech recognition rates than MLP, but the training algorithm is again more complex and dynamically sensitive, which can cause problems [15].

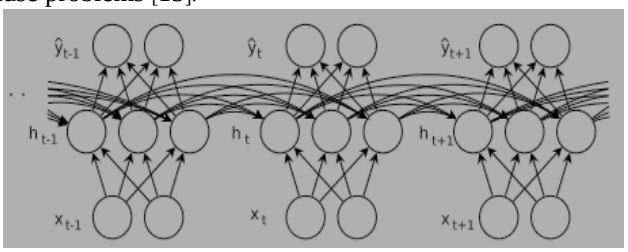

Fig. 4: Structure of A Recurrent Neural Network.

4.3 Modular Neural Network

A Modular Neural Network (MNN) consist of several modules, each module carrying out one sub task of the neural network's global task, and all module functionally embedded. The global task can be any NN application, e.g., mapping, clustering, function approximation or associative memory application [16].

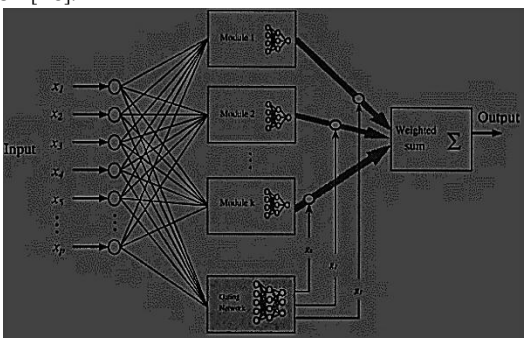

Fig. 5: Module Neural Network Architecture.

4.4 Kohonen Self Organizing Maps

Kohonen self-organizing maps are a type of neural network. They require no supervision and hence called as "Self-organizing". They learn on their own unsupervised competitive learning. They are called as "Maps" because they attempt to map their weight to conform to the given input data [17].

V. Advantages of Artificial Neural Network

- ANN have the ability to learn how to do task based on the data given for training, learning and initial experience.
- ANN can create their own organisation and require no supervision as they can learn on their own unsupervised competitive learning.
- Computations of ANN can be carried out in parallel.
- ANN can be used in pattern recognition which is a powerful technique for harnessing the data and generalizing about it.

- The development of system is through learning instead of programming.
- ANN are flexible in changing environments.
- ANN can build informative model when conventional model fails. They can handle very complex interactions.
- ANN is a nonlinear model which is easy to use and understand than statistical methods.

VI. Limitations of Artificial Neural Network

- It is not a daily life problem solving approach.
- No structured methodology is available in ANN.
- ANN may give unpredictable output quality.
- Problem solving methodology of many ANN system is not described.
- Black box nature.
- Empirical nature for model development.

VII. Conclusion & Future Scope

ANN are one of the promises for the future computing. This paper shows that they can be very useful in speech signal classification. They operate more similarly to human brain than a conventional computer logic. Different types of ANN are shortly discussed in this paper and it can be concluded that RNN have achieved better speech recognition rates than MLP, but the training algorithm is again more complex and dynamically sensitive, which can cause problems. Speech recognition has attracted many scientists and has created technological influence on society. Hope this paper brings out the basic understanding of ANN and inspire the research group working on Automatic Speech Recognition. The future of this technology is very promising and the whole key lies in hardware development as ANN need faster hardware.

References

[1] Wouter Gevaert, Georgi Tsenov, Valeri Mladenov, "Neural Network used for Speech Recognition", Journal of Automatic Control, University of Belgrade, Vol. 20, pp. 1-7, 2010. http://dx.doi.org/10.2298/JAC1001001G

[2] Vimal Krishnan VR, Athulya Jayakumar, Babu Anto P, "Speech Recognition of Isolated Malyalam Words Using Wavelet Feature and Artificial Neural Networks", 4th IEEE International Symposium on Electronic Design, Test and Application, 2008. http://dx.doi.org/10.1109/DELTA.2008.88

[3] Ganesh Tiwari,"Text Prompted Remote Speaker Authentication: Joint Speech & Speaker Recognition/Verification System.

[4] http://www.guidogybels.eu/asrp4.html

[5] Yashwanth H, Harish Mahendrakar and Suman Davia, " Automatic Speech recognition Using Audio Visual Cues", IEEE India Annual Conferencec pp. 166-169, 2004.

[6] G. Saha, Sandipan Chakroborty, Suman Senapati, "A New Silence Removal and Endpoint Deletion Algorithm for Speech and Speaker Recognition Applications.

[7] Urmila Shrawankar, Dr. Vilas Thakare, "Techniques for Feature Extraction in Speech Recognition System: A Comparative Study.

[8] Lei Xie, Zhi-Qiang Liu, "A Comparative Study of Audio Feature for Audio Visual Conversion in MPEG-4 Compliant Facial Animation", Proc. of ICMLC Dalian, 13-16, August 2006. http://dx.doi.org/10.1109/icmlc.2006.259085

[9] Honig, Florian Stemmer, George Hacker, Christian Brugnara, Fabio, "Revising Perceptual Linear Prediction", In interspeech – 2005, pp. 2997-3000.

[10] Santosh K. Gaikwad, Bharti W. Gawali, Pravin Yennawar, "A Review on Speech Recognition Techniques", IJCA Vol. 10, No. 3, pp. 16-24, November 2010. http://dx.doi.org/10.5120/1462-1976

[11] Santosh K. Gaikwad, Bharti W. Gawali, Pravin Yennawar, "A Review on Speech Recognition Techniques", IJCA Vol. 10, No. 3, pp. 16-24, November 2010 http://dx.doi.org/10.5120/1462-1976.

[12] Lindasalva Muda, "Voice Recognition Algorithm Using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", Journal of Computing, Vol. 2, Issue 3, March 2010.

[13] Singh Satyanand, Dr. E. G. Rajan, "Vector Quantization Using MFCC and Inverted MFCC", International Journal of Computer Applications, Vol. 17, No. 1, pp. 1-7, March 2011.

[14] Sonali B. Maind, Priyanka Wankar, "Research Paper on Basic of Artificial Neural Network ", International Journal on Recent & Innovation Trends in Computing & Communication, Vol. 1, Issue 1, pp. 96-100.

[15] Robison, A. J. Cook, G. D. Ellis, D. P. W. Fosteruissier, E., Renals, S. J., Williams, D. A. G., "Connectionist Speech Recognition of Broadcast News", Speech Commnication 37: 27-45, 2000.

[16] James Martens, Ilya Sutskever, "Learning Recurrent Neural Network with Hessian-Free Optimization", University of Toronto, Canada.

[17] Gasser Auda, Mohamed Kamel, "Modular Neural Network: A Survey", International Journal of Neural System, Vol. 9, No. 2, pp. 129-151, April 1999. http://dx.doi.org/10.1142/S0129065799000125

[18] Shyam M. Guthikonda, "Kohonen Self-Optimizing Maps", Wittensberg University, December 2005.