# Modern Speech Identification Model using Acoustic Neural approach

**Mr. Ashish Uplenchwar**
*Senior Software Developer, Oakland Systems Pvt. Ltd. Nagpur India*
*uplenchwar.ashish@gmail.com*

***Article History***

***Article Submission***
*19 April 2017*
***Revised Submission***
*12 June 2017*
***Article Accepted***
*13 August 2017*
***Article Published***
*30 September 2017*

***Abstract***

*Modern day technology demands sophisticated technology to give input commands to computational devices. Prominent techniques has to be introduced to make human machine interface smooth and compatible especially speech signals. Establishing efficient communication between computer and machine plays a vital role in speech processing. This article uses one of current technologies in the Continuous Speech Recognition systems which is Reservoir Computing based Neural Network followed by likelihood conversion. Our aim is to build a stand-alone system which understands the terminology of languages. Throughout the development, measures will be taken to keep the memory requirement and the processing time of the software as small as possible.*

***Keywords:*** *Speech processing, Neural network, memory*

## I. Introduction

Speech recognition is the vital and important task in modern day engineering. The system comprises of pre-processing stage which deals with speech input and conversion of the input signal into respective waveform. In next step, HMM models which are termed as Markov models are used for decoding process. Next, acoustic models are formed which recognize the word sequence. In case of huge vocabulary recognition involving 1000 words, subordinate speech units are used with training sequences [1][2].

A modular decoding scaffold for LVCSR with search strategy is preferred. WFSMs are employed to achieve acoustic mappings from speech attributes [3]. The speech knowledge is programmed through phone lattice and probabilistic approach. The final output is obtained through lexical and syntactical approaches. The final decoded sentence is obtained by performing lexical access and applying syntactical constraints. The Speech Recognition Systems using Bottom Up approach is shown in Fig. 1. The proposed system allows us to get speech recognition at all phases such as word, attribute and phone levels [4].

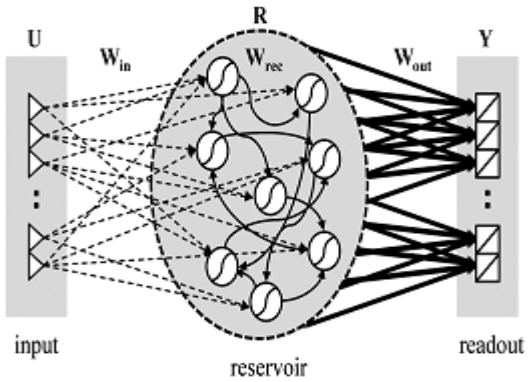


*Fig 1: Reservoir Network based LVCSR Model*

ISSN: 2250-0839

Acoustic Modeling using Hierarchical Reservoirs is used in continuous speech recognizes. The AM model differentiates observable and non-observable speech signals given at the input [5][6]. The HMM models are combined with Gaussian models for greater efficiency. In this work, we investigate how Continuous Speech Recognition based RC networks can yield good Recognition accuracy. In what follows, we first analyze the HMM based hybrid Architectures and applied to Neural Networks. Based on gained knowledge we conclude it with a future idea [7][8].

## II. PROPOSED NEURAL NETWORK BASED SPEECH RECOGNITION SYSTEM

The detailed analysis of the Large Vocabulary based Continuous Speech Recognition Systems using Neural Networks could be made by the following block diagram. The implementation begins with preprocessing and then the training. From the above figure it is very clear that RC-HMM based Acoustic modeling is a technology which could be used only after the front end processing. So, it is an intermediate process which could design in such a way that it must provide an efficient speech recognition strategy for the real time applications [9].

The front end processing involves three processes namely; Frame blocking, Windowing and FFT. In this project, some of the speaker's commands are recorded and then fed as an input to our RC-HMM Model. In order to perform the front end processing, we have taken the data samples of about 60 voice commands. All the voices are recorded at the frequency of 8000Hz through Microphone. In total voice data 20 data are voice samples of command GEAR DOWN and similarly for GEAR UP and LAUNCH THE MISSILE, in which 15 of them are used for training from each command and 5 voice samples are used for testing the accuracy. Since Speech is a nonlinear data, the length of each sample differs from one another. So length has to be made equal in dimension before starting the process and is shown in figure 2.

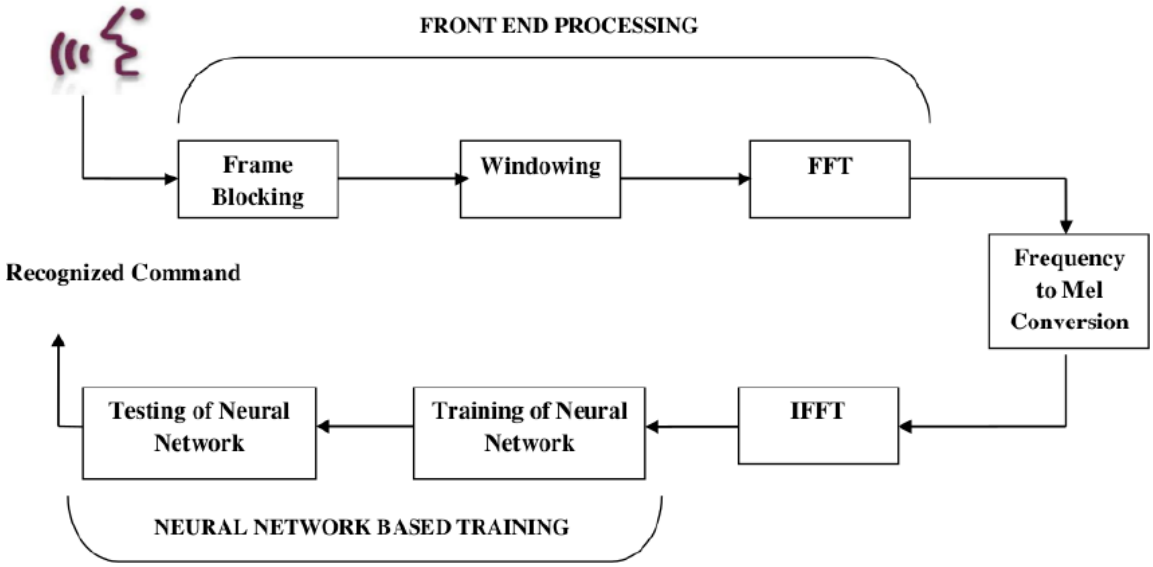


*Fig 2: proposed neural network based speech recognition system*

The following table illustrates the comparative study of different existing systems which in turn pictures the best suitable methodology for future work. The following table illustrates the comparative study of different from the above comparison,[5] it is very clear that Reservoir based LVCSR system has great advantage than the other two existing methodologies. Since Word Error Rate is found minimum, it is a best suited approach for large terminology based speech appreciation systems. But the Neural network based LVCSR models provide better performance in Voice Recognition systems than the generic methods and unidirectional approaches but still provides a lower Word Error Rates [10].

*TABLE 1: Comparison of various speech recognition systems*

| Analyzing Parameters | Advances in LVCSR | Bottom- Up Approach of LVCSR | Reservoir Network based LVCSR |
|---|---|---|---|
| **Recognition Technique** | Continuous Speech Recognition | Continuous Speech Recognition | Continuous Speech Recognition |
| **Acoustic Modeling** | Not Specific | Hidden Markov Model (HMM) | Hidden Markov Model (HMM) |
| **Technology Used** | Speaker Adaptation Technique | Speaker Adaptation + Modular Search | Speaker Adaptation + Reservoir Computing |
| **Adopted Method** | Generic method | Artificial Neural Networks | Recurrent Neural Networks |
| **System Nature** | Speaker Independent and dependent | Speaker Independent | Speaker Independent |
| **Word Error Rate** | Above 25% | 13.3% | 6.2% |

## III. ANALYSIS AND SIMULATION RESULTS

The typical speech signal for the command GEAR DOWN with duration of 8100ms by a MALE Speaker is shown in the following Fig 3 which was generated using MATLAB.

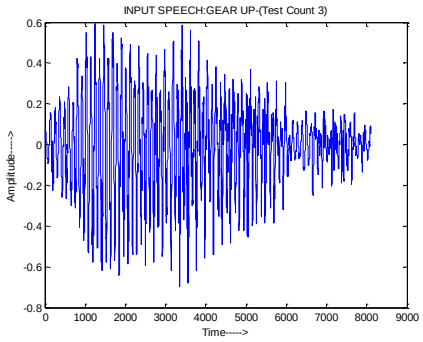


*Fig 3: Speech Signal: GEAR UP*

The digitized speech signal is processed by a first-order digital network in order to spectrally flatten the signal. Sections of NA consecutive speech samples are used as a single frame. Consecutive frames are spaced MA samples apart. The Pre-emphasized signal is now applied for frame blocking and the resultant signal is shown in the following Fig 4.

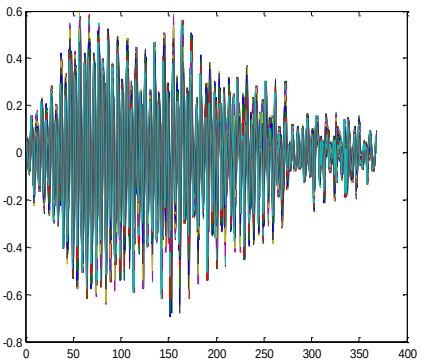


*Fig 4: Frame Blocked Speech Signal: GEAR UP.*

ISSN: 2250-0839

In the process of frame blocking, we have fixed Original Frame Size: 32ms and Consecutive Frame Size: 22ms.In order to calculate the Total No of frames in a given speech duration, we have formula of :

Total No of frames = (8100-32)/22 =366.72~ 367frames

Window used: Hamming Window and it is the process of Convolution of Framed Speech with the Window function.

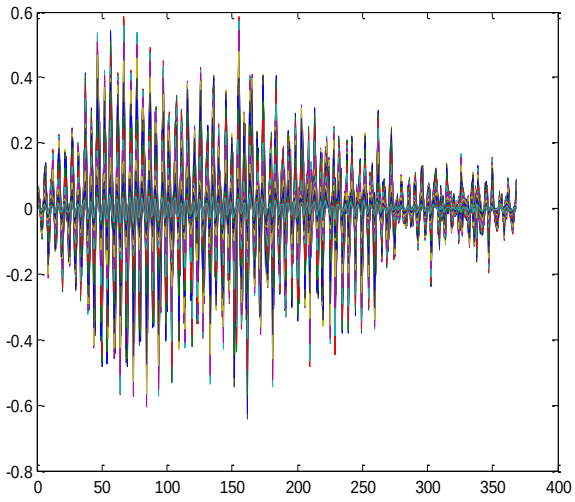


*Fig 5: Frame Blocked Speech Signal: GEAR UP after HMM*

Once the signal is converted into numerous frames, the next step would be windowing. Using the Hamming window when it is convolved with the framed signal, a clear range of a signal for a particular period could be analyzed. It helps to determine the energy of the particular speech command. The following figure shows the Windowed signal for the Command GEAR UP: MALE.

$$D_k = \sum_{m=0}^{N_m - 1} D_m \, e^{\frac{-j2\pi km}{N_m}} \quad (1)$$

Where k= 0, 1, 2 ….. Nm-1

On another, in case of FFT this frame will be divided into small

The following neural network is taken as the training network and it is shown below in the Fig.7. The type of neural network used here is a Custom Neural Network and it has taken 100 coefficients of original training data samples as its input. The output is recognized commands either GEAR UP or GEAR DOWN or LAUNCH THE MISSILE. It has two layers with soft threshold and output layer of 3. The layers can be changed accordingly which differs according to the application which the user is interested. If one has to understand the model connections and other parameters of a network in a better manner, the usage of lower level layers are advisable. Fixing of epochs are very important in training of speech samples. No of epochs differs for different training periods. In this case, we have fixed the epoch value as 50.

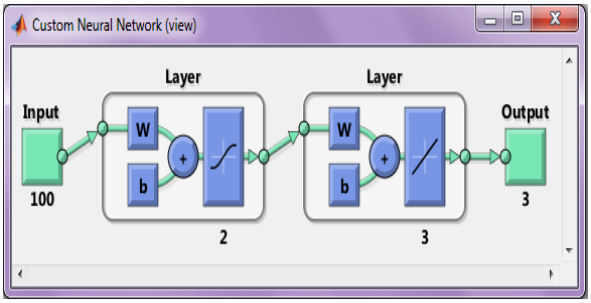


*Fig 6: Overview of Neural Network*

Targets are assigned based on the input training data. Once the target is fixed, the network has to train to meet the target. The target for the command Gear Up is fixed. This is because the Gear up commands are trained with the above target vectors with the following specifications. After fixing the neural network design and targets, we are ready to do training of the Voice Samples which are taken for the analysis and it is shown in the following Fig.7.From the above simulated result the performance of the training Neural network has been converged at 7 iterations (epochs) with the MSE (Mean Square Error) value of 1.37*10-15. In this case we have taken the 5 different sets of Gear down Command and trained with the size of inputs 100*5=500 input Coefficients.

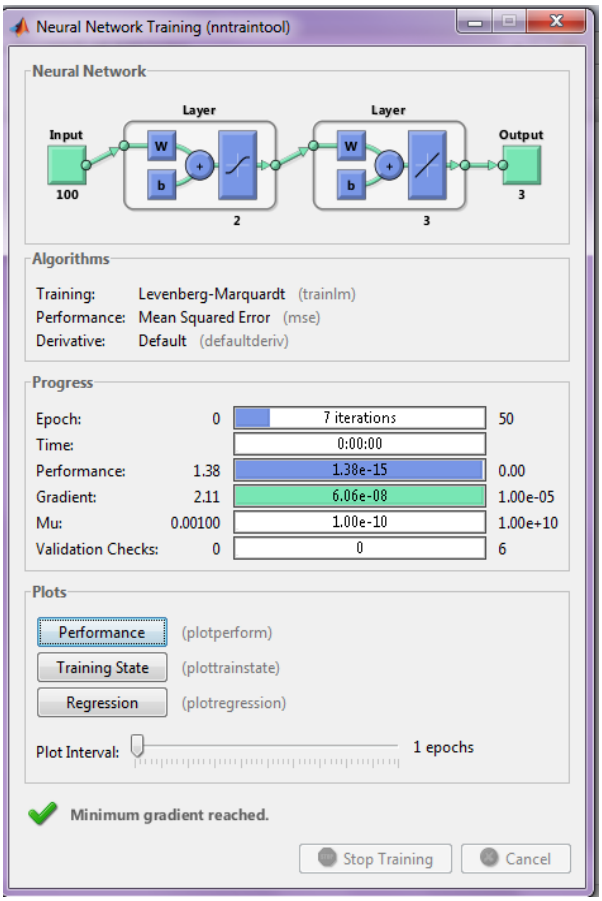


*Fig 7: Training of Neural Network*

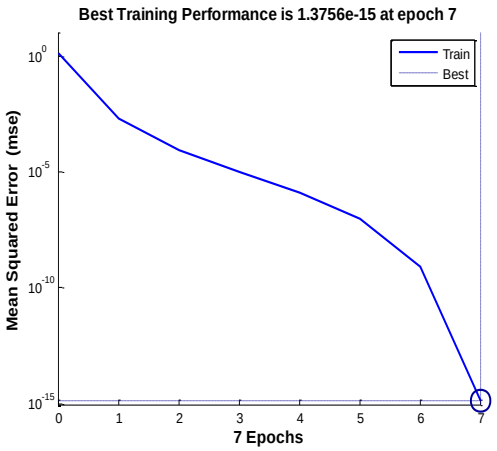


*Fig 8: Performance Estimation of Trainingof Neural Network*

After the completion of the training of neural network, one has to test the Recognition accuracy of network by providing test data which are not used for previous training process. While testing the data with an untrained data set which includes all the three commands, the recognition accuracy gets reduced due to the less input data sample training. Here the Recognition accuracy is about 50% due to less data intake for training.

## IV. CONCLUSION

The Neural Network based Continuous Speech Recognition Systems yields better performance when compared to other generic HMM based models when large numbers of training data are used. In neural network, a large amount of data has to be handled and training time varies with respect to it. So in turn it varies the recognition accuracy of the system. From the experiments the managing the data samples which in turns increases the training time of the samples. The Performance goes down when noise get added with the input voice samples. So in future a new algorithm has to be proposed in order to overcome the above disadvantages.

## References

[1] G. Saon and J.-T. Chien, “Large-vocabulary continuous speech recog-nition systems: A look at some recent advances,” IEEE Signal Process.Mag, vol. 29, no. 6, pp. 18–33, 2013.

[2] G. Hinton, S. Osindero, and Y. Teh, “A fast learning algorithm for deep belief nets,” Neural Comput., vol. 18, no. 7, pp. 1527–1554, Jul 2006.

[3] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” IEEE Signal Process. Mag., vol. 29, no. 6, pp. 82–97, Nov. 2012.

[4] L. Deng, G. Hinton, and B. Kingsbury, “New types of deep neural network learning for speech recognition and related applications: An overview,” in Proc. ICASSP, 2013, pp. 8599–8603.

[5] T. N. Sainath, B. Kingsbury, B. Ramabhadran, P. Fousek, P. Novák, and A. rahman Mohamed, “Making deep belief networks effective for large vocabulary continuous speech recognition,” in Proc. ASRU Work-shop, 2011, pp. 30–35.

[6] A. Mohamed, G. Dahl, and G. Hinton, “Acoustic modeling using deep belief networks,” IEEE Trans. Audio, Speech, Lang. Process., vol. 20, no. 1, pp. 14–22, Jan 2012.

[7] B. Hutchinson, L. Deng, and D. Yu, “A deep architecture with bilinear modeling of hidden representations: Applications to phonetic recogni-tion,” in Proc. ICASSP, 2012, pp. 4805–4808.

[8] V. V. V. Raju, P. Gangamohan, S. V. Gangashetty and A. k. Vuppala, "Application of prosody modification for Speech Recognition in different Emotion conditions," 2016 IEEE Region 10 Conference (TENCON), Singapore, 2016, pp. 951-954, doi: 10.1109/TENCON.2016.7848145.

[9] J. Kim and I. Lane, "Accelerating multi-user large vocabulary continuous speech recognition on heterogeneous CPU-GPU platforms," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, 2016, pp. 5330-5334, doi: 10.1109/ICASSP.2016.7472695.

[10] A. Prodeus and K. Kukharicheva, "Training of automatic speech recognition system on noised speech," 2016 4th International Conference on Methods and Systems of Navigation and Motion Control (MSNMC), Kiev, 2016, pp. 221-223, doi: 10.1109/MSNMC.2016.7783147.