A Comparative Study on Ann and Hmm Based Automatic Speech Recognition Systems for Controlling Micro Air Vehicles

[1] Pragathi G, [2]Veena S, [3]Roopa S
[1] Pursuing master's degree program in Signal Processing in Electronics and Communication Engineering in Siddaganga Institute of Technology, Tumakuru, Karnataka
[2] Principal Scientist in Aerospace and Systems Division CSIR-NAL,
[3] Assistant Professor in Department of Electronics and Communication Engineering in Siddaganga Institute of Technology, Tumakuru, Karnataka
[1] pragathi7124@gmail.com [2] veenas@nal.res.in, [3] roopa01@gmail.com

***Abstract*- Speech is one of the effective modes of communication and when made to be recognized by a computer, it can be used in many different areas of application. This paper makes a comparison between Hidden Markov Model Artificial Neural Networks (ANN) used in controlling of Micro Air Vehicle (MAV) based on speech-activated commands from Ground Control Station (GCS). Therefore, Automatic Speech Recognition (ASR) Systems are developed based on ANN and HMM separately and the recognition accuracies obtained in both the cases are validated against each other.**

***Index Terms*—Artificial Neural Networks, Automatic Speech Recognition, Hidden Markov Model, Micro Air Vehicles**

I. INTRODUCTION

Micro Air Vehicles (MAVs) are a smaller class of Unmanned Aerial Vehicle (UAV) and are used in variety of missions like surveillance in hazardous conditions. Micro Air Vehicles are the new development of technology by which many operations can be done. The MAV operations are controlled by Ground Control Station software. The GCS software is being installed on a computer helps in planning and control of MAV operations. These commands are sent to MAV through a wireless link. Speech being a convenient means of interaction between a human and a machine automatic speech recognition is the key for Human-to-Machine communication technology[1,2] .

The different training methods used for automatic speech recognition are artificial neural networks[3], Hidden Markov Model[4]. The hidden Markov model shows only 0-40% recognition accuracy in few small vocabulary words. To increase this recognition accuracy an artificial neural network is considered in this paper.

II. AUTOMATIC SPEECH RECOGNITION

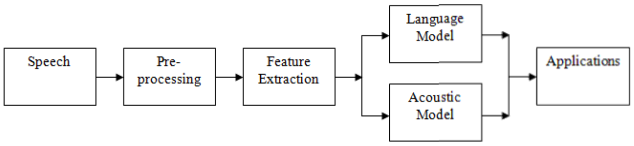

Fig. 2.1: A typical Speech Recognition System

Fig. 2.1 shows the architecture of a typical speech recognition system. Speech signal represents to the original speech used for both training and testing is recorded and stored. The pre-processing stage includes Pre-emphasis and Voice-activity Detection which makes the speech more compact and efficient. The extraction of Mel Frequency Cepstral Coefficients (MFCCs) is done in Feature Extraction stage. The acoustic model represents the speech in terms of gender, acoustics, phonetics and other related parameters. The language model represents the semantics, syntax and pragmatic aspects related to speech.

In Automatic Speech Recognition(ASR) systems, an individual speaker reads the isolated vocabularies or text into the system. The system analyzes a specific voice and uses this to recognize the person's speech. Applications of automatic routing, voice dialing, speech-to-text processing and aircraft.

A. *Acoustic Modeling using HMM*

HMM is a statistical tool used for characterization of spectral features of a speech frame. HMM makes an assumption that a speech signal can be characterized as a parametric random process, and these parameters of a stochastic process can be well predicted[7].

B. Acoustic Modeling using ANN

Artificial Neural Networks (ANNs) are used with the desire to achieve human-like performance in the application of ASR. An ANN model is composed of many non-linear elements working in parallel similar to a biological neuron[5]. Non linearity and fault tolerance are the two main characteristics of ANNs used in solving a speech recognition problem [6].

Multilayer Perceptions (MLPs) are an important class of neural networks which are feed-forward networks with one or more hidden layers[5]. The hidden layers form a group of layers between input and output layers

III. EXPERIMENTAL FRAMEWORK

In this section the two different systems designed for ASR are presented in detail. Same data sets are used for training and testing of both the systems. The steps for ASR are same for both ANN and HMM systems, except that acoustic modeling are different.

The 12 flight command which makes up the database are:

Actions, Config, Donate, Gauges, Help, Messages, Quick, Scripts, Servo, Simulation, Status and Terminal. The speech signal used for training and testing were recorded and stored. A database of each word is created with many utterances of the same word by unspecified speakers since it is speaker-independent mode. The pre-emphasizer is a first order high pass FIR filter. This filter is used to emphasize the high frequency components and this brings in spectral flattening. The main purpose of Voice Activity Detection (VAD) is to select a signal above a threshold level as speech. Frame blocking divides the signal into many number of frames. Each frame has an appropriate time-length. After the frame blocking, a Hamming window is applied to ever other frame which reduces the discontinuity of signal at the end of each block. Since the human perception is non-linear, there is a need for a mapping scale and thus Mel scale is used. The triangular Mel scale filter bank warps the frequency in Hz to frequency in Mel scale. The number of filter taps considered in Mel scale filter bank is 26. Hence the output from Mel filter bank will be a 26 sample long vector. Out of these 26 samples, only first 13 are considered in the DCT signal, and again the first bank is ignored. Thus for a single frame, 12

MFCC features are extracted. The same is repeated for every other frame resulting in a set of MFCC feature matrix of dimension (12*number of frames) for a single utterance.

A. ANN System Overview

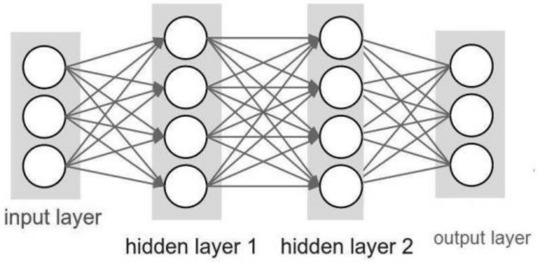

Fig. 3.1: Architecture of a general back propagation feed forward neural network

A neural network based ASR system is designed to carry out this research. The system is partitioned into different modules according to their functionalities. The neural network used for the purpose is Back Propagation Feed Forward Neural Network (BPFFNN)[8]. The back propagation algorithm assumes feed-forward network architecture.

Fig 3.1 shows an architecture of a general back propagation feed forward neural network. In this architecture, the nodes are partitioned into L layers. The input layer being the lowermost layer and the output layer being the topmost layer or layer L. The layers between input layer and output layer are hidden layers. The number of nodes present in input, hidden and output layers depends on the complexity of the problem for which the neural network is used. The initial weights in the network are taken to be random values.

The operation of a back propagation neural network has two phases: feed-forward and back propagation phases. In feed forward phase, the outputs at all the nodes are calculated by applying sigmoid function to input at each node. This output values obtained are compared with the desired output values. This difference between the obtained actual output values and desired output values is the error. In back propagation phase, this error is propagated from output layer to hidden layers and then from hidden layers to input layer and thus the weights are updated. Fig 3.2 shows a ASR system based on ANN and Table I describes the parameters used for the ANN based ASR system.

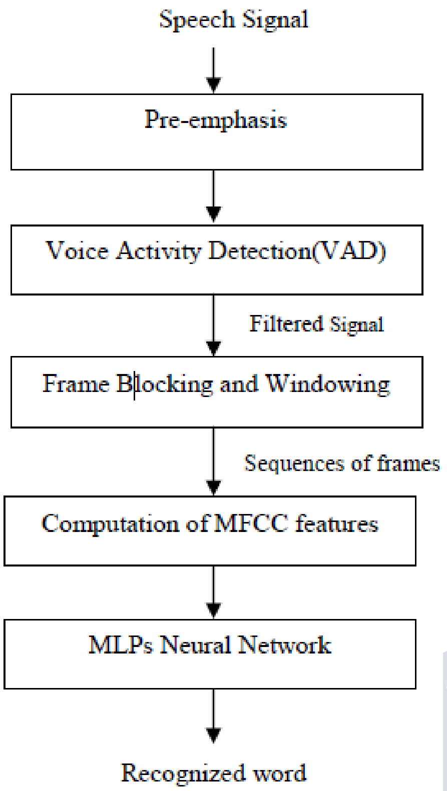

Fig. 3.2: Block Diagram of ANN based System

Table I: System parameters for ANN based ASR system

Parameter	Value
Sampling rate	16KHz
Database	Isolated 12 Flight Commands
Speakers	10
Repetitions	5
Pre-emphasis filter	[1,0.95]
Window size and type	512, Hamming
MFCC filter bank	26 triangular filters
MFCC coefficients	13
Number of input nodes	260 nodes (13 MFCC coefficients×20 frames)
Number of hidden and output nodes	50 nodes in first hidden layer, 20 in second hidden layer and 12 output nodes

B. HMM System Overview

A HMM based ASR system is also designed. According to the functionalities, the system was partitioned into different modules. A HMM model represents the utterances in a compact manner.

HMM is represented by λ=(π, A, B), where:

π = Initial state probability matrix,

A = State transition probability matrix,

B = Continuous observation probability density function matrix.

The number of states considered in HMM model is 5.

Hence matrix A is of size (5×5) with the probability of transition to other state is 0.15 and probability of transition to same state is 0.85. π is a matrix of size (1×5) with probability of state one to be 1 and probabilities of other states to be 0.

The size of matrix B is (number of frames×5). The matrix B is initialized using the feature matrix.

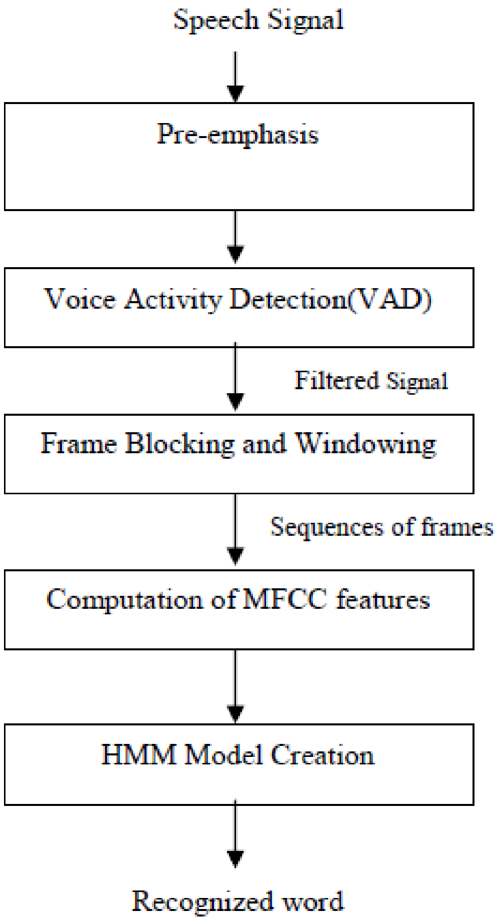

Fig. 3.2: Block Diagram of HMM based System

Table II: System parameters for HMM based ASR system

Parameter	Value
Sampling rate	16KHz
Database	12 Flight Commands
Speakers	10
Repetitions	5
Pre-emphasis filter	[1,0.95]
Window size and type	512,Hamming
MFCC filter bank	26 triangular filters
MFCC coefficients	13
Number of HMM states considered	5 states

IV. RESULTS

To carry out the research, seven speakers (speaker one to seven) were used for training purpose. Hence the total samples dedicated to training phase is 420 samples (7 speakers × 5 utterances or repetitions × 12 flight commands). The testing set consists of utterances of speaker eight to ten with 5 repetitions and 12 flight commands. Hence a total of 180 samples are used for testing. The same data sets were used for both the ANN and HMM based systems

$$\%\ Recognition\ accuracy = \frac{Number\ of\ words\ correctly\ recognized}{Total\ number\ of\ words\ used\ for\ testing}$$

Table III. Obtained Recognition results

Word Model	% Recognition accuracy for ANN based system	% Recognition accuracy for HMM based system
Actions	100%	100%
Config	93.33%	86.67%
Donate	100%	100%
Gauges	86.67%	93.33%
Help	93.33%	60%
Messages	100%	100%
Quick	86.67%	40%
Scripts	86.67%	93.33%
Servo	93.33%	100%
Simulation	100%	100%
Status	86.67%	100%
Terminal	100%	86.67%

From the results, it can be deduced that though the HMM model works accurately for few words. Even though the number of states is increased from 5 to 10 in HMM model, the percentage accuracy remains almost same for the small utterances in this model. This issue is addressed in ASR model. For the words with small vocabulary (eg., Help and Quick), HMM model shows lower accuracy, whereas, ANN model works more accurately for all the words. The average accuracies of ANN model and HMM model is obtained to be 93.89% and 88.33% respectively. Therefore, we can say that the accuracy of the ASR model is better in ANN case than compared to HMM case in most of the utterances

V. CONCLUSION

Two Automatic Speech Recognizers namely, ANN based and HMM based were designed and the process of automatic speech recognition was investigated. The performances of these two systems were compared. It is found that ANN based recognizer performs better than HMM based system. Therefore, we may conclude that the ANN approach is better than HMM approach in controlling Micro Air Vehicles using Automatic Speech Recognition due to the simplicity of those recognizers.

REFERENCES

[1] Mark Draper, Gloria Calhoun, Heath Ruff, David Williamson and Timothy Barry, "Manual versus Speech input for Unmanned Aerial Vehicle Control Station operation," Proceedings of the Human Factors Ergonomics Society's 47th Annual meeting, pp. 109-113, October 2003.

[2] B. H. Juang and S. Furui, "Automatic recognition and understanding of spoken language-a first step toward natural human-machine communication," Proceedings of the IEEE, vol. 88, pp. 1142-1165, 2000.

[3] Austin Marshall, "Artificial Neural Network for Speech Recognition," 2nd Annual Student Research Showcase, March 3, 2005.

[4] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE, vol.77, no.2, pp.257-286, February 1989.

[5] Lippmann R., "Review of neural networks for Speech Recognition," Neural Computation, pp.1-38, MIT press, 1989.

[6] Haykin, S., "Neural Networks: A Comprehensive Foundation,"Second Edition, Prentice Hall 1999.

[7] Loizou P. C. and Spanias A. S., "High-Performance AlphabetRecognition," IEEE Trans. on Speech and Audio Processing, pp. 430-445, 1996.

[8] Siddhant C. Joshi and Dr. A. N. Cheeran, "MATLAB Based Back-Propagation Neural Network for Automatic Speech Recognition," International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, vol.3, no.7, July 2014.