Hybrid Artificial Neural Network and Hidden Markov Model (ANN/HMM) for Speech and Speaker Recognition

Kapure Vijay Ramesh
Saraswati college of Engg.
Khaghar ,Navi-Mumbai

Sonal Gahankari
Saraswati college of Engg.
Kharghar ,Navi-Mumbai

ABSTRACT
Speech recognition is an important component of biological identification which is an integrated technology of acoustics, signal processing and artificial intelligence. Recognition systems based on hidden Markov models are effective under particular circumstances, but do suffer from some major limitations that limit applicability of ASR technology in real-world environments. Attempts were made to overcome these limitations with the adoption of artificial neural networks as an alternative paradigm for ASR, but ANNs were unsuccessful in dealing with long time sequences of speech signals. So taking the limitations and advantages of both the systems it was proposed to combine HMM and ANN within a single, hybrid architecture. The goal in hybrid systems for ASR is to take advantage from the properties of both HMM and ANNs, improving flexibility and ASR performance For Speech recognition features from speech sample are extracted & mapping is done using Artificial Neural Networks. Multilayer pattern mapping neural network, which works on the principle of back propagation algorithm is proposed. Finally Speaker Recognition is done using Hidden Markov Model (HMM). The specialty of this model is the flexible and expandable hidden layer for recognition

Keywords-Speech Recognition, Artificial Neural network, Speaker Recognition, Hidden Markov Model

1. INTRODUCTION

The goal of automatic speaker recognition is to extract, characterize and recognize the information about speaker identity to identify a speaker by his or her voice. This technology can be used as a biometric feature for verifying the identity of a person in applications like banking by telephone and voice mail. Speech recognition systems were introduced after the discovery of Alexander Graham Bell about the process of converting sound waves into electrical impulses & the first speech recognition system was developed by Davis et al.[1]. Speech is the most efficient way to train a machine or communicate with a machine. This work focuses on the objective to recognize the word or the phase spoken by human, keywords in high speed.Speaker Recognition approaches can be subdivided into two approaches: text-dependent and text-independent approaches[8]. In the first approach, the speaker is asked to utter a specific phrase pin-code, password, etc.; while in the second approach, the speaker identification engine should catch the characteristics of the uttered speech irrespective of the spoken text. Hidden Markov model (HMM) is the most popular parametric model at the acoustic level. A brief Although HMM is effective approaches to the problem of acoustic modeling in ASR, allowing for good recognition performance under many circumstances, it also suffers from some limitations. The limitations of HMM is that it does not include the way or method of integrating new features into a framework in a consistent and meaningful. Moreover HMM does not properly make use of information in state transitions & even does not provide the efficient way to model the training data in case of Splitting & Clustering. In order to overcome these limitations of HMM in late 1980s, many researchers began to use artificial neural networks (ANNs) for ASR. ANN was expected to carry out the recognition task. In spite of their ability to classify short-time acoustic phonetic units, such as individual phonemes, ANNs failed as a general framework for ASR, especially with long sequences of acoustic observations like those required in order to represent words from a dictionary or whole sentences. This is mainly due to the lack of ability to model long-term dependencies in ANNs. In order to make the Recognition more efficient &accurate, led to the idea of combining HMM and ANNs within a single, novel model, known as hybrid ANN/ HMM [7,8].

In this approach the input speech signal acquired at first. After acquisition spectrogram is computed &analyzed . Most of the electronic recording equipment has an effect of noise on the recorded sound signal. But the captured sound signals are varies speaker to speaker by age, sex, anatomic variation and emotion. this added emotions & noise has to be neutralized otherwise it makes the system unstable in recognition of speech.the signal is then neutralized which reduces the emotional effects of the speech signals with noises.

2. PROPOSED METHOD

Here we propose a methodology to identify speaker and detection of speech [4]. The Fig. 1.Demonstrates the process flow.In this approach the input speech signal is acquired at first. After acquisition spectrogram is computed & analyzed. the most common format is to carry out STFT (Short Time Fourier Transform) which is a graph with two geometric dimension as horizontal axis represents time, vertical axis is frequency & third axis represents amplitude. this amplitude of a particular frequency at a particular time represents the intensity or color of each point in the image. MFCC's are calculated which provides as an input for further processing[4,5].Mel-frequency cepstral coefficients (MFCCs) are the coefficients that collectively represent the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency[9-11].Most of the electronic recording equipment has an effect of noise on the recorded sound signal. But the captured sound signals are varies speaker to speaker by age, sex, anatomic variation and emotion. This added emotions & noise has to be neutralized otherwise it makes the system

unstable in recognition of speech.when speech recognition is carried out,the minimum noise also can weigh down the process of neural networkduring training and processing. the signal is then neutralized which reduces the emotional effects of the speech signals with noises. Normalization of signal is essential after spectrogram analysis before the recognition process actually starts. Normalization can be either peak normalization or Loudness normalization.

The algorithm is divided into five parts as follows:

- Acquisition of speech signals
- Spectrogram Analysis
- Reduction of Noise
- Normalization of signal
- Recognition of speech using ANN
- Recognition of speaker using HMM

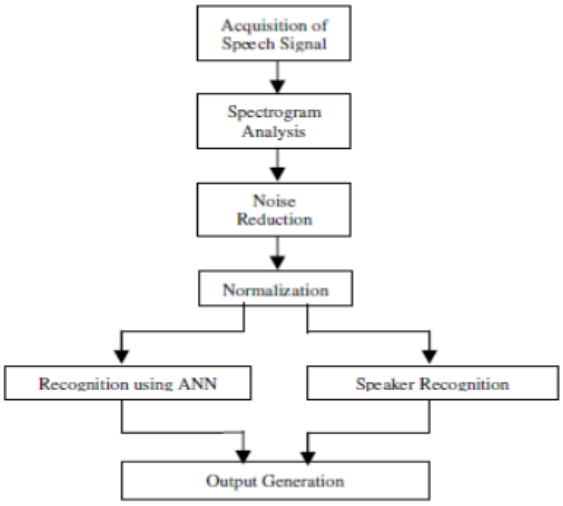

Fig 1: Process Flow

2.1 Speech Recognition Using ANN

After Normalization,the next important step is to recognize the speech using Artificial Neural Networks . In this we propose a Multilayer Mapping Network .The Fig. 2 Showsthe Multilayer Pattern Mapping Neural Network. This Multilayer Mapping network works on the principle of BackPropagation algorithm[6]. The advantage of this model is that is its flexibility & expandability of hidden layers for recognition.

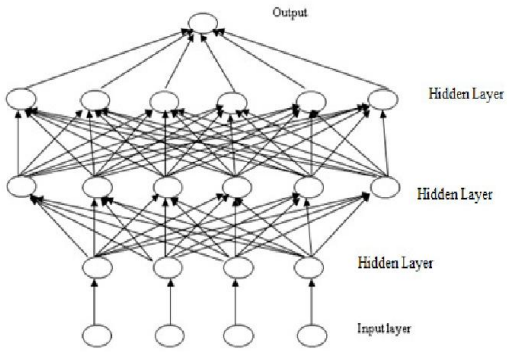

Fig 2:Multilayer Pattern Mapping Network

2.2 Speaker Recognition using HMM

Hidden Markov Model is used to recognize the speaker after speech is recognized using ANN. HMM is an statistical tool for modeling generative sequences to generate Observable sequence when characterized by an underlying process.the input to the HMM is the data rejected by ANN [4]. The calculation of cut-off value is pre-decided & the calculation with the input data will be done in the processing phase. HMM finally decides the addition & Rejection of new parameters in hidden layers. During processing,the data set will be justified with the cut-off value. If not rejected then the analysis of frequency, time, and amplitude will be performed and extraction of new features will be calculated. The Fig. 3. Shows Proposed Hidden Markov Model.

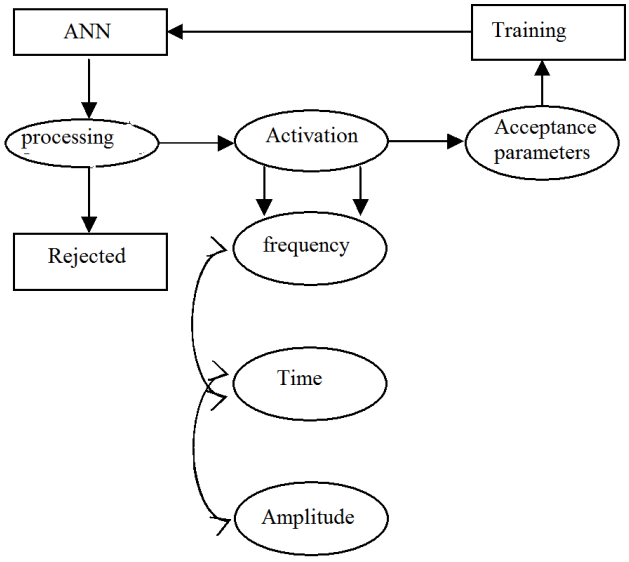

Fig 3: Proposed Hidden Markov Model

3. RESULTS & DISCUSSION

In this paper we have tried to recognize speech of users by storing the voice samples in database as well as accepting real time voice samples as an input to the system. One input can be considered at a time, this input is preprocessed & given to ANN. Sufficient no of samples for particular speaker are stored in the database. This

is done for storing samples with different pitch, emotions etc. The Fig. 4 shows the GUI model for storing samples in the database .The Fig. 5 shows the selection of a sample .once the samples are stored in database ,they are trained and features are extracted. In Recognition Phase the a single speech sample is taken as an input. The Fig. 6.shows the speech recognition GUI model. The features extracted for stored speech samples are then compared with the features of speech input sample. ANN uses multilayer mapping network which uses Back propagation algorithm for training the samples to find the best match in the comparison. During the comparison if input sample matches with the samples in the database then the system provides access to that particular speaker. The Fig. 7 shows this condition. On the contrary if input sample does not match with the samples stored in the database then access to that particular speaker will be denied by the system. The spectrogram is also plotted for the speech sample used in Recognition. The Fig. 8 shows this illustration. The same procedure is repeated for real timespeech processing by recording the samples in real time with the help of microphone. This is how speech is recognized using ANN. The data now rejected by ANN is used as an input to the HMM system.

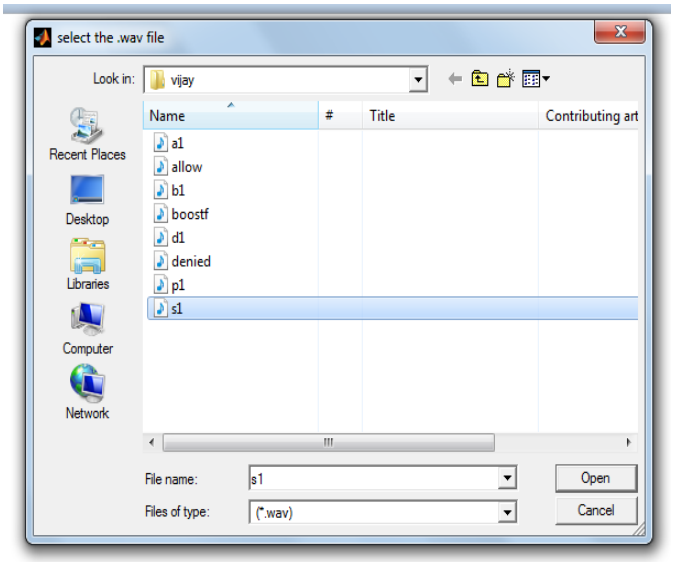

Fig 5: Selection of Speech Sample

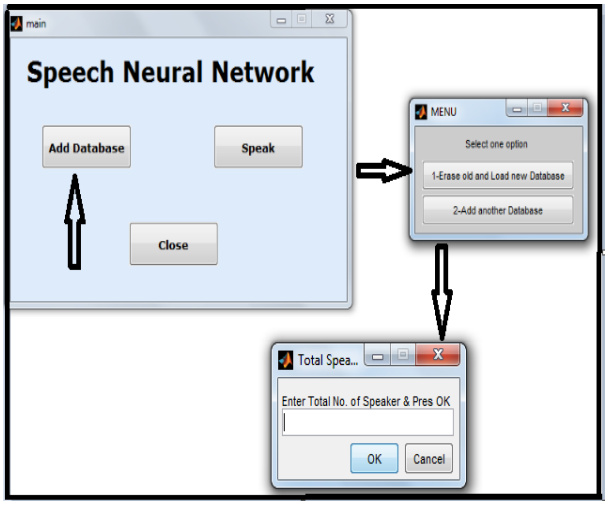

Fig 4: GUI model for Adding samples in Database

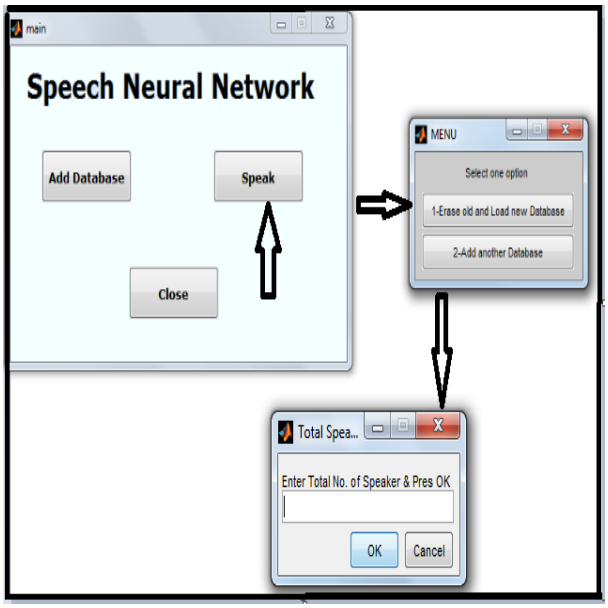

Fig 6: GUI for Speech Recognition

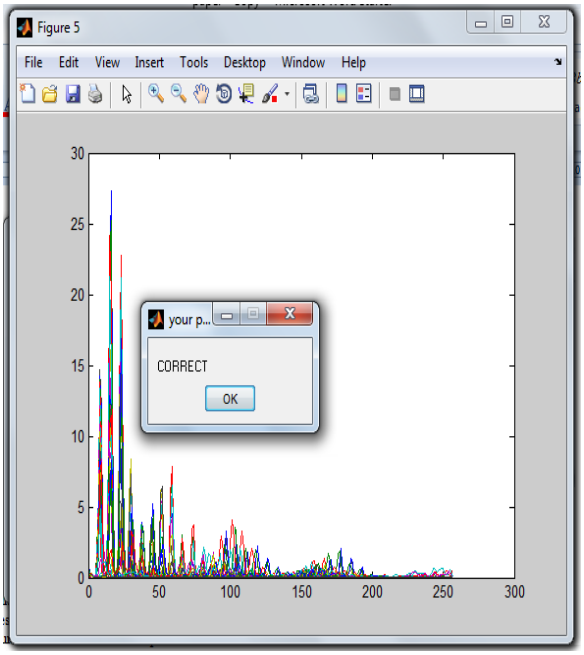

Fig 7:GUI for Access Granted

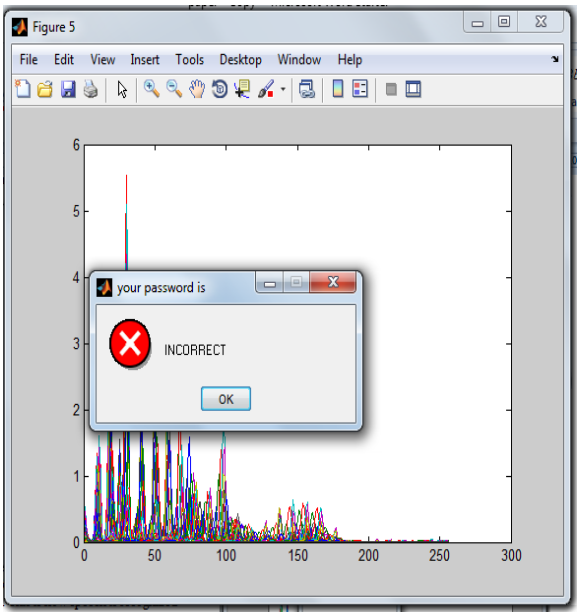

Fig 8: GUI for Access Denied

The initial steps of this paper i.e speech acquisition, preprocessing, training & recognition of speech is successfully implemented .The Accuracy observed for speech recognition using ANN is about 90–95 % for stored database samples and slightly less for real time samples .Currently the work related to this paper is focused on recognition of speaker using HMM. In Speaker Recognition, according to the algorithm shown in fig 1, the initial steps which were carried out during speech recognition need not to be carried out. During processing, the data set will be justified with the cut-off value. If not rejected then the analysis of frequency, time, and amplitude will be performed and extraction of new features will be calculated. & neural network will be retrained and expanded.

4. CONCLUSION

In this paper we have tried to apply Hybrid ANN/HMM models for speech and speaker Recognition respectively. The work was initially focused on speech acquisition, Spectrogram analysis, Neutralization, Normalization, Features Extraction and Mapping using Artificial Neural Network. Moreover additionally Speaker Recognition is also done with the help of Hidden Markov model which generates additional features for Recognition.

5. REFERENCES

[1] Davis K.,Biddulp,R., and Balashek, S.," Automatic Recognition of Spoken Digit," J. Acoust.Soc .Am.24: Nov 1952,p.637

[2] H.F. Ong and A.M. Ahmad " Malay Language Speech Recogniser with Hybrid Hidden MarkovModel & Artificial Neural Network(HMM/ANN)" InternationalJournal of Information and Education Technology, Vol.1, No.2, June 2011.

[3] NiladriSekharDey, RamakantaMohanty,K.L. Chugh"Speech & Speaker Recognition System using Artificial Neural Networks and Hidden Markov Model"2012 International Conference on Communication System and Network Technologies.

[4] Surendra P. Ramteke, GunjalOza, Nilima P.Patil "Development of TTS for Marathi Speech Signal Based on Prosody and Concatenation Approach" International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 10, December- 2012

[5] Li Bo, Wang Dong-xia, Zou De-jun, Hu Tie-sen"On Speech Recognition Access Control System Based on HMMI ANN"

[6] R. Rojas " Neural Networks" Springer Verlag, Berlin,1996

[7] "Hybrid Hidden Markov Model and Artificial Neural Network for Automatic Speech Recognition" Xian Tang , 2009 Pacific-Asia Conference on Circuits

[8] A Comparative Study to Evaluate a Text-Independent Speaker Identification Engine for Arabic Speakers Using a CHMM-Based Approach

[9] 'Voice command recognition based on MFCC and DTW' Anjali Bala et al. / International Journal of Engineering Science and Technology Vol. 2 (12), 2010, 7335-7342

[10] "A Comparative Study of Filter Bank Spacing for Speech Recognition" Ben J. Shannon, Kuldip K. Paliwal Microelectronics Engineering Research Conference 2003

[11] "Speech Recognition using MFCC" Chadawan Ittichaichareon, SiwatSuksri and Thaweesak YingthawornsukInternational Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012) July 28-29, 2012 Pattaya (Thailand)

[12] "Speech recognition using Back Propagation Algorithm" 1991 TECNON IEEE Region 10 International Conference on Energy ,Comminication and Control.