International Journal of Innovations in Engineering and Technology (IJIET)
http://dx.doi.org/10.21172/ijiet.122.12

Speech Recognition using AANN

R.Thiruvengatanadhan[1]
[1]*Department of Computer Science and Engineering*
Annamalai University, Annamalainanar, Tamilnadu, India

Abstract- Human beings communicate in either of the two ways, through speech or by writing. The most common mode of communication is speech, because it doesn't take much time to transmit information from one source to receptor. This paper describes a technique that uses Autoassociative Neural Network (AANN) to recognized speech based on features using Mel Frequency Cepstral Coefficients (MFCC). Modeling techniques such as AANN were used to model each individual word which is trained to the system. Each isolated word Segment using Voice Activity Detection (VAD) from the test sentence is matched against these models for finding the semantic representation of the test input speech. Experimental results of AANN shows good performance in recognized rate.
Keywords – Speech Recognition, VAD, MFCC, AANN

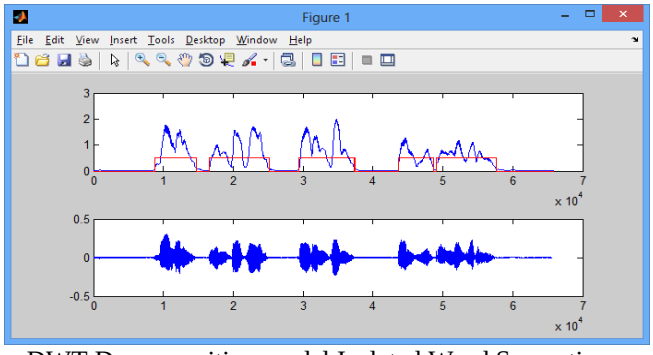
I. INTRODUCTION

Speech signal is produced as a result of time varying excitation of the vocal tract system. Speech is produced because of the vibrating source of sound that is present in the larynx of humans. This vibrating source of sound is none other than the vocal cord located in the larynx [1]. The path from larynx to the lips acts as the air column and is referred to as the vocal tract. When the velum is lowered, the nasal tract is acoustically coupled to the vocal tract to produce the nasal sounds of speech. But it is a known fact that no sound can be produced without a supply of force or energy [2].

Proposed work aims to develop a system which has to convert spoken word into text using AANN modeling technique using acoustic feature namely MFCC. In this work the temporal envelop through RMS energy of the signal is derived for segregating individual words out of the continuous speeches using voice activity detection method. Features for each isolated word are extracted and those models were trained. During training process each isolated word is separated into 20ms overlapping windows for extracting 13 MFCCs features. AANN modeling technique is used to model each individual utterance. Thus each isolated word segment from the test sentence is matched against these models for finding the semantic representation of the test input dialogue.

II. VOICE ACTIVITY DETECTION

Voice Activity Detection (VAD) is a technique for finding voiced segments in speech and plays an important role in speech mining applications [3]. VAD ignores the additional signal information around the word under consideration. It can be also viewed as a speaker independent word recognition problem. The basic principle of a VAD algorithm is that it extracts acoustic features from the input signal and then compares these values with thresholds usually extracted from silence. Voice activity is declared if the measured values exceed the threshold. Otherwise, no speech activity is present [4]. VAD finds its usage in a variety of speech communication systems like coding of speech, recognizing speech, hands free telephony, audio conferencing, speech enhancement and cancellation of audio [5]. It identifies where the speech is voiced, unvoiced or sustained and makes smooth progress of the speech process [6]. A frame size of 20 ms, with an overlap of 50%, is considered for VAD. RMS is extracted for each frame. Figure 1 shows the isolated word separation.



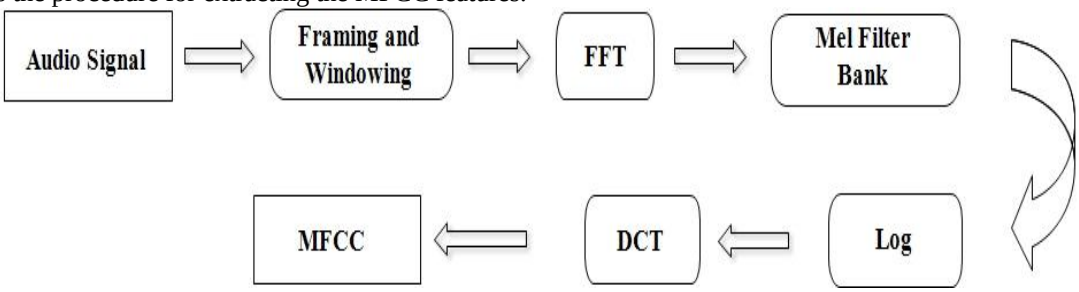
DWT Decomposition model Isolated Word Separations.

III. MEL FREQUENCY CEPSTRAL COEFFICIENTS

An important objective of extracting the features is to compress the speech signal to a vector that is representative of the meaningful information it is trying to characterize. In these works, acoustic features namely MFCC features are extracted.

Mel Frequency Cepstral Coefficients (MFCCs) are short-term spectral based and dominant features and are widely used in the area of audio and speech processing. The mel frequency cepstrum has proven to be highly effective in recognizing the structure of music signals and in modeling the subjective pitch and frequency content of audio signals [7]. The MFCCs have been applied in a range of audio mining tasks, and have shown good performance compared to other features. MFCCs are computed by various authors in different methods. It computes the cepstral coefficients along with delta cepstral energy and power spectrum deviation which results in 26 dimensional features. The low order MFCCs contains information of the slowly changing spectral envelope while the higher order MFCCs explains the fast variations of the envelope [8].

MFCCs are based on the known variation of the human ears critical bandwidths with frequency. The filters are spaced linearly at low frequencies and logarithmically at high frequencies to capture the phonetically important characteristics of speech and audio. To obtain MFCCs, the audio signals are segmented and windowed into short frames of 20 ms. Magnitude spectrum is computed for each of these frames using Fast Fourier Transform (FFT) and converted into a set of mel scale filter bank outputs. The human ear resolves frequencies non-linearly across the audio spectrum and empirical evidence suggests that designing a front-end to operate in a similar non-linear manner improves the performance. A popular solution is therefore filterbank analysis since this provides a much more straightforward route to obtain the desired non-linear frequency resolution. However, filterbank amplitudes are highly correlated and hence, the use of a cepstral transformation in this case is virtually mandatory. Figure 2 describes the procedure for extracting the MFCC features.



3.1 Extraction of MFCC from Audio Signal.

Mel frequency to implement this filterbank, the window of audio data is transformed using a Fourier transform and the magnitude is taken. The magnitude coefficients are then binned by correlating them with each triangular filter. Here, binning means that each FFT magnitude coefficient is multiplied by the corresponding filter gain and the results are accumulated. Thus, each bin holds a weighted sum representing the spectral magnitude in that filterbank channel.

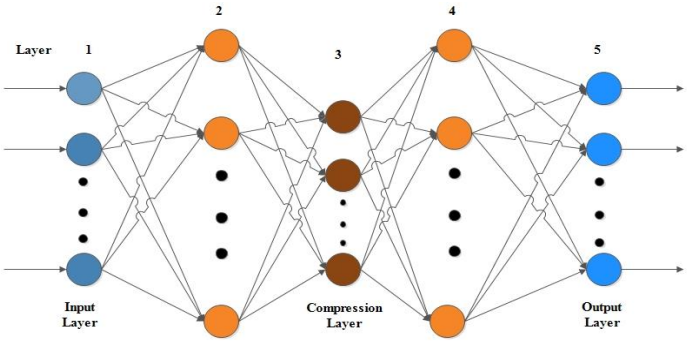
Logarithm is then applied to the filter bank outputs. Discrete Cosine Transformation (DCT) is applied to obtain the MFCCs. Since the mel spectrum coefficients are real numbers, they are converted to the time domain using the DCT. In practice, the last step of taking inverse Discrete Fourier Transform (DFT) is replaced by taking DCT for computational efficiency. The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Typically, the first 13 MFCCs are used as features.

IV. AUTOASSOCIATIVE NEURAL NETWORK (AANN)

Autoassociative Neural Network (AANN) model consists of five layer network which captures the distribution of the feature vector as shown in Figure 3. The input layer in the network has less number of units than the second and the fourth layers. The first and the fifth layers have more number of units than the third layer [9]. The number of processing units in the second layer can be either linear or non-linear. But the processing units in the first and third layer are non-linear. Back propagation algorithm is used to train the network [10].

The activation functions at the second, third and fourth layer are nonlinear. The structure of the AANN model used in our study is 13L 26N 4N 26N 13L for capturing the distribution of acoustic features, where L denotes a linear unit, and N denotes anon-linear unit. The integer value indicates the number of units used in that layer [11]. The non-linear units use tanh(s) as the activation function, where s is the activation value of the unit. Back propagation

learning algorithm is used to adjust the weights of the network to minimize the mean square error for each feature vector [12].



Autoassociate neural network

V. EXPERIMENTAL RESULTS

5.1. Dataset Collection

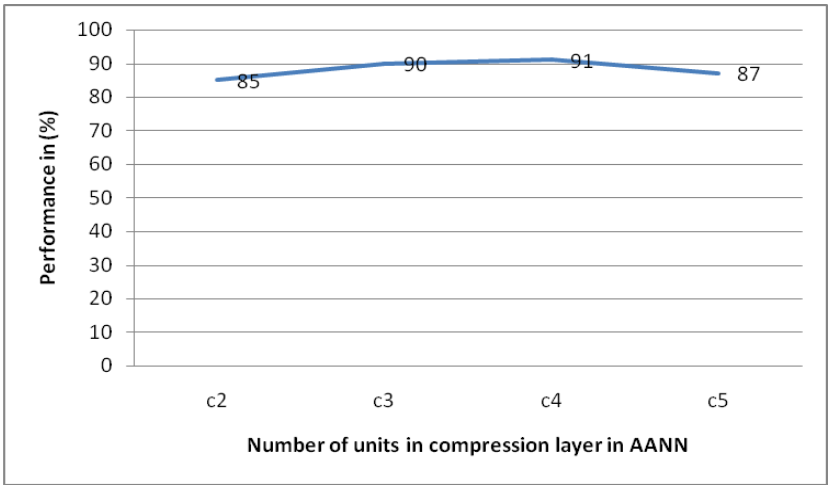
Experiments are conducted for speech recognition audio using Television broadcast speechdata collected from Tamil news channels using a tuner card. A total dataset of 100 different speech dialogue clips, ranging from 5 to 10 seconds duration, sampled at 16 kHz and encoded by 16-bit is recorded. Voice activity detection is performed to isolate the words in each speech file using RMS energy envelope.

5.2. Feature Extraction

In this work the pre-emphasized signal containing the continuous speech is taken for testing. Through VAD the isolated words are extracted from the sentences. Thus frames which are unvoiced excitations are removed by thresholding the segment size. Feature MFCC are extracted from each frame of size 320 window with an overlap of 120 samples. Thus it leads to 13 MFCCs respectively which are used individually to represent the isolated word segment. During training process each isolated word is separated into 20ms overlapping windows for extracting 13 MFCCs features.

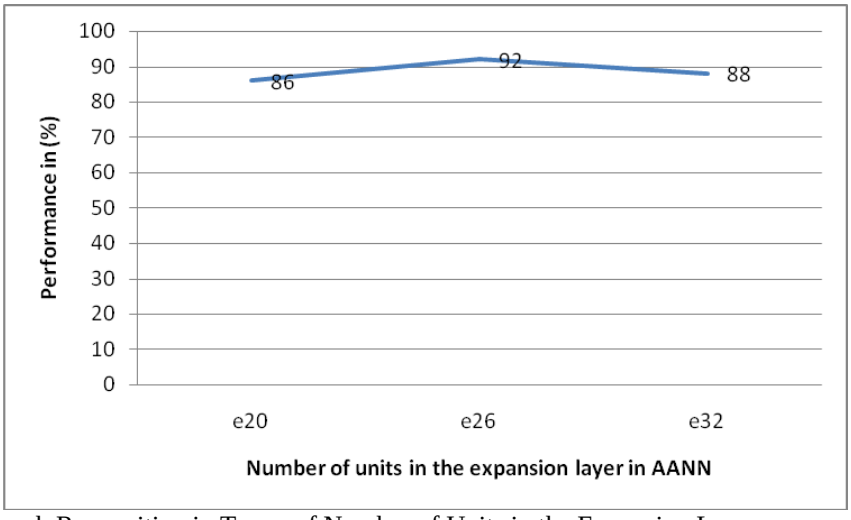
5.3. Classification

Using VAD isolated words in a speech is separated. For training, isolated words from were considered. The training process analyzes speech training data to find an optimal way to classify speech frames into their respective classes. The feature vectors are given as input and compared with the output to calculate the error. In this experiment the network is trained for 500 epochs. The confidence score is calculated from the normalized squared error and the category is decided based on highest confidence score.



Performance of Speech Recognition in Terms of Number of Units in the Compression Layer.

The performance of speech recognition is studied by varying the number of units in the compression layer as shown in Figure 4.



Performance of Speech Recognition in Terms of Number of Units in the Expansion Layer.

The performance of speech recognition in terms of number of units in the expansion layer is shown in Figure 5.The network structures 13L 26N 4N 26N 13L gives a good performance and this structure is obtained after some trial and error

VI. CONCLUSION

In this paper, we have proposed speech recognition system using AANN. Voice Activity Detection (VAD) is used for segregating individual words out of the continuous speeches. Features for each isolated word are extracted and those models were trained successfully. AANN is used to model each Individual utterance. MFCC is calculated as features to characterize audio content. AANN learning algorithm has been used for the recognized speech by learning from training data. Experimental results show that the proposed audio AANN learning method has good performance in 92% speech recognized rate.

VII. REFERENCE

[1] L. Rabiner and B.H. Juang, Fundamentals of Speech Recognition, Pearson Education, Singapore, 2003.
[2] S. Palanivel, Person Authentication using Speech, Face and Visual Speech, Ph.D. thesis, IIT, Madras, September 2004.
[3] Ivan Markovi, Sre´ ckoJuri´ Kavelj and Ivan Petrovi, "Partial Mutual Information Based Input Variable Selection for Supervised Learning Approaches to Voice Activity Detection," Applied Soft Computing Elsevier, vol. 13, pp. 4383-4391, 2013.
[4] Khoubrouy, S. A. and Panahi, I.M.S., "Voice Activation Detection using Teager-Kaiser Energy Measure," International Symposium on Image and Signal Processing and Analysis, pp. 388-392, 2013.
[5] Saleh Khawatreh, Belal Ayyoub, Ashraf Abu-Ein and Ziad Alqadi. A Novel Methodology to Extract Voice Signal Features. International Journal of Computer Applications 179(9):40-43, January 2018.
[6] Tayseer M F Taha and Amir Hussain. A Survey on Techniques for Enhancing Speech. International Journal of Computer Applications 179(17):1-14, February 2018.
[7] O.M. Mubarak, E. Ambikai rajah and J. Epps, "Novel Features for Effective Speech and Music Discrimination," IEEE Engineering on Intelligent Systems, pp. 342-346, 2006.
[8] Meng and J. Shawe-Taylor, "An Investigation of Feature Models for Music Genre Classification using the Support Vector Classifier," International Conference on Music Information Retrieval, Queen Mary, University of London, UK, pp. 604-609, 2005.
[9] ShaojunRen, Fengqi Si, Jianxin Zhou, Zongliang Qiao, Yuanlin Cheng, "A new reconstruction-based auto-associative neural network for fault diagnosis in nonlinear systems," Chemometrics and Intelligent Laboratory Systems, Volume 172, 15 January 2018, Pages 118-128N.
[10] Nitananda, M. Haseyama, and H. Kitajima, "Accurate Audio-Segment Classification using Feature Extraction Matrix," IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 261-264, 2005.
[11] G. Peeters, "A Large Set of Audio Features for Sound Description," Technical representation, IRCAM, 2004.
[12] K. Lee, "Identifying Cover Songs from Audio using Harmonic Representation," International Symposium on Music Information Retrieval, 2006.