International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Issue 01, Volume 6 (January 2019) www.ijirae.com

SPEECH RECOGNITION USING SONOGRAM AND AANN

R.Thiruvengatanadhan
Department of Computer Science and Engineering, **Annamalai University**,
Annamalainagar, Tamilnadu, India
thiruvengatanadhan01@gmail.com

Manuscript History
Number: **IJIRAE/RS/Vol.06/Issue01/JAAE10086**
Received: 02, January 2019
Final Correction: 13, January 2019
Final Accepted: 20, January 2019
Published: **January 2019**
Citation: Thiruvengatanadhan(2019), .Speech Recognition Using Sonogram and AANN. IJIRAE:: International Journal of Innovative Research in Advanced Engineering, Volume VI, 08-12. **doi://10.26562/IJIRAE.2019.JAAE10086**
Editor: Dr.A.Arul L.S, Chief Editor, IJIRAE, AM Publications, India

Abstract— Automatic recognition of speech using computers is a challenging issue. This paper describes a technique that uses Auto associative Neural Network (AANN) to recognized speech based on features using Sonogram. 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—Feature Extraction; Voice Activity Detection (VAD); Sonogram and support vector machines (SVM);

I. INTRODUCTION

An audio signal represents the sound as an electrical voltage. Signal flow is nothing but a route taken by an audio signal for travelling towards the speaker from the source. Audio signal is characterized by bandwidth, power and voltage. Impedance of the signal path determines the relation between power and voltage [1]. Electrical signal is used by analog processors but digital signals are mathematically deals by the digital processors. Due to storage constraints, research related to speech indexing and retrieval has received much attention [2]. As storage has become cheaper, large collection of spoken documents is available online, but there is a lack of adequate technology to explain them. Manual transcription of speech is costly and also has privacy constraints [3]. Hence, the need to explore automatic approaches to search and retrieve spoken documents has increased. Moreover, a wide variety of multimedia data is available online and paves the way for development of new technologies to index and search such media [4]. Speech recognition is a main core of spoken language systems. Speech recognition is a complex classification task and classified by different mathematical approaches: acoustic-phonetic approach, pattern recognition approach, artificial intelligence approach, dynamic dime warping, connectionist approaches and support vector machine.

Proposed work aims to develop a system which has to convert spoken word into text using AANN modeling technique using acoustic feature namely Sonogram. 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. 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 [5]. 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 [6].

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 [7]. It identifies where the speech is voiced, unvoiced or sustained and makes smooth progress of the speech process [8]. A frame size of 20 ms, with an overlap of 50%, is considered for VAD. RMS is extracted for each frame. Fig. 1 shows the isolated word separation.

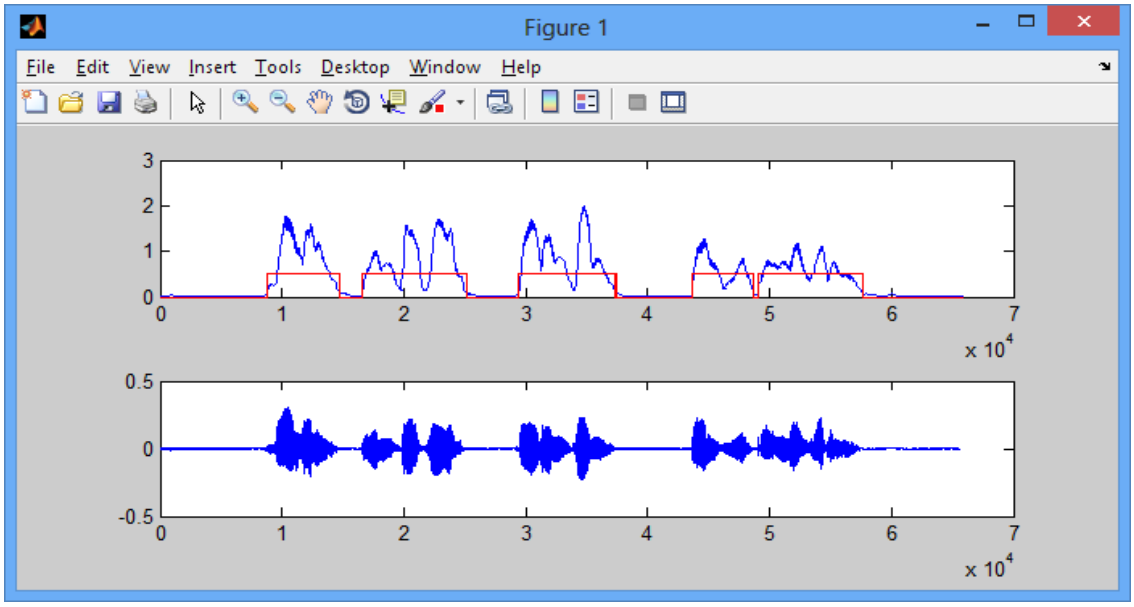

Fig. 1. Isolated Word Separations.

III. SONOGRAM

Pre-emphasis is performed for the speech signal followed by frame blocking and windowing. The speech segment is then transformed using FFT into spectrogram representation [9]. Bark scale is applied and frequency bands are grouped into 24 critical bands. Spectral masking effect is achieved using spreading function. The spectrum energy values are transformed into decibel scale [10]. Equal loudness contour is incorporated to calculate the loudness level. The loudness sensation per critical band is computed. STFT is computed for each segment of pre-processed speech. A frame size of 20 ms is deployed with 50% overlap between the frames. The sampling frequency of 1 second duration is 16 kHz. The block diagram of sonogram extraction is shown in Fig. 2.

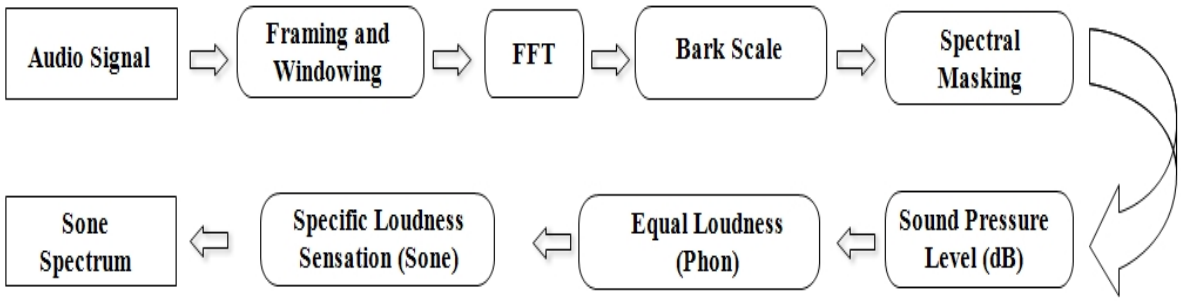

Fig. 2. Sonogram Feature Extractions.

A perceptual scale known as bark scale is applied to the spectrogram and it groups the frequencies based upon the perceptive pitch regions to critical bands. The occlusion of one sound to another is modelled by applying a spectral masking spread function to the signal [11]. The spectrum energy values are then transformed into decibel scale. Phone scale computation involves equal loudness curve which represents different perception of loudness at different frequencies respectively. The values are then transformed into a sone-scale to reflect the loudness sensation of the human auditory system [12].

IV. AUTOASSOCIATIVE NEURAL NETWORK (AANN)

Auto associative 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 [13]. 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 [14]. 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 [15]. 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 [16].

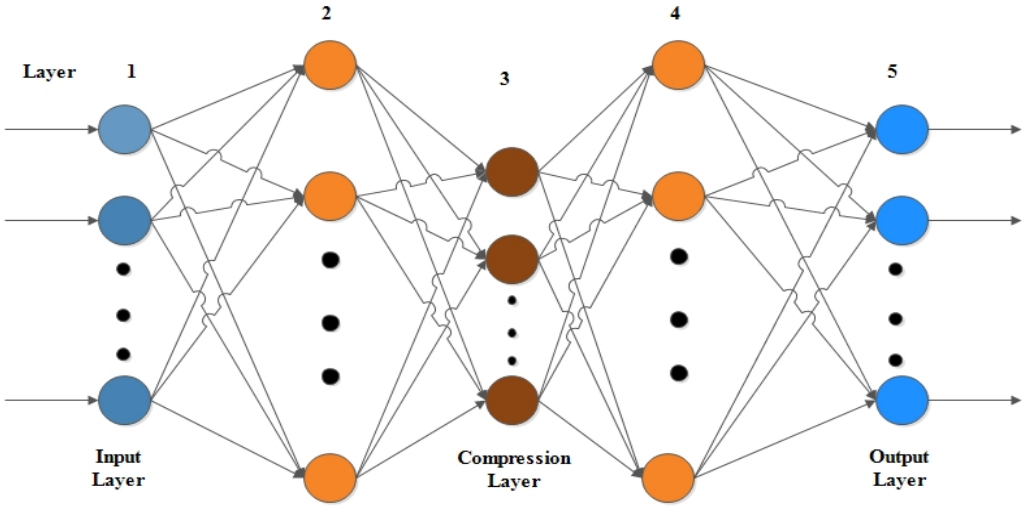

Fig. 3. Auto associate neural network

V. EXPERIMENTAL RESULTS

A. Dataset Collection

Experiments are conducted for indexing speech 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. For each speech file, a database of the isolated words is obtained using VAD.

B. Feature Extraction

VAD the isolated words are extracted from the sentences. Thus frames which are unvoiced excitations are removed by thresholding the segment size. Feature Sonogram are extracted from each frame of size 320 window with an overlap of 120 samples. During training process each isolated word is separated into 20ms overlapping windows for extracting 22 Sonogram features.

C. Classification

Using VAD isolated words in a speech is separated. AANN are created for each isolated word. 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. For testing 22 dimensional Sonogram feature vectors were given as input. 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. The performance of speech recognition is studied by varying the number of units in the compression layer as shown in Fig. 4.

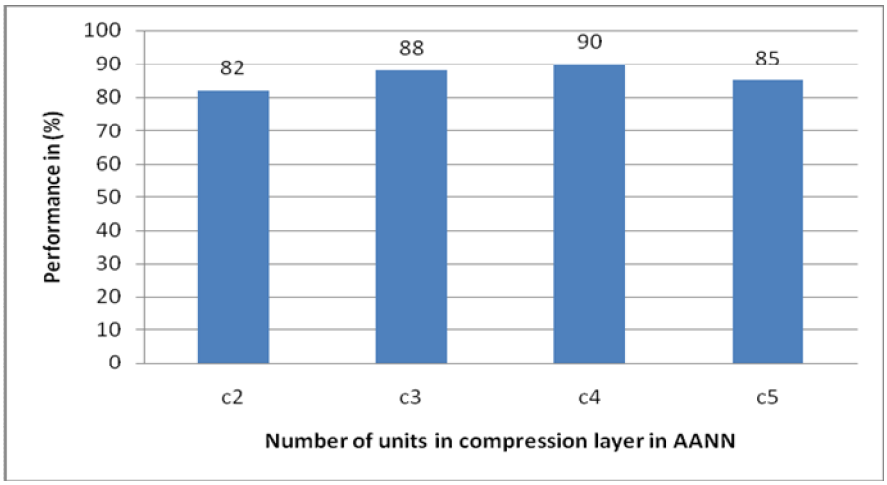

Fig. 4. Performance of Speech Recognition in Terms of Number of Units in the Compression Layer.

The performance of speech recognition in terms of number of units in the expansion layer is shown in Fig. 5.The network structures 22L 44N 4N 44N 22L gives a good performance and this structure is obtained after some trial and error

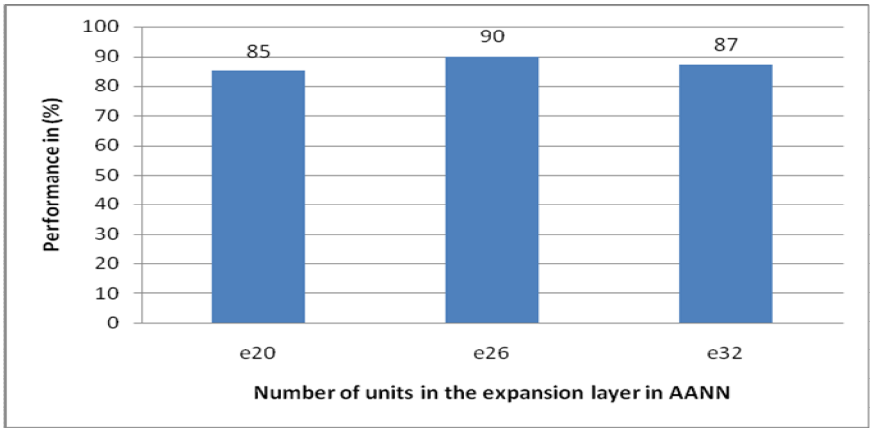

Fig. 5. Performance of Speech Recognition in Terms of Number of Units in the Expansion Layer.

VI. CONCLUSIONS

In this paper, 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. Sonogram is calculated as features to characterize audio content. AANN is used to model each Individual utterance. Sonogram 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 90% speech recognized rate.

REFERENCES

1. Albert Bregman, Auditory Scene Analysis, MIT Press, Cambridge, 1990.
2. Tsung-Hsien Wen, Hung-Yi Lee, Pei-hao Su and Lin-shan Lee, "Interactive Spoken Content Retrieval by Extended Query Model and Continuous State Space Markov Decision Process," IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8510-8514, 2013.
3. Iswarya, P. and Radha, V, "Speech and Text Query Based Tamil - English Cross Language Information Retrieval system," International Conference on Computer Communication and Informatics, pp. 1-4, Coimbatore, 2014.

4. Chien-Lin Huang, Chiori Hori and Hideki Kashioka, "Semantic Inference Based on Neural Probabilistic Language Modeling for Speech Indexing," IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8480-8484, 2013.
5. 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.
6. 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.
7. 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.
8. 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.
9. Xiaowen Cheng, Jarod V. Hart, and James S. Walker, "Time-frequency Analysis of Musical Rhythm," Notices of AMS, vol. 56, no. 3, 2008.
10. Ausgef¨uhrt, Evaluation of New Audio Features and Their Utilization in Novel Music Retrieval Applications, Master's thesis, Vienna University of Technology, December 2006.
11. Eberhard Zwicker and Hugo Fastl, "Psychoacoustics-Facts and Models," Springer Series of Information Sciences, Berlin, 1999.
12. M. R. Schroder, B. S. Atal, and J. L. Hall, "Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear," Journal of the Acoustical Society of America, vol. 66, pp. 1647-1652, 1979.
13. 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.
14. 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.
15. G. Peeters, "A Large Set of Audio Features for Sound Description," Technical representation, IRCAM, 2004.
16. K. Lee, "Identifying Cover Songs from Audio using Harmonic Representation," International Symposium on Music Information Retrieval, 2006.

