ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

Speech recognition using MFCC and RBFNN

R. Thiruvengatanadhan

Assistant Professor, Dept. of Computer Science and Engineering, Annamalai University, Annamalainagar, Tamilnadu, India

ABSTRACT: Speech Recognition approach intends to recognize the text from the speech utterance which can be more helpful to the people with hearing disabled. This paper describes a technique that uses support vector machines (SVM) to recognized speech based on features using Mel Frequency Cepstral Coefficients (MFCC). Modeling techniques such as Radial Basis Function Neural Network (RBFNN) 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 RBFNN shows good performance in recognized rate.

KEYWORDS: Speech Recognition, VAD, MFCC, RBFNN

I.INTRODUCTION

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 time warping, connectionist approaches and support vector machine. There have also been applications to speech recognition problems, namely phonetic classification and post-classification of speech recognition hypotheses. Large speech databases such a Television program, radio broadcasts, CDs and DVDs are available online and the necessity to organize such huge databases becomes essential these days [1]. As Large Vocabulary Continuous Speech Recognition (LVCSR) is imperfect, automatic speech transcripts contain errors. 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].

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 [4]. 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 [5].

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 [6], [7]. It identifies where the speech is voiced, unvoiced or sustained and makes smooth progress of the speech process. Fig. 1 shows the isolated word separation.

DOI: 10.15680/IJIRCCE.2018.0606066

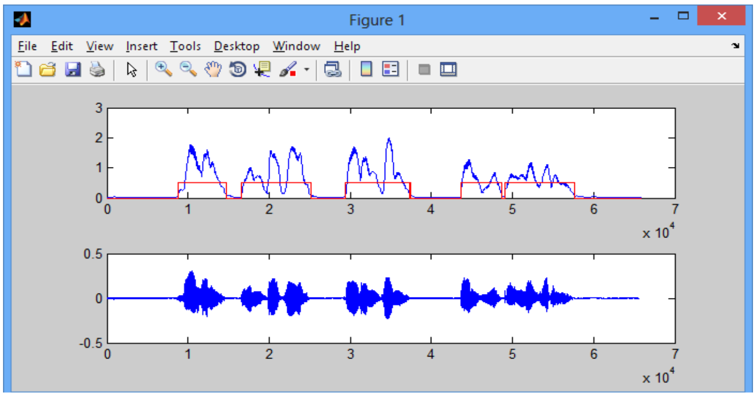

Fig. 1 Isolated Word Separations.

III. MEL-FREQUENCY CEPSTRAL COEFFICIENTS (MFCC)

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 [8], [9]. The MFCCs have been applied in a range of audio mining tasks, and have shown good performance compared to other features.

MFCCs are short-term spectral 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. Fig. 2 describes the procedure for extracting the MFCC features.

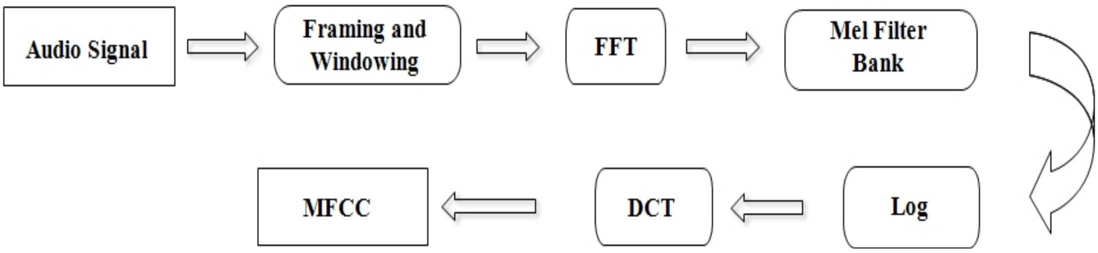

Fig. 1 Extraction of MFCC from Audio Signal.

IV. RADIAL BASIS FUNCTION NEURAL NETWORK (RBFNN)

Radial basis function neural network (RBFNN) forms a special architecture with several distinctive features. A typical RBF neural network classifier has three layers, namely input, hidden, and output layer. The input layer of the network is made of source nodes that connect the coordinates of the input vector to the nodes in the second layer. The second layer, the only hidden layer in the network, includes processing units called the hidden basis function units which are located on the centers of well chosen clusters. Each hidden layer node adopts a radial activated function, and output nodes implement a weighted sum of hidden unit outputs [10]. The output layer is linear, and it produces the predicted class labels based on there sponse of the hidden units. The structure of multi-input and multi-output RBF neural network is represented by Fig. 2.The parameters of an RBF type neural network consist of the centers spread the basis functions at the hidden layer nodes and the synaptic weights of the output layer nodes. The RBF centers are also points in the input space. It would be ideal to have them at each distinct point on the input space, but for any realistic problem, only a few input points from all available points are selected using clustering.

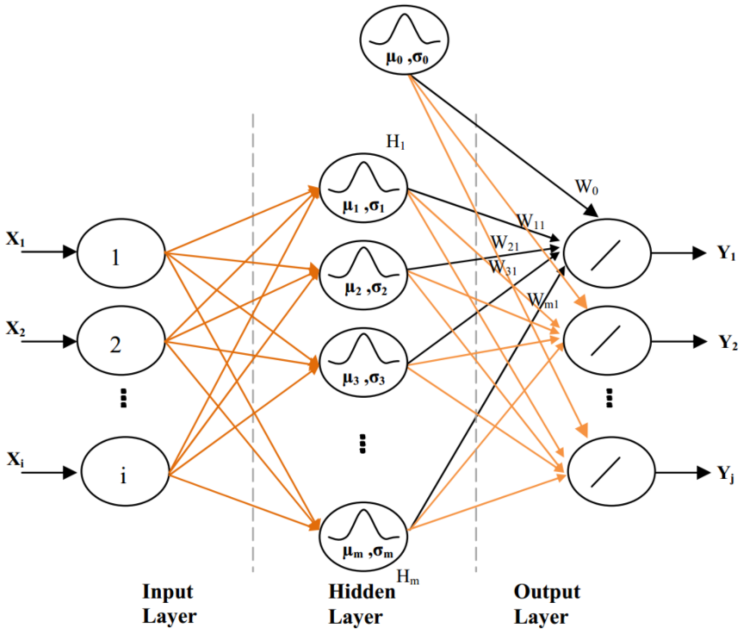

Fig. 2 RBFNN Architecture

V. EXPERIMENTAL RESULTS

A. The database

Experiments are conducted for speech recognition audio using Television broadcast speech data 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.

B. Acoustic 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.

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 RBFN is trained by adaptively updating the free parameters, i.e. center and width of the basis function, and the weight between the hidden and output neurons of the network. To select an optimal RBFN model, the number of

neurons in the hidden layer was varied from 2 to 30, and the learning rate was varied between 0.05 and 0.5. The initial basis function centers were chosen randomly from the input space, and the initial weight values were chosen randomly between ±0.9. Normalized datasets were used for the training, testing, and validation of the RBFN model. The best network was found to be one having 26 basis functions with a learning rate of 0.9 and 0.05 for center and weight respectively. The prediction errors of the validation patterns are larger because these patterns are outside the training space. The Fig. 3 shows the comparison of various means in RBFNN.

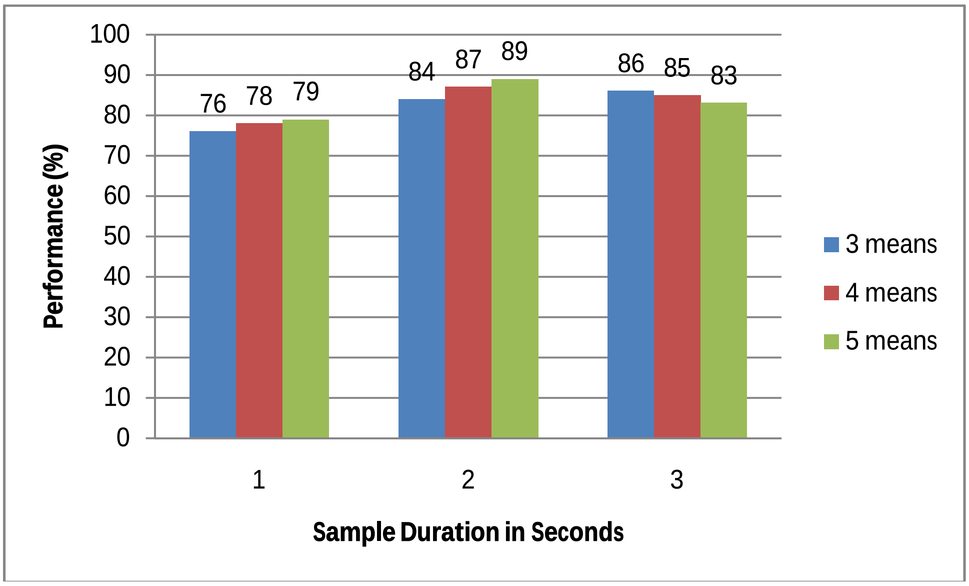

Fig. 3 Comparison graph for various means in RBFNN

VI. CONCLUSION

In this paper, we have proposed speech recognition system using RBFNN. 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. RBFNN is used to model each Individual utterance. MFCC is calculated as features to characterize audio content. RBFNN learning algorithm has been used for the recognized speech by learning from training data. Experimental results show that the proposed audio RBFNN learning method has good performance in 89% speech recognized rate.

REFERENCES

[1] YaliZheng, Chisaki, Y. and Usagawa T., “Speech/Music Indexing for Audio Life Logs from Portable Device Record,” IEEE International Conference on Advanced Computer Science and Information Systems, pp. 173 -178, 2013.

[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] 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.

[5] 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.

[6] El Bachir Tazi, Abderrahim Benabbou and Mostafa Harti, “Voice Activity Detection for Robust Speaker Identification System,” *IJCA Special Issue on Software Engineering, Databases and Expert Systems*, pp. 35-39, 2012.

[7] Nitin N Lokhande, Navnath S Nehe and Pratap S Vikhe, "Voice Activity Detection Algorithm for Speech Recognition Applications," *IJCA Proceedings on International Conference in Computational Intelligence*, vol. 6, pp. 5-7, 2012.

[8] 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.

[9] Ahmad R. Abu-El-Quran, Rafik A. Goubran, and Adrian D. C. Chan, "Security Monitoring using Microphone Arrays and Audio Classification," *IEEE Transaction on Instrumentation and Measurement*, vol. 55, no. 4, pp. 1025-1032, August 2006.

[10] D.Tjondronegoro, Y.Chen, and B.Pham, "The power of play break for automatic detection and browsing of self consumable sport video highlights", In Proceedings of the ACM Workshop on Multimedia Information Retrieval, pp. 267-274, 2004.