Wavelet – Neural Data Mining Approach for Spoken Keyword Spotting

[1]K. A. Senthil Devi, [2]Dr. B. Srinivasan
[1]Assistant Professor, [2]Associate Professor
[1]Department of Computer Science ,
[1]Gobi Arts & Science College, Tamil Nadu, India.

***Abstract*— Spoken keyword spotting is a technologically relevant problem in speech data mining. It is essential to identify the occurrences of specified keywords expertly from lots of hours of speech contents such as meetings, lectures, etc. In this paper, Wavelet Packet Decomposition (WPD) and Neural Network (NN) based data mining model (WPDNNM) is explored for keyword spotting. Speech data is first decomposed with Haar, Daubechies2 and Simlet4 wavelets packets. Then, some significant features are extracted from the decomposed speech data. Back Propagation Neural Network (BPNN) is trained with three predefined spoken keywords based on known features and finally, input speech features are compared with keyword features in the trained BPNN for spotting the occurrences of the specified keyword. The method of this paper is tested with 5 minutes lecture data. This method is compared with Discrete Wavelet Transformation (DWT) feature extraction based keyword spotting. Experimental results show that the wavelet - neural method with WPD of Daubechies2 wavelet is more accurate than with Haar and Simlet4 wavelets.**

***Index Terms*— Spoken keyword spotting, Speech data mining, Wavelet Packet Decomposition, Discrete Wavelet Transformation, BPN neural network, word detection.**

I. INTRODUCTION

With the advent of inexpensive storage space and faster processing over the past decade, data mining research has started to penetrate new grounds in areas of speech and audio processing as well as spoken language dialog [9]. Speech mining helps in the areas of prediction, search, word spotting, explanation, learning, and language understanding. Effective techniques for mining speech data can impact numerous business and government applications.

Spoken keyword spotting is a very crucial and promising branch in speech mining and it is useful to retrieve the speech files which enclose the words associated with an application-specific domain [6]. It is essential to classify expertly lots of hours of speech contents such as meetings, lectures, etc. Keyword spotting technologies are widely used in the security services, telecommunication companies, radio stations, call-centers, broadcasting companies and other organizations that use a large stream or archive of speech information. Wavelet theory could naturally play an important role in data mining because wavelets could provide data presentations that enable efficient and accurate mining [8].

Artificial neural networks have been already proposed in many speech mining applications. In paper [3], Jothilakshmi at el proposed an approach for spoken keyword detection using Auto Associative Neural Networks. The work concerned the use of the distribution capturing ability of the auto associative neural network for spoken keyword detection. It involves sliding a frame-based keyword template along the speech signal and using confidence score obtained from the normalized squared error of AANN to efficiently search for a match. Another approach with Support Vector Machine (SVM) was proposed by J. Sangeetha and S.Jothilakshmi [6]. This work concerned sliding a frame-based keyword template along the speech signal and using SVM misclassification rates obtained from the hyperplane of two classes efficiently search for a match.

A new word spotting approach in continuous speech is introduced in [4] that use wavelet transform based feature extraction and Euclidean distance. The system is capable of identifying and localizing a target word in a continuous speech of any length. In our previous work, we developed an approach for keyword spotting using wavelet packet transformation and sliding frame method with Euclidean distance calculation [7]. It consumes more time to identify the occurrences of a keyword.

The paper work involves in designing a new method which combines wavelet packet and neural network techniques for identifying occurrences of keywords in the input speech contents. WPD based feature extraction and the BPNN is used for identifying keyword match. Performance of the overall system depends on signal decomposition, feature extraction and classification. Here, accuracy has been increased by the combination of wavelet and artificial neural network. This approach is also compared with Discrete Wavelet Transformation (DWT) feature extraction based keyword spotting.

II METHODOLOGY

Wavelet Analysis

The wavelet transform is a recently developed mathematical tool for signal processing. It has been applied successfully in speech and image processing. Wavelets are functions which represents speech signals with good time and frequency resolution. The basic concept in wavelet analysis is to select a proper wavelet (mother wavelet), then perform an analysis using its translated and dilated versions. There are many kinds of wavelets which can be used as a mother wavelet, such as the Haar wavelet, Meyer wavelet, Coiflet wavelet, Daubechies wavelet, Morlet wavelet and etc.

Discrete Wavelet Transformation

Filters are one of the most widely used signal processing functions. Wavelets can be realized by iteration of filters with rescaling. In the discrete wavelet transform, a speech signal can be analyzed by passing it through an analysis filter bank followed by a decimation operation. When a signal passes through these filters, it is split into two bands [7]. The low pass filter performs an averaging operation which extracts the coarse information of the signal. The high pass filter performs a differencing operation which extracts the detail information of the signal. The output of the filtering operations is then decimated by two.

Wavelet Packet Decomposition

The wavelet packet method is a generalization of wavelet decomposition that offers a richer range of possibilities for signal analysis and which allows the best matched analysis to a signal [8]. The WPD divides the low and also the high frequency subband of signal. In wavelet analysis, a signal is split into an approximation and a detail coefficient. The approximation coefficient is then itself split into a second-level approximation coefficients and detail coefficients, and the process is repeated. In wavelet packet analysis, the details as well as the approximations can be split. The top level of the WPD tree is the time representation of the signal. As each level of the tree is traversed there is an increase in the tradeoff between the time and frequency resolution. The bottom level of a fully decomposed tree is the frequency representation of the signal. Figure 1 shows the level decomposition using wavelet packet transform.

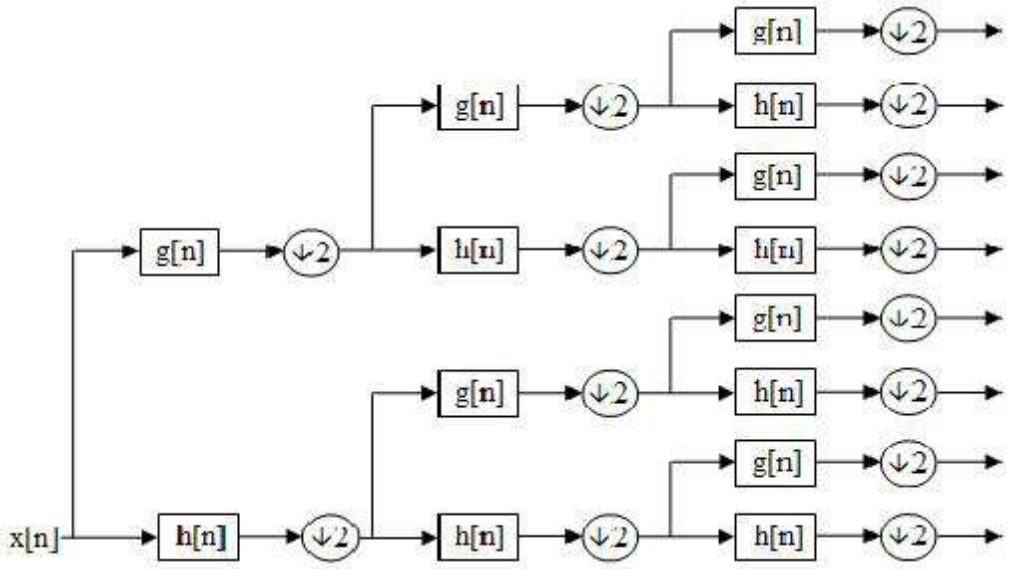

Figure 1. Level 3 decomposition using wavelet packet decomposition.

Neural Network Classifier

Recently, neural network (NN) based methods have shown tremendous success on speech processing and data mining tasks. The neural network model is a powerful tool used to perform keyword spotting tasks as performed by human brain. The neural network approach for keyword spotting is based on the type of the learning mechanism applied to generate the output from the network [5]. Among number of neural networks, the Multi-Layer Perceptron (MLP) with back propagation (BP) neural network algorithm is found to be effective for solving a number of real world problems.

Back Propagation Neural Network

This section presents the architecture of the back propagation algorithm. Input vectors and corresponding target vectors are used to train a network until it can approximate a function, associate input vectors with specific output vectors, or classify input vectors in an appropriate way as defined in this study. The network consists of three layers: input layer, output layer and the intermediate layer i.e. the hidden layer [2]. These layers comprises of the neurons which are connected to form the entire network. Weights are assigned on the connections which marks the signal strength. The weight values are computed based on the input signal and the error function back propagated to the input layer. Networks with biases, a sigmoid layer and a linear output layer are capable of approximating any function with a finite number of discontinuities. The back propagation algorithm consists of two paths; forward path and backward path. Forward path contain creating a feed forward network, initializing weight, simulation and training the network. The network weights and biases are updated in backward path**.** The neural network model is shown in figure 2.

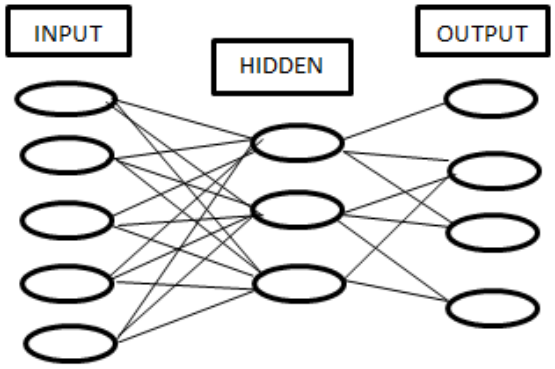

Figure 2. The neural network model.

III. PROPOSED WAVELET-NEURAL (WPDNNM) APPROACH FOR KEYWORD SPOTTING

The block diagram shown in Figure 3 gives the actual implementation of the method proposed in this paper.

Preprocessing

Speech signal pre-processing covers digital filtering, to enhance the speech quality in terms of silence removal, noise reduction, resampling and segmentation. In this proposed system moving – average filter function is used to filter the input speech and keyword given. The moving average filter is a simple Low Pass FIR (Finite Impulse Response) filter commonly used for smoothing signals. The moving average filter takes average of samples for filtering the noise from signal. The preprocessed output is then passed to the next stage wavelet decomposition.

Wavelet based Decomposition

The wavelet packet decomposition is applied to the enhanced speech signal to acquire its frequency domain spectrum and filter out unwanted frequencies from input and template speech. The selected frequency spectrum is passed to feature extraction process that extracts some important features out of time and frequency domain speech signal. These wavelets have different specificities. In this work, Haar, Daubechies2 and Simlet4 wavelet packets are applied for speech decomposition.

Feature Extraction

The decomposed frequency spectrum is passed to feature extraction process that extracts some important features out of time and frequency domain speech signal. The features which are extracted and used for the test and template frame matching are listed below:

- RMS (Root Mean Square level)
- Correlation
- Homogeneity
- Standard Deviation
- Variance
- Smoothness
- Kurtosis
- Skewness

3.2 Neural Network Training

The performance of the system depends on the neural network model deployed to identify the words in the input data. Back Propagation algorithm which is based on the concept of improving the network performance by reduction of error from the output data is used to train the network in this system. This algorithm works in batch mode in which the weight updates take place after much propagation. The implementation of this algorithm is faster and efficient depending upon the amount of input-output data available in the layers.

Before training the feed forward network, the weight and biases are initialized. Once the network weights and biases have been initialized, the network is ready for training. We used random numbers around zero to initialize weights and biases in the network. The training process requires a set of proper inputs and targets as outputs. During training, the weights and biases of the network are iteratively adjusted to minimize the network performance function. The default performance function for feed forward networks is mean square errors, the average squared errors between the network outputs and the target output.

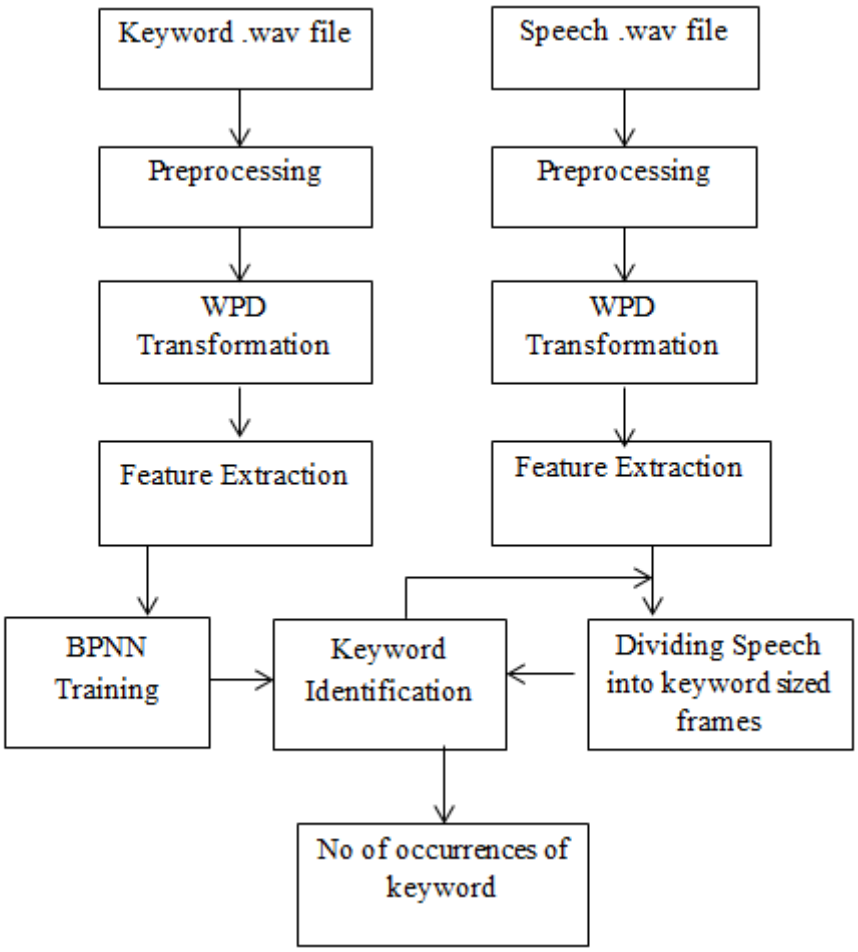

Figure 3. Block diagram of proposed wavelet-neural (WPDNNM) keyword spotting approach

Keyword Detection

The proposed method tries to detect the predefined keyword in the given audio stream by splitting the input speech content into blocks in the size of keyword. Sliding frame method is used splitting the speech stream [4]. In this process, initially a block of frames such that the number of frames in the block is equal to number of frames of the keyword signal are selected from the input signal starting from the first frame. This block of feature vectors is then matched in neural network classifier with the trained keyword. The process is repeated to the next block of frames in the input speech.

IV EXPERIMENTAL RESULTS

In the proposed wavelet – neural approach, Wavelet packet decomposition based features are used with back propagation neural network classifier. Speech file of 5 minutes length is used in this experiment. Three spoken keywords are used in the experiment for training th network. Recording is done in a silent room environment with a PC computer with AUDACITY sound recording package in frequency 8000Hz. The predefined keywords are shown the tables I. The wave form of recorded keyword "ondru" is shown in the figure 4. The speech content and keywords are decomposed with Haar, Daubechies2 and Simlet4 wavelet packets. The wavelet packet decomposition of the word "ondru" is shown is the figure 5.

In this work, a three – layer network is developed with back propagation algorithm. An input vector and the corresponding desired output are considered first. The input is propagated forward through the network to compute the output vector. The output vector is compared with the desired output, and the errors are determined. The errors are then propagated back through the network from the output to input layer. The process is repeated until the errors being minimized. The input layer of network contains 18 neurons and the output layer contains 6 neurons corresponding to 3 pre-defined keywords.

The hidden layer is responsible for internal representation of the data and the information transformation input and output layers. If there are too few neurons in the hidden layer, the network may not contain sufficient degrees of freedom to form a representation. If too many neurons are defined, the network might become over trained. Therefore, an optimum design for the number of neurons in the hidden layer is required. In this research, we used one hidden layer with 15 neurons.

After obtaining the speech features for each frame of the given keyword signal, BPNN model is created to capture the distribution of this keyword signal. Likewise the speech features are obtained for each frame of the given input signal in which the given keyword should be detected. Initially a block of frames such that the number of frames in the block is equal to number of frames of the keyword signal are selected from the input signal starting from the first frame. This block of feature vectors are used for testing the model. If the word corresponding to the block of frames is same as the keyword then the score for the block will be very high. If the word corresponding to the block of frames is completely different from the keyword, the feature vectors from the block may not fall into the distribution and the model gives low score. The approach is compared with Discrete Wavelet Transformation (DWT) feature extraction based keyword spotting. The accuracy of the models is shown in the table II.

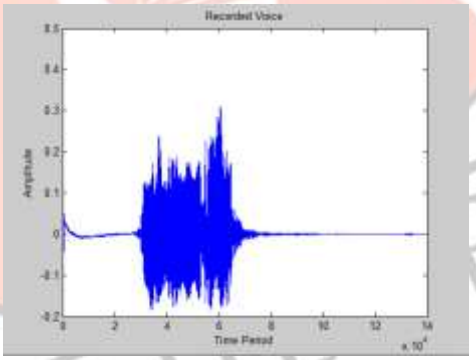

Figure4. Recorded keyword “ondru”

TABLE I

Keywords trained on BPNN
Ondru (1)
Irandu (2)
Mundru (3)

TABLE II

Name of mother wavelet used for WPD	Keyword	Accuracy in %	
		DWT Model	WPD Model
Haar	Ondru	90	94
	Irandu	85	92
Daubechies2	Ondru	90	94
	Irandu	92	95
Simlet4	Ondru	86	90
	Irandu	92	92

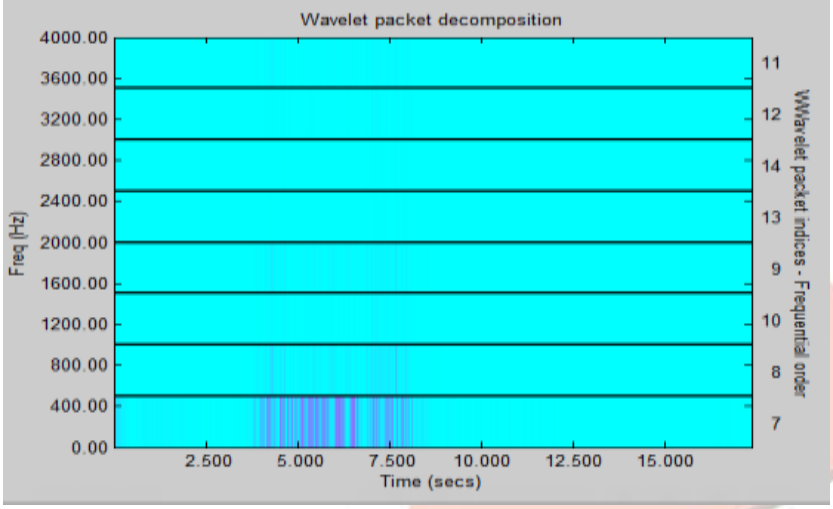

Figure 5. WPD with Haar wavelet for the word "ondru".

Conclusion

This paper presents a resourceful new technique for the keyword spotting system using wavelet packet and neural network. Features have been extracted using wavelet packet decomposition and discrete wavelet transformation. These features extracted in this way are more suitable for keyword spotting than MFCC. The performance of wavelet packet with neural network is appreciable while comparing with the DWT based method since wavelet packet analysis can provide a more precise frequency resolution than the wavelet analysis. It also has compact support in time as well as in frequency domain and adapts its support locally to the signal which is important in time varying signal. The results show that the WPDNNM model is more accurate with Daubechies2 wavelet than Haar and Simlet4 wavelets.

References

1. Gokhale, M. Y.DaljeetKaurKhanduja, 2010, "Time Domain Signal Analysis Using Wavelet Packet Decomposition Approach", Int. J. Communications, Network and System Sciences,2010, 321 – 329.
2. Heerman P.D. and N.Khazenie, "Classification of multispedtral remote sensing data using a bach propagation neural network," IEEE Trans, Geosci. Remote Sensing, vol.GE_30,no.1,1992, pp.81-88.
3. Jothilakshmi, S., Spoken keyword detection using autoassociative neural networks, International Journal Speech Technology, *Springer*, 2013, pp. 83-89.
4. Khan, W. and Holton, R., Word spotting in continuous speech using wavelet transform, IEEE International Conference on Electro/Information Technology, 2014, pp.275-279.

5. Dr. Rama Kishore, Taranjit Kaur, "Backpropagation Algorithm: An Artificial Neural Network Approach for Pattern Recognition", International Journal of Scientific & Engineering Research, Volume 3, Issue 6, June-2012 1 ISSN 2229-5518.
6. Sangeetha, J. and Jothilakshmi, S., "A novel spoken keyword spotting system using support vector machine", Engineering Applications of Artificial Intelligence, Springer, 2014, pp. 287–293.
7. Senthil devi K.A., Dr.Srinivasan B., "A novel Keyword Spotting Algorithm in speech mining using wavelet", International Journal of Current Research Vol. 8, Issue, 08, pp.36943-36946, August, 2016.
8. Tao Li, Sheng Ma, Mitsunori Ogihara, " Wavelet methods in data mining", Chapter 27 Data Mining and Knowledge Discovery Handbook ,Springer, 2005, pp 603-626..
9. Thambiratnam, Albert J. K," Acoustic keyword spotting in speech with applications to data mining." PhD Thesis Published in Queensland University of Technology, 2005.