Volume 4, Issue 3, March-2017, pp. 111-118ISSN (O): 2349-7084

International Journal of Computer Engineering In Research Trends

Available online at: www.ijcert.org

Spoken Keyword Spotting System Design Using Various Wavelet Transformation Techniques with BPNN Classifier

Senthil Devi K. A.*[1], Dr. B. Srinivasan [2]

[1]Assistant Professor, Gobi Arts & Science College, Tamil Nadu, India.[2]Associate Professor, Gobi Arts & Science College, Tamil Nadu, India. senthildevigasc@gmail.com[1], srinivasan_gasc@yahoo.com[2].

Corresponding author*: Phone: +0-9486399353

Abstract:-SpokenKeyword spotting is a speech data mining task which is used to search audio signals for finding occurrences of a specified spoken word in the given speech file.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, keyword spotting system designed with various wavelet transformation techniques and BackpropagationNeural Network (BPNN). Back Propagation Neural Network (BPNN) is trained with two 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 tested with ten speech content often different speakers. Various statistical features extraction techniques with wavelet transformation are used. Performance comparison is done among these methods with Haar, Daubechies2 and Simlet 4 wavelets.

Keywords:-Spoken keyword spotting, Speech data mining, MFCC, Wavelet Packet Decomposition, Discrete Wavelet Transformation, BPN neural network, wavelet families.

---

1. Introduction

Keyword spotting in continuous speech considered as a challenging issue due to dynamic nature of speech.To identify the keyword from speech signals and reveal the underlying dynamics that corresponds to the speech signals, theproper signal processing technique is needed. Typically, the process of signal processing transforms a time-domain signal into another domain, with the purpose of extracting the characteristic information embedded within the signal that is otherwise not readily observable in its original form.

Wavelet theory has become one of the most important and powerful tools of signal representation [1]. Wavelet transformsproper tools for analyzing non-stationary signals. Since speech consists of both High and low-frequency components, short and long duration sounds, the wavelet transform is well suited to this type of analysis. In many speech related applications, wavelets are used to perform preprocessing (e.g. noise filtering), dimensionality reduction and data transformation. Wavelet theory could naturally play a major role in data mining because wavelets could provide data presentations that enable efficient and accurate mining [15].

Wavelet theory could naturally play a major role in data mining since it is well founded and of very practical use. Wavelets have many favorable properties, such as vanishing moments, hierarchical and multiresolution decomposition structure, linear time and space complexity of the transformations, decorrelated coefficients, and a wide variety of basis functions. These properties could provide considerably

more efficient and effective solutions to many data mining problems. As keyword spotting is a technologically relevant problem in speech data mining, wavelet theory is a suitable tool to perform the spotting task from lots of hours of speech contents

Artificial neural networks have been used as an important classification tool in many speech mining applications. In paper [7], an approach for spoken keyword detection was proposed using Auto Associative Neural Networks. The same work was again implemented J. Sangeetha and S.Jothilakshmi [12] with Support Vector Machine (SVM). Wavelet transformation was first used in spoken keyword spotting in continuous speech by Khan at el [8]. Wavelet transformed features of desired keyword and testing input speech are compared by calculating 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 in sliding frame method with Euclidean distance for similarity measure [13] and BPNN network [14].

The paper work involves inthedesign of keyword spotting method which uses various statistical feature extraction techniques with wavelet transformation methods. BPNN neural network is used for identifying occurrences of keywords in the input speech contents. The performance of each feature extraction techniques used in this keyword spotting system is compared.

2. Speech Transformation

To extract information from speech signals and reveal the underlying dynamics thatcorresponds to the signals, theproper signal processing technique is needed. Typically, the process of signal processing transforms a time-domain signal into another domain, with the purpose of extracting the characteristic information embedded within the signal that is otherwise not readily observable in its original form. Mathematically, this can be achieved by representing the time-domain signal as a series of coefficients, based on a comparison between the signal x (t) and a set of known, template functions $\{\Psi_n(t)\}n \square z$ as [2, 9]

$$C_n = \int x(t)\, y_n(t)\, dt \quad \text{------- (1)}$$

Where (·)* stands for the complex conjugate of the function (·). The inner product between the two functions x(t) and $y_n(t)$ is defined as

$$\langle x, y_n \rangle = \int x(t)\, y_n(t)\, dt$$

2.1 Fourier Transform

The Fourier transform is the most widely applied signal processing tool in science and engineering. It reveals the frequency composition of a time series x (t) by transforming it from the time domain into the frequency domain. In 1807, the French mathematician Joseph Fourier [4], found that any periodic signal can be presented by a weighted sum of a series of sine and cosine functions.

The Fourier transform of a signal x (t) can b e expressed as

$$X(f) = \langle x e^{2\pi f} \rangle = \int x(t) e^{2\pi f}\, dt$$

2.2 Wavelet Transform

A wavelet is used to analyze a given function or continuous-time signal at a specified scale. This function plays the role of the window from the case of STFT, but it has a second parameter, additional to the position, the scale. It can be moved to various locations of the signal as shown in Figure 1.

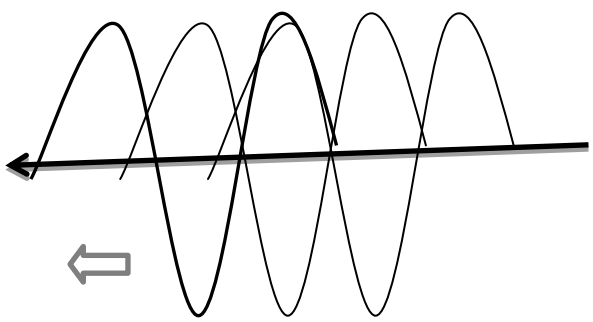

Figure 1 Location in atime of a wavelet with a given scale.

2.3 Need forWavelet Transform in Speech Signal Processing

The Wavelet Transform provides a time-frequency representation of the signal. It was developed to overcome the shortcoming of the Short-Time Fourier Transform (STFT), which can also be used to analyze non-stationary signals. While STFT gives a constant resolution at all frequencies, the Wavelet Transform uses themulti-resolution technique by which different frequencies analyzed with various resolutions.

The wavelet transform enables variable window sizes in analyzing various frequency components within a signal [9]. Thisis realized by comparing the signal with a set of template functions obtained from the scaling (i.e., dilation and contraction) and shift (i.e., translation along the time axis) of a base wavelet and looking for their similarities. The wavelet transform of a signal x(t) can be expressed as

$$w(s,t) = \langle x, y_{s,t} \rangle = \frac{1}{\sqrt{s}} \int x(t)\, y^{*}\left(\frac{t-t}{s}\right) dt$$

Where the symbols> 0 represents the scaling parameter, which determines the time and frequency resolutions of the scaled base wavelet $y((t-t)/s)$. The specific values of s are inversely proportional to the frequency. The symbol t is the shifting parameter, which translates the scaled wavelet along the time axis. The symbol y *(t) denotes the complex conjugation of the base wavelet y (t).

3. Wavelet Transformation

3.1 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 [13]. 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.

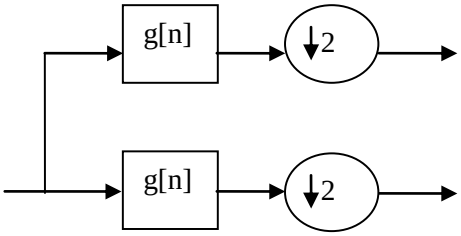

Figure. 2 DWT highpassand lowpass filters

3.2 Wavelet Packet Decomposition

The Discrete Wavelet Transform (DWT) is obtained by the discretization of the CWT in the time-frequency plane [3] and is used to decompose discrete time signals. The result achieved at each decomposition level is composedof two types of coefficients: approximation coefficients and detail coefficients. The approximation coefficients are obtained by low-pass filtering the input sequence, followed by down-sampling. The detail coefficients are obtained by high-pass filtering the input sequence followed by down-sampling. The sequence of approximation coefficients constitutes the input for the next iteration. Each decomposition level corresponds to a specified resolution. The resolution decreases with the increasing of the number of de composition levels. The DWT is invertible. Its inverse is named Inverse DWT (IDWT). At each resolution level, the approximation and the detail sequences are needed for the reconstruction of the approximation signal from the previous resolution level.

The Discrete Wavelet Transform has two features: the wavelet mother and the number of decomposition levels. Discrete wavelets can be scaled

and translated in discrete steps, and a wavelet representation is the following:

$$y_{j,n}(t) = \frac{1}{\sqrt{2^j}} y\left[\frac{t - 2^j n}{2^j}\right]$$

where j is the scale factor and n is the translation index.

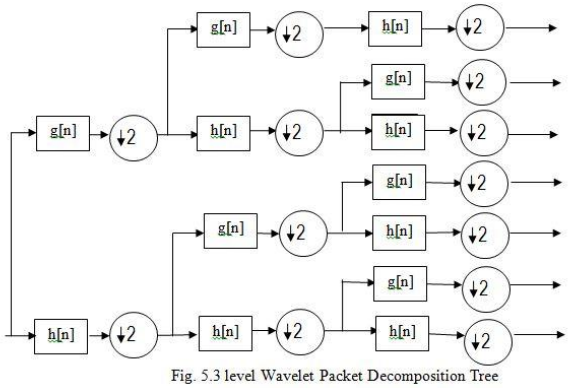

Fig. 5.3 level Wavelet Packet Decomposition Tree

Figure 3. Level 3 decomposition using wavelet packet decomposition.

3.3 Multiwavelets

, α≥0 and α ≠1.

Here X is a discrete random variable with possible outcomes 1, 2…n. α is the order, and when it equals to 1, it is Shannon entropy. An entirely homogeneous sample has the entropy of 0.Equally divided sample has the entropy of 1.

3.4 Multiwavelet Packets

Just as with scalar wavelets, the Multiwavelet filter bank procedure involves iterating the filtering operations on the low pass channel of the filter bank. Moreover, just as with scalar wavelets, iterating on the high pass channel as well can produce new basis function, this approach combines the wavelet packet decomposition with Multiwavelet filter, and hence we call it the Multiwavelet Packet in a manner analogous to the wavelet packet in the last section [6]

4. Feature Extraction Techniques

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 discussed below:

Mean

An Arithmetic Mean is amathematical representation of the typical value of a set of data, computed as the sum of all the numbers in the dataset divided by the size of the dataset. Suppose there is sample space {x1, x2, x3….xn} then the arithmetic mean m_x is defined as the means of the raw signals:

$$m_x = \frac{1}{N}\sum_{n=1}^{N} X_n$$

Standard Deviation

Standard deviation shows how much variation or dispersion exists from the average. A low standard deviation indicates that the data points tend to be very close to the mean, whereas the high standard deviation indicates that the data points are spread out over a large range of values. Let X is a random variable with mean value μ. Here the operator E denotes the average or expected value of X. Then the standard deviation of X is:

$$s_x = \frac{1}{N-1}\sum_{n=1}^{N}(X_n - m_x)$$

Variance

Variance is a statistical parameter which gives information about data distribution from its mean or expected value. It is the one type of probability distribution which measures how far a set of numbers get spread. The variance is calculated as:

$$s_i^2 = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} P_{i,j}(i - m_i)^2$$

$$s_j^2 = \sum_{j=0}^{N-1} \sum_{i=0}^{N-1} P_{i,j}(i - m_j)^2$$

Skewness

Skewness measures the symmetry of the data. If data distribution seems to be similar the same to the left and right of the center point, it shows symmetric property, and always symmetric data has skewness near zero whereas skewness for anormal distribution is zero. A negative value of skewness shows data that are

skewed left and positive values of the skewness show data that are skewed right. Skewness is represented by the formula:

$$S = \sum_{j=0}^{N-1} (i - m)^3 P(i)$$

Entropy

Entropy is a numerical measure of the randomness of a signal. Entropy can act as a feature and is used to analyze time series data such as speech signal. The Entropy is the statistical descriptor of the variability within the speech signal and is a strong feature for emotion classification. It can be mathematically represented as

$$e = \sum_{1}^{n} x^2 \log(x^2)$$

The simplest and the most common approach use histogram-based estimation, but other approaches have been developed and used, each with its benefits and drawbacks. The main factor in choosing a method is often a trade-off between the bias and the variance of the estimate although the nature of the (suspected) distribution of the data may also be a factor.

Power

Power is Measure of the amplitude of speech signal, and the power can be calculated as:

$$Pow = \sum \frac{X^2}{L(X)}$$

Where, X = is the signal values and L(x) = Length of the signal.

Power is also defined as the amount of energy consumed per unit time. The unit of power is the joule per second (J/s), known as the watt. Energy transfer can be used to do work, so power is the rate at which this work is performed.

Root Mean Square (RMS) Value

The RMS value of a set of values is the square root of the arithmetic mean (average) of the squares of the original values (or the square of the function that defines the continuous waveform). In the case of a set of n values{x1, x2, x3….xn}, the RMS value is given by the formula:

$$X_{RMS} = \sqrt{\frac{1}{n}(X_1^2 + X_2^2 + \ldots + X_n^2)}$$

Smoothness

Smoothness is a measure of the rhythmic pattern of acceleration and deceleration of signal. The measure of smoothness is calculated as:

$$R = 1 - \frac{1}{1 + S^2}$$

Where the bounded measure 0 <= S^2 <= 1.

5. Back Propagation Neural Network (BPNN) Classifier

Recently, neural network-based methods have shown tremendous success in speech processing and data mining tasks. A neural network model is a powerful tool used to perform keyword spotting tasks as performed by thehuman brain. Among a number of neural networks, the Multi-Layer Perceptron (MLP) with back-propagation (BP) neural network algorithm is found to be effective for solving some real world problems.

5.1 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: aninput layer, anoutput layer and the intermediate layer i.e. the hidden layer [2]. These layers comprise of the neurons which are connected to form the entire network. Weights are assignedto the connections which mark the signal strength. The weight values are computed based on the input signal, and the error function back propagated to the layer of entry. 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. The forward path contains creating a feed-forward network, initializing weight, simulation and training the network. The network weights and biases are updatedin backward path. The neural network model is shown in figure 2.

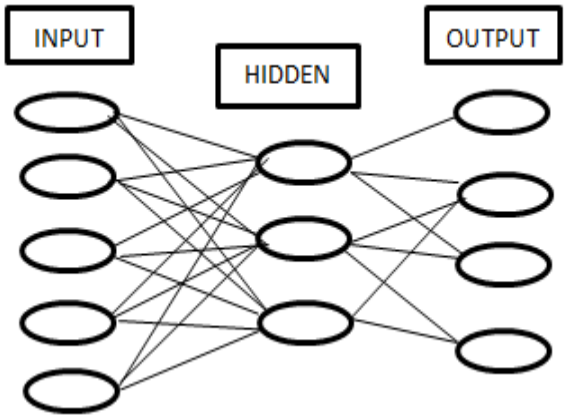

Figure 4. Neural network model

5.2 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 feedforward 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.

6. Framework of Keyword Spotting System.

6.1 Preprocessing

The speechsignal is pre-processing covers digital filtering, to enhance the speech quality regarding silence removal, noise reduction, resampling, and segmentation. In this proposed system moving – normal 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 anaverage of samples for filtering the noise from thesignal. The preprocessed output is then passed to the next stage wavelet decomposition.

6.2 Statistical Features Extraction

The various wavelet transforms like DWT, WPD, Multiwavelet and Multiwavelet Packets are applied to transform the uttered speech and test keyword to acquire its frequency domain spectrum and filter out unwanted frequencies. 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.

A lot of statistical features like RMS (Root Mean Square level), Correlation, Homogeneity, Standard Deviation, Variance, Smoothness, Kurtosis, Skewness are extracted. The keyword features are trained with BPNN and speech input features are tested by splitting the speech into keyword sized blocks using sliding window method.

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.

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 thekeyword. Sliding frame method is used splitting the speech stream [8]. In this process, initially, a block of frames such that the number of frames in the block is equal to some frames of the keyword signal are selected from the input signal starting from the first frame. This block of feature vectors then matched in neural network classifier with the trained keyword. The process repeatedfor the next block of frames in the input speech.

7. Experimental Results

In this experiment, first the speech signal is preprocessed, and statistical features are extracted based on DWT, WPD, Multiwavelet and Multiwavelet Packets decomposition. Multi-resolution features of the speech signal can easily be extracted using the wavelet transformations. The signals are decomposed at level 3 using Haar, db7 and simplest wavelets which are suitable for the problem of keyword spotting.

Sliding frame method is used to split the speech stream into blocks of the size of keyword size[4]. That is, the block of frames of speech splitintothe number of frames of the keyword signal. This block of feature vectors then matched in neural network classifier with the trained keyword. The process is repeatedfor the next block of frames in the input speech.

Figure 4 shows the waveform of the Tamil word "Ondru". Figure 5 shows the wavelet packet decomposition of the same word. The table I shows the selected keywords for training the neural network BPNN. Table II shows the accuracies of the spotting of the keyword "ONDRU" with various wavelet transformations.

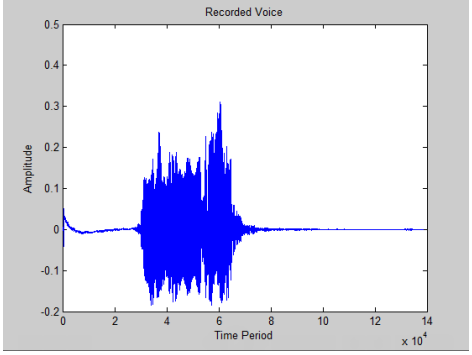

Figure5. Recorded keyword "ONDRU"

TABLE I.Keywords trained on BPNN

Ondru(1)

Irandu (2)

Mundru (3)

TABLE II. Results of KWS with various wavelet transformations

Input Speech Files	Accuracy in %											
	DWT			WPD			Multiwavelets			Multiwavelet Packets		
	Haar	Db6	Sim4	Haar	Db5	Sim4	Haar	Db6	Sim4	Haar	Db6	Sim4
Speech1. Wav	92.8	93.2	93.7	92.5	94.2	93.4	92	93.2	94.7	92.8	94.2	93.7
Speech2. Wav	87.9	93	92.4	87.9	93	92.4	87.9	93	92.4	87.9	93	94.4
Speech3. Wav	90	92.2	93.4	90	92	93.4	90	92.2	93	90	93.2	90.4
Speech4. Wav	90.3	91.5	93.2	90.3	91.5	93.7	90.5	91.5	92.2	90.3	92	93.2
Speech5. Wav	92.3	90.8	92.2	92.3	90.8	92	92.8	90.8	93.2	92.3	92.8	92

7. Conclusion

This paper presents the keyword spotting system using various wavelet transformations and BPNN neural network. Features have been extracted based on DWT, WDP, multiwavelets and multiwavelet packets transformations. The performance of multiwavelet packet transformation with neural network is appreciable while comparing with the other transformation methods since multiwavelet packet analysis provides a more precise frequency resolution than the wavelet analysis. It also has small support in time as well as in frequency domain and adapts its support locally to the signal which is relevant in time different signal.

References

1. Bahi, H., and Benati, N., [2009], "A new keyword spotting approach", IEEE International Conference on Multimedia Computing and Systems, pp. 77-80.
2. Chui CK, [1992], "An introduction to wavelets". Academic, New York.
3. P. Flandrin. Representation temps-fréquence. Hermes, 1993
4. Fourier J, [1822], "The analytical theory of heat. (trans: Freeman A)", Cambridge University Press, London, pp. 1878
5. 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.
6. Jo Yew Tham, Lixin Shen, Seng Luan Lee and Hwee Huat Tan (2000) "A General Approach for analysis and application of Discrete Multiwavelet Transforms", IEEE Transaction on Signal Processing ,48(2), 457- 464.
7. Jothilakshmi, S., Spoken keyword detection using autoassociative neural networks, International Journal Speech Technology, Springer, 2013, pp. 83-89.
8. Khan, W. and Holton, R., Word spotting in continuous speech using wavelet transform, IEEE International Conference on Electro/Information Technology, 2014, pp.275-279
9. Mallat S., [1999], "A wavelet Tour of Signal Processing", Academic Press, New York, 1999.
10. Qian S, [2002], "Time-frequency and wavelet transforms", Prentice Hall PTR, Upper Saddle, River, NJ.
11. 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.
12. 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.
13. 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.
14. Senthil devi K A., Dr.Srinivasan B., "Wavelet – Neural Network Approach for keyword spotting in Speech Mining", International Journal of Trends and Technologies", Vol 43, Issue 3, pp 160-165, 2017.
15. Tao Li, Sheng Ma, Mitsunori Ogihara, " Wavelet methods in data mining", Chapter 27 Data Mining and Knowledge Discovery Handbook ,Springer, 2005, pp 603-626.

