Global Journal of Enterprise Information System

DOI: 10.18311/gjeis/2017/16120

Speech Recognition using Wavelet based Feature Extraction Techniques

Pardeep Sangwan[1*], Dinesh Sheoran[1] and Saurabh Bhardwaj[2]

[1]Department of ECE, Maharaja Surajmal Institute of Technology, Delhi, India; sangwanpardeep@gmail.com

[2]Department of ICE, Thapar University, Patiala, Punjab, India ; bsaurabh2078@gmail.com

Abstract

Speech recognition by machine may be defined as the conversion of human speech signal into textual form automatically by the machine without any human intervention. Two feature extraction techniques utilizing DWT (Discrete Wavelet Transform) and WPD (Wavelet Packet Decomposition) for speech recognition are discussed in the present article. The comparison of two speech recognizer, first, based on Discrete Wavelet Transform and the second based on Wavelet Packet Decomposition, and with four classifiers has been done in this paper. The proposed method is implemented for a database consisting of ten digits and two hundred speakers, making it a database of 2000 speech samples. The results present the accuracy rate of the respective speech recognizers.

Keywords: ANN, DWT, Feature Extraction, HMM, Speech Recognition, Wavelet-Packet Decomposition

Paper Code: 16120; **Originality Test Ratio:** 7%; **Submission Online:** 01-March-2017; **Manuscript Accepted:** 12-March-2017; **Originality Check:** 18-Mar-2017; **Peer Reviewers Comment:** 04-April-2017; **Double Blind Reviewers Comment:** 14-May-2017; **Author Revert:** 19-May-2017; **Camera-Ready-Copy:** 22-May-2017

1. Introduction

Speech is a source of oral communication in which humans express their ideas, emotions and thoughts with each other. Human speech is a complex signal contains rich information. Thus to understand this complex information by computer is known as speech recognition. Developing an ASR (Automatic Speech Recognizer) consists of two major parts one known as front-end processor and other known as back-end processor. Former is utilized for building individual templates of all speech signals and latter performs the task of template matching.[1,2] Figure 1 shows the basic block diagram for an ASR system.

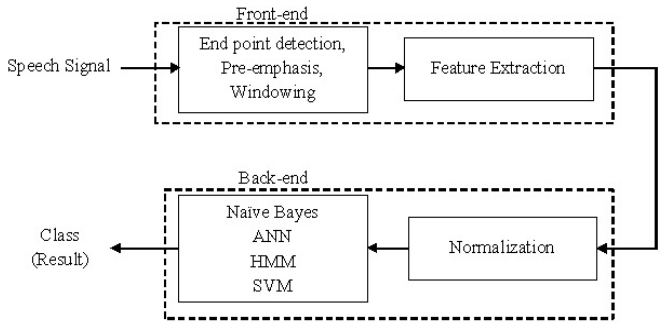

Figure 1. Block diagram for automatic speech recognition system.

Front-end processor performs several tasks, like, end-point detection, pre-emphasis, windowing, framing etc., known as pre-processing of speech signal, and also extracting features from this pre-processed speech. Back-end processor compares individual template generated by front-end processor with the template of unknown speech signal generated by the same front-end processor.[3] Features are the parametric representation of speech signal, which can be utilized for grouping same kind of patterns.[4] MFCC (Mel-Frequency Cepstral Coefficient), LPC (Linear Predictive Coding), PLP (Perceptual Linear Predictive) coefficient, wavelet and auditory features are the generally used features. Despite of availability of several feature extraction techniques, the efficiency of a speech recognition system largely depends upon selection of appropriate features.[4] Various parameters can deteriorate the efficiency of the system. A very efficient method for speech recognizing application is wavelet transformation.

In this paper, Section-2 presents two wavelets based feature extraction techniques. Section-3 provides the result and discussions and lastly, Section-4 presents conclusion and future scope of the paper.

2. Wavelet based Feature Extraction Techniques

The general architecture of automatic speech recognition system to recognize voice signal involves different stages and is shown in

*Author for correspondence

Figure 1. The sound signal captured by microphone is preprocessed for making it compatible, noise robust and appropriate to extract features from the sound signal. In the present work, four tasks are performed to pre-process the sound signals: detecting end-point of sound, Pre-emphasizing the sound signal, Framing and Windowing. After that, extracting features is the next step.

Most of the speech-based studies are based on Fourier Transforms (FTs), Short Time Fourier Transforms (STFTs), Mel-Frequency Cepstral coefficients (MFCCs), Linear Predictive Coding (LPCs), and prosodic parameters. Various researches have used MFCC, LPC, PLP etc. for speech processing but size of feature vectors and thus, the computing costs are very high for these parameters. Wavelet transform can reduce size of feature vectors in comparison to these techniques and hence, reduces the computing costs several folds.

2.1 Discrete Wavelet Transform (DWT)

Wavelet transform is a technique for decomposing the signals in a set of basic functions, which are known as wavelet. DWT, a specific type of WT, is a tool to represent signal in time and frequency in a compact manner for efficient computation. DWT can more efficiently process the sound signal due to its efficient time-frequency localization[5] along with multi resolution, multi scale analyzing features.

The DWT can be represented as:

$$W(j,k) = \sum_{j}\sum_{k} X(k) 2^{-j/2} \varphi(2^{-j} n - k)$$

Where $\varphi(t)$ is basic analysis function of "mother wavelet". Further functions can be obtained translating and dilating the mother wavelet. In DWT, sound signal is passed from low-pass and high-pass filter, the outputs of which are known as approximation and detailed coefficients respectively.[6] In sound signal, low-frequency component termed as approximation "h[n]" is more important as compared to high-frequency signal termed as detailed "g[n]" because low-frequency component characterizes a sound signal as compared to its high-frequency component.[7] The successive coefficients can be given by:

$$Y_{low}[k] = \sum_{n} x[n] h[2k - n]$$

$$Y_{high}[k] = \sum_{n} x[n] g[2k - n]$$

Where, Y_{low} (approximation coefficient) and Y_{high} (detail coefficient) are the output of low-pass and high-pass filters respectively after sub-sampling with two. This process is performed till the achievement of desired level.[8] WT has a variable window-size, which is broad at low-frequencies and narrow at high-frequencies. Due to this, the output is an optimized time-frequency resolution for all frequencies.[9] Figure 2 shows the wavelet tree decomposition of speech signal up to 5 levels.

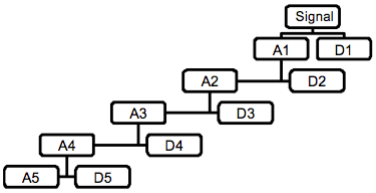

Figure 2. Wavelet tree decomposition for 5 levels.

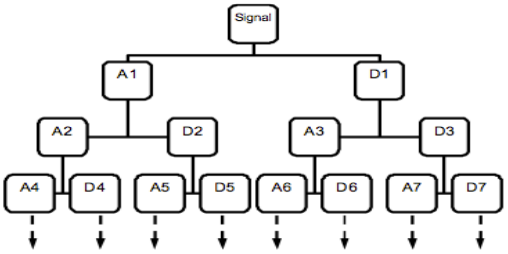

Figure 3. Wavelet Packet Decomposition (A_i & D_i – Approximation & Detailed co-efficients at ith level).

2.2 Wavelet Packet Decomposition (WPD)

This transformation is an expansion of DWT. WPD also provides a multi level time-frequency division of sound signal. The A_i and D_i are further decomposed into low-resolution approximation and detailed co-efficient. DWT is performed to low-pass filter output only while WPD is applied to low-pass and high-pass filter outputs both. Hence, WPD utilizes long time intervals for lower frequencies and short time intervals for higher frequencies.[10] In WPD, every detailed co-efficient is decomposed in two parts. Approximation co-efficient consist properties of a sound signal and detailed co-efficient consist noise and disturbance present in the sound signal.[11] Although, majority of signal characteristics are contained in approximation co-efficient but sometimes detail co-efficient may also have useful characteristics of speech signal. So, both A_i and D_i are decomposed in WPD. Figure 3 shows the WPD of speech signal. The classifiers used are Naive Bayes, ANN, HMM and SVM.

3. Result and Discussion

Various experiments performed to develop a speech recognition system using wavelet are presented here. Table 1 provides the performance evaluation results to select optimal wavelet.

Figure 4 provides the comparison of recognition accuracy for various wavelets over the digits database. From the experiments conducted, it is concluded that result with higher accuracy

were achieved with db1 or haar. Sym4 gave optimal results for symlets and Coif4 for coiflets. The Daubechies wavelet performs better than others with accuracy rate of 87.0% using db1 or Haar wavelet. Db4 wavelets also demonstrated a high accuracy rate of 86.9% with digit database. Hence, db1 or Haar wavelet is selected and decomposition is done till 8 levels. The originally recorded sound signal and its end point detected speech are shown in Fig 5. The first level and 8th level DWT decomposition of speech signal with approximation and details coefficients are provided in Figure 6 and Figure 7 respectively.

Table 1. Selection of Optimal Wavelet

Wavelet	Accuracy	Wavelet	Accuracy	Wavelet	Accuracy
db1 or Haar	87.0	sym2	79.7	coif1	75.4
db2	84.3	sym3	80.1	coif2	73.4
db4	86.9	sym4	82.0	coif3	76.1
db6	84.7	sym5	78.4	coif4	76.4
db8	83.4	sym6	78.3	coif5	74.2
db10	82.8	sym7	80.3		
db12	81.7	sym8	80.4		
db14	78.4				
db16	81.2				
db18	76.9				
db20	79.4				

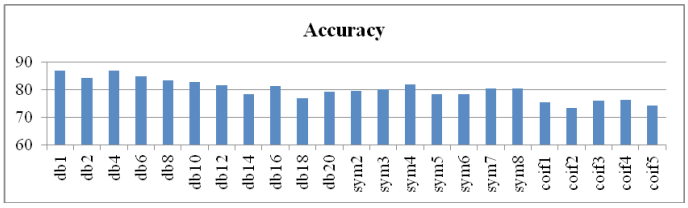

Figure 4. Recognition accuracy for digits database.

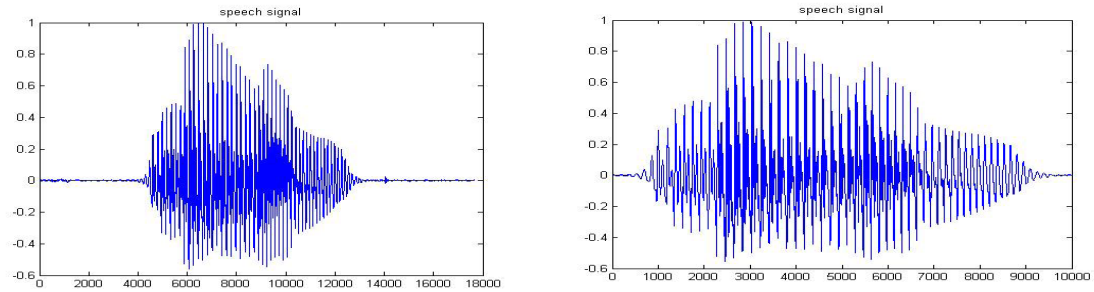

Figure 5. Original Speech signal and End point detected speech.

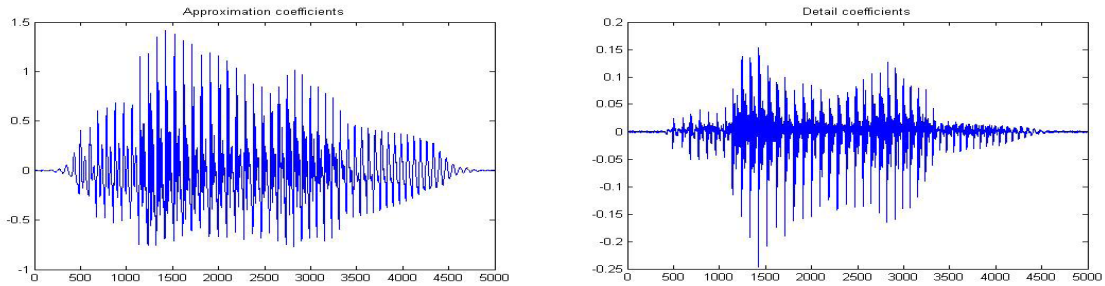

Figure 6. DWT decomposition for level 1 (Approximation and Detail Coefficients).

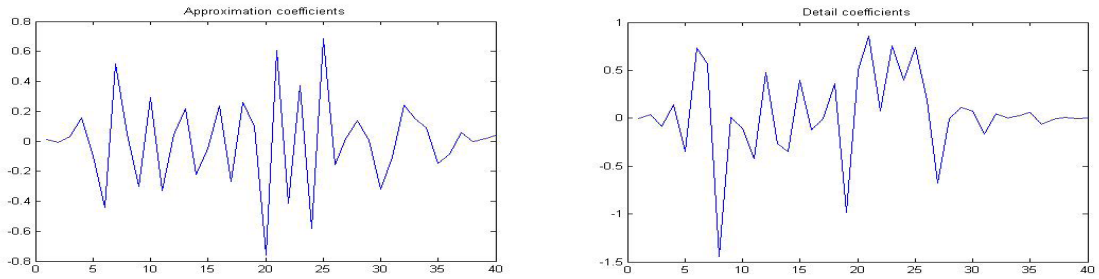

Figure 7. DWT decomposition for level 8 (Approximation and Detail Coefficients).

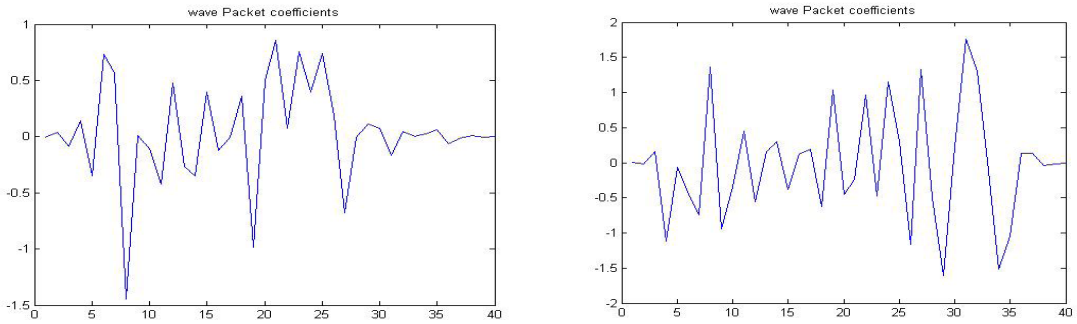

Figure 8. Wavelet Packet Decomposition for Digit 9.

The selected database has ten digits and two-hundred speakers making it 2000 samples in all. Here, 16 experiments were conducted. Following the front-end processing, resultant feature vectors are normalized and given as input to classifiers. The number of speakers is varied during the course of experiment for determining the effect of this factor on accuracy rate. Generally, smaller database with few numbers of speakers results in high recognition rate. Table 2 shows the results of DWT based features of digits database after classification using four different classifiers.

The results for WPD and four classifiers on digit database along with comparison graph are presented in Table 3 and figure 10.

Figure 9 and Figure 10 provides the comparison between classification techniques of wavelet and wavelet packet based features for digits based speech respectively. From the presented results it is observed that both the methods perform well for speech digits database and DWT perform better than WPD. The ANN classifier performs better recognition rate. Table 4 below shows the comparison for performance for both the DWT and WPD.

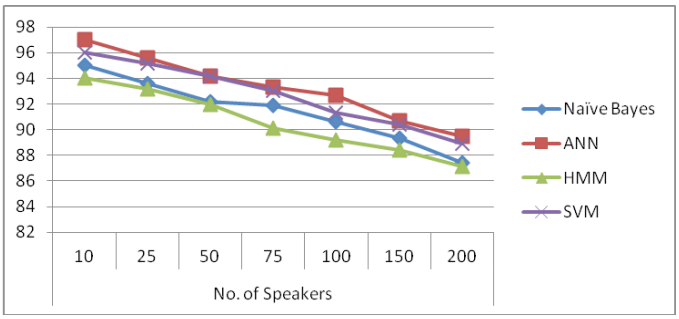

Figure 9. Recognition accuracy of DWT based features for speech recognition.

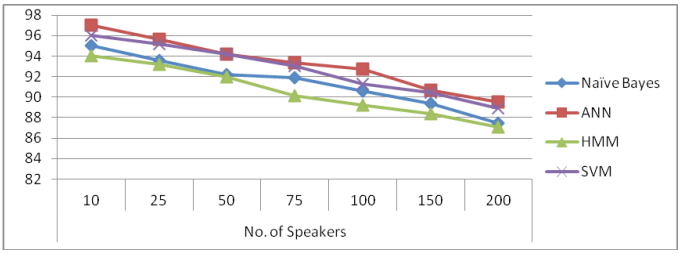

Figure 10. Recognition accuracy of wavelet packet features for speech recognition.

Table 2. Results for wavelets based features of digits database with four classifiers

No. of Speakers	Total Samples	Naïve Bayes		ANN		HMM		SVM	
		Correct	Accuracy	Correct	Accuracy	Correct	Accuracy	Correct	Accuracy
10	100	95	95	97	97	96	96	97	97
25	250	237	94.8	245	98	238	95.2	242	96.8
50	500	468	93.6	480	96	469	93.8	476	95.2
75	750	699	93.2	705	94	692	92.26667	709	94.53333
100	1000	920	92	936	93.6	918	91.8	932	93.2
150	1500	1365	91	1395	93	1352	90.13333	1377	91.8
200	2000	1782	89.1	1821	91.05	1769	88.45	1800	90

Table 3. Classification results for wavelet packet based features of digits database

No. of Speakers	Total Samples	Naïve Bayes		ANN		HMM		SVM	
		Correct	Accuracy	Correct	Accuracy	Correct	Accuracy	Correct	Accuracy
10	100	95	95	98	98	93	93	97	97
25	250	235	94	241	96.4	232	92.8	238	95.2
50	500	466	93.2	477	95.4	463	92.6	472	94.4
75	750	697	92.93333	709	94.53333	687	91.6	701	93.46667
100	1000	909	90.9	937	93.7	908	90.8	924	92.4
150	1500	1353	90.2	1384	92.26667	1350	90	1365	91
200	2000	1775	88.75	1812	90.6	1774	88.7	1795	89.75

Table 4. Comparison of performance for both DWT and WPD

Feature Extraction	Accuracy	Precision	Recall
Wavelet	93.48	0.91	0.894
Wavelet Packet	92.8	0.897	0.877

4. Conclusion and Future Scope

Based on experimental results, it is observed that DWT performs better than WPD and accuracy decreases with the increase in number of speakers. Moreover, the results show that ANN has given highest accuracy rate among the four chosen classifiers. So, the best results are obtained by using DWT for feature extraction followed by ANN as classifier. One extension of the current work can be enhancing the database of speech samples for higher accuracy rate. Also, other classifiers such as Genetic algorithm, Fuzzy logic etc. can also be utilized for analyzing the performance of the speech recognition system.

5. References

1. Rabiner LR, Juang BH. Fundamentals of speech recognition. 1993
2. Quatieri TF. Discrete-Time Speech Signal Processing Principles and Practice. Upper Saddle River, NJ: Prentice-Hall; 2001.
3. Kumar K, Aggarwal RK. Hindi speech recognition system using Htk. IJCBR. 2011; 2(2).
4. Hai J, Joo EM. Improved linear predictive coding method for speech recognition. Proc. ICICSP. 2003; 1614–8.
5. Mallat S. A Wavelet Tour of Signal Processing. U.S.: Academic Press; 2008.
6. Chan Woo S, Lin CP, Osman R. Development of a speaker recognition system using wavelets and artificial neural networks. Proc. ISIMVSP. Hong Kong. 2001; 413–6. https://doi.org/10.1109/isimp.2001.925421
7. Kadambe S, Srinivasan P. Application of adaptive wavelets for speech. Optical Engineering. 1994; 33(7):2204–11. https://doi.org/10.1117/12.172410
8. Mallat SG. A theory for multi resolution signal decomposition: the wavelet representation. IEEE Trans PAMI. 1989; 11:674–93. https://doi.org/10.1109/34.192463
9. Ubeyil ED. Combined neural network model employing wavelet coefficients for ECG signals classification. DSP. 2009; 19(2):297–308.
10. Ting W, Zheng YG, Bang-Hua Y, Hong S. EEG feature extraction based on wavelet packet decomposition for brain computer interface. Measurement. 2008; 41(6):618–25. https://doi.org/10.1016/j.measurement.2007.07.007
11. Li BC, Luo JS. Wavelet Analysis and its Applications. Beijing: Electronics Engineering Press; 2003 https://doi.org/10.1142/5251

Citation:

Pardeep Sangwan, Dinesh Sheoran and Saurabh Bhardwaj
"**Speech Recognition using Wavelet based Feature Extraction Techniques**",
Global Journal of Enterprise Information System. Volume-9, Issue-2, April-June, 2017. (http://informaticsjournals.com/index.php/gjeis)

Conflict of Interest:

Author of a Paper had no conflict neither financially nor academically.