Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 3, No. 2, March : 2023

AN ANALYSIS OF LPC, RASTAT AND DWT TECHNIQUES IN AUTOMATIC SPEECH RECOGNITION SYSTEM

BANOTH SUSHMITHA, PG Scholar, Dept. of ECE, JNTUH College of Engineering, Jagtial, Telangana, India 505501
Dr. D. NAGA SUDHA, Associate Professor, Dept. of ECE, JNTUH College of Engineering, Jagtial, Telangana, India 505501

Abstract
In an effort to provide a more efficient representation of the speech signal, the application of the wavelet analysis is considered. This research presents an effective and robust method for extracting features for speech processing. Here, this work proposed a new improved method for text dependent speaker recognition system using combination of discrete wavelet transform (DWT) and relative spectra algorithm and linear predictive coding (RASTA-LPC). First, we will apply the proposed techniques to the training speech signals and then form a train feature vector which contains that Wavelet and RASTA-LPC coefficients. Afterwards, the same process will be applied to the testing speech signals and will form a test feature vector. Now, we will compare the two feature vectors by calculating the Euclidean distance between the vectors to identify the speech and speaker. If the distance between two vectors is near to zero, then the tested speech/speaker will be matched with the trained speech/speaker. Simulation results have been compared with LPC scheme and shown that the proposed scheme has performed superior to the existing technique, and the verification tests have been carried and an accuracy rate of approximately 90% has been achieved.

Keywords: Speech recognition, speaker recognition, linear predictive coding, relative spectral algorithm, wavelet decomposition.

1. INTRODUCTION

In our everyday lives the audio signal especially the voice signal has become one of the major part, because it can be used as a one of the major tool for communicating each other. However, by using some software applications, the voice signal has modified or processed further due to its technological advancement, and can be utilized in various applications such as security applications. In many applications these speech processing systems plays a vital role such as speech recognition, voice communication. Speech recognition is the process of automatically extracting and determining linguistic information conveyed by a speech signal using computers or electronic circuits. Automatic speech recognition methods, investigated for many years have been principally aimed at realizing transcription and human computer interaction systems. The first technical paper to appear on speech recognition has since then intensified the researches in this field, and speech recognizers for communicating with machines through speech have recently been constructed, although they remain only of limited use. Automatic speech recognition (ASR) features some of the following advantages:

- Speech input is easy to perform because it does not require a specialized skill as does typing or pushbutton operations.
- Information can be input even when the user is moving or doing other activities involving the hands, legs, eyes, or ears.
- Since a microphone or telephone can be used as an input terminal, inputting information is economical with remote inputting capable of being accomplished over existing telephone networks and the Internet.

However, the task of ASR is difficult because:

- Lot of redundancy is present in the speech signal that makes discriminating between the classes difficult.

- Presence of temporal and frequency variability such as intra speaker variability in pronunciation of words and phonemes as well as inter speaker variability e.g. the effect of regional dialects.
- Context dependent pronunciation of the phonemes (co-articulation).
- Signal degradation due to additive and convolution noise present in the background or in the channel.
- Signal distortion due to non-ideal channel characteristic.

2. SPEECH RECOGNITION

Most speech recognition systems can be classified according to the following categories:

2.1 Speaker Dependent vs. Speaker Independent

A speaker-dependent speech recognition system is one that is trained to recognize the speech of only one speaker. Such systems are custom built for just a single person, and are hence not commercially viable. Conversely, a speaker-independent system is one that is independence is hard to achieve, as speech recognition systems tend to become attuned to the speakers they are trained on, resulting in error rates that are higher than speaker dependent systems.

2.2 Isolated vs. Continuous

In isolated speech, the speaker pauses momentarily between every word, while in continuous speech the speaker speaks in a continuous and possibly long stream, with little or no breaks in between. Isolated speech recognition systems are easy to build, as it is trivial to determine where one word ends and another starts, and each word tends to be more cleanly and clearly spoken. Words spoken in continuous speech on the other hand are subjected to the co-articulation effect, in which the pronunciation of a word is modified by the words surrounding it. This makes training a speech system difficult, as there may be many inconsistent pronunciations for the same word.

3. WAVELET ANALYSIS

The basic idea of this proposal is to use wavelets as a mean of extracting features from a voice signal. The wavelet technique is considered a relatively new technique in the field of signal processing compared to other methods or techniques currently employed in this field. Fourier Transform (FT) and Short Term Fourier Transform (STFT) [1] [2] are the current methods used in the field of signal processing. However due to severe limitations imposed by both the Fourier Transform and Short Term Fourier Transform in analyzing signals deems them ineffective in analyzing complex and dynamic signals such as the voice signal [3][4]. In order to substitute the shortcomings imposed by both the common signal processing methods, the wavelet signal processing technique is used. The wavelet technique is used to extract the features in the voice signal by processing data at different scales. The wavelet technique manipulates the scales to give a higher correlation in detecting the various frequency components in the signal. These features are then further processed in order to construct the voice recognition system. Extracting the features of the voice signal does not limit the capabilities of this technique to a particular application alone, but it opens the door to a wide range of possibilities as different applications can benefit from the voice extracted features. Applications such as speech recognition system, speech to text translators, and voice based security system are some of the future systems that can be developed.

3.1 Fourier transform

The signal can be analyzed more effectively in frequency domain than the time domain, because the characteristics of a signal will be more in frequency domain. One possible way to convert or transform the signal from time to frequency domain is Fourier transform (FT). FT is an approach which breaks down the signal into different frequencies of sinusoids and it is defined as a mathematical approach for transforming the signal from time domain to frequency domain.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 3, No. 2, March : 2023

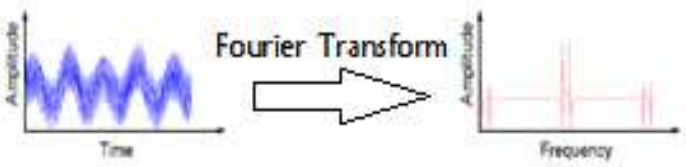


Fig. 1. Analysis of FT with an example

FT has a drawback that it will work out for only stationary signals, which will not vary with the time period. Because, the FT applied for the entire signal but not segments of a signal, if we consider non-stationary signal the signal will vary with the time period, which could not be transformed by FT. and one more drawback that we have with the FT is we cannot say that at what time the particular event will has occurred.

3.2 *Short-time fourier analysis*

To correct the deficiency in FT, Dennis Gabor in 1946 introduced a new technique called windowing, which can be applied to the signal to analyze a small section of a signal. This adaptation has been called as the Short-Time Fourier Transform (STFT), in which the signal will be mapped into time and frequency information.

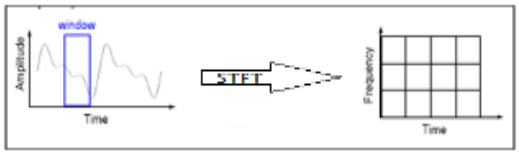


Fig. 2. STFT analysis of a signal

In STFT, the window is fixed. So, we this window will not change with the time period of the signal i.e., for both narrow resolution and wide resolution. And we cannot predict the frequency content at each time interval section. To overcome the drawbacks of STFT, a wavelet technique has been introduced with variable window size. Wavelet analysis allows the use of long time intervals where we want more precise low-frequency information, and shorter regions where we want high-frequency information.

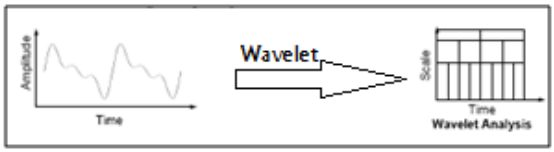


Fig. 3. Wavelet analysis with an example

In fig.4 it is shown that the comparison of FT, STFT and wavelet transform by considering an example input signal and how the analysis of transformation techniques will apply to get the frequency information of input signal. We can observe that in wavelet analysis the graphical representation shows that the wavelet has more number of features than the FT and STFT. Wavelet is also called as multi resolution analysis (MRA). Here's what this looks like in contrast with the time-based, frequency-based, and STFT views of a signal:

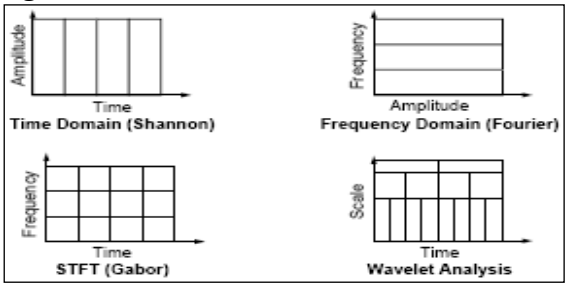


Fig. 4. Comparison of FT, STFT and Wavelet analysis of a signal

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 3, No. 2, March : 2023

3.3 Discrete wavelet transform

Discrete Wavelet Transform (DWT) is a revised version of Continuous Wavelet Transform (CWT). The DWT compensates for the huge amount of data generated by the CWT. The basic operation principles of DWT are similar to the CWT however the scales used by the wavelet and their positions are based upon powers of two. This is called the dyadic scales and positions as the term dyadic stands for the factor of two [9]. As in many real world applications, most of the important features of a signal lie in the low frequency section. For voice signals, the low frequency content is the section or the part of the signal that gives the signal its identity whereas the high frequency content can be considered as the part of the signal that gives nuance to the signal. This is similar to imparting flavor to the signal. For a voice signal, if the high frequency content is removed, the voice will sound different but the message can still be heard or conveyed. This is not true if the low frequency content of the signal is removed as what is being spoken cannot be heard except only for some random noise. The wavelet function is defined as follows:

$$W(\tau,s)=\frac{1}{\sqrt{s}}\int_{-\infty}^{\infty}x(t)\psi\left(\frac{t-\tau}{s}\right)dt$$

$$\int_{-\infty}^{\infty}\psi(t)dt=0$$

$$\int_{-\infty}^{\infty}\left|(\psi(t))^2\right|dt<\infty$$

The basic operation of the DWT is that the signal is passed through a series of high pass and low pass filter to obtain the high frequency and low frequency contents of the signal. The low frequency contents of the signal are called the approximations [10]. This means the approximations are obtained by using the high scale wavelets which corresponds to the low frequency. The high frequency components of the signal called the details are obtained by using the low scale wavelets which corresponds to the high frequency. From Figure 5, demonstrates the single level filtering using DWT. First the signal is fed into the wavelet filters. These wavelet filters comprises of both the high-pass and low-pass filter. Then, these filters will separate the high frequency content and low frequency content of the signal. However, with DWT the numbers of samples are reduced according to dyadic scale. This process is called the sub-sampling. Sub-sampling means reducing the samples by a given factor. Due to the disadvantages imposed by CWT which requires high processing power [11] the DWT is chosen due its simplicity and ease of operation in handling complex signals such as the voice signal. B. Wavelet Energy Whenever a signal is being decomposed using the wavelet decomposition method, there is a certain amount or percentage of energy being retained by both the approximation and the detail. This energy can be obtained from the wavelet bookkeeping vector and the wavelet decomposition vector. The energy calculated is a ratio as it compares the original signal and the decomposed signal.

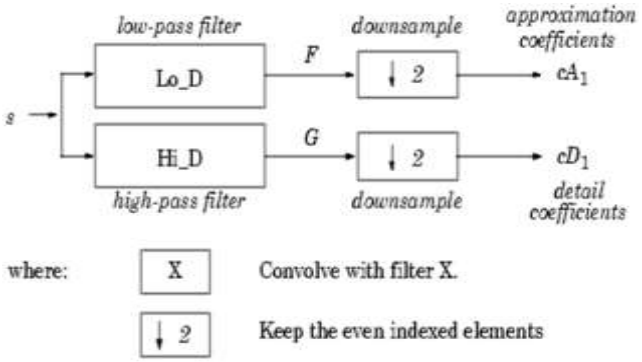


Fig. 5. Demonstration of single level wavelet decomposition

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 3, No. 2, March : 2023

4. EXISTING ALGORITHM

4.1 LPC algorithm

The LPC (Linear Predictive Coding) method is derived from the word linear prediction. Linear prediction as the term implies is a type of mathematical operation. This mathematical function which is used in discrete time signal estimates the future values based upon a linear function of previous samples [8].

$$\hat{x}(n) = -\sum_{l=1}^{P} a_l x(n-l)$$

$\hat{x}(n)$ is the predicted or estimated value and$x(n-l)$is the previous value. By expanding this equation

$$\hat{x}(n) = -[a_1 x(n-l) - a_2 x(n-l) - a_3 x(n-l) \ldots.$$

The LPC will analyze the signal by estimating or predicting the formants. Then, the formants effects are removed from the speech signal. The intensity and frequency of the remaining buzz is estimated. So by removing the formants from the voices signal will enable us to eliminate the resonance effect. This process is called inverse filtering. The remaining signal after the formant has been removed is called the residue. In order to estimate the formants, coefficients of the LPC are needed. The coefficients are estimated by taking the mean square error between the predicted signal and the original signal. By minimizing the error, the coefficients are detected with a higher accuracy and the formants of the voice signal are obtained.

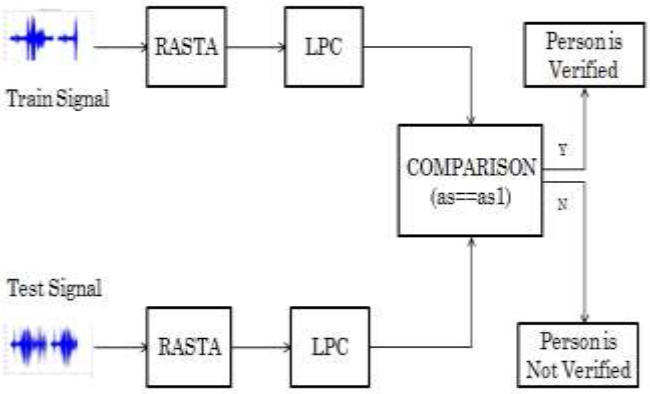


Fig. 6. Block diagram of LPC based recognition system

5. PROPOSED ALGORITHM

5.1 *RASTA*

RASTA or Relative Spectral Algorithm as it is known is a technique that is developed as the initial stage for voice recognition **[**13**]**. This method works by applying a band-pass filter to the energy in each frequency sub-band in order to smooth over short-term noise variations and to remove any constant offset. In voice signals, stationary noises are often detected. Stationary noises are noises that are present for the full period of a certain signal and does not have diminishing feature **[**14**]**. Their property does not change over time. The assumption that needs to be made is that the noise varies slowly with respect to speech. This makes the RASTA a perfect tool to be included in the initial stages of voice signal filtering to remove stationary noises **[**15**]**. The stationary noises that are identified are noises in the frequency range of 1Hz - 100Hz.

5.2 Formant estimation

Formant is one of the major components of speech. The frequencies at which the resonant peaks occur are called the formant frequencies or simply formants [12]. The formant of the signal can be obtained by analyzing the vocal tract frequency response. Figure 7 shows the vocal tract frequency response. The x-axis represents the frequency scale and the y-axis represents the magnitude of the signal. As it

can be seen, the formants of the signals are classified as F1, F2, F3 and F4. Typically a voice signal will contain three to five formants. But in most voice signals, up to four formants can be detected.
In Order to obtain the formant of the voice signals, the LPC (Linear Predictive Coding) method is used. The LPC (Linear Predictive Coding) method is derived from the word linear prediction. Linear prediction as the term implies is a type of mathematical operation.
This mathematical function which is used in discrete time signal estimates the future values based upon a linear function of previous samples [8].

5.3 RASTA-LPC and DWT implementation

In order to implement the system, a certain methodology is implemented by decomposing the voice signal to its approximation and detail. From the approximation and detail coefficients that are extracted, the methodology is implemented in order to carry out the recognition process. The proposed methodology for the recognition phase is the statistical calculation. Four different types of statistical calculations are carried out on the coefficients. The statistical calculations that are carried out are mean, standard deviation, variance and mean of absolute deviation. The wavelet that is used for the system is the symlet 7 wavelet as that this wavelet has a very close correlation with the voice signal. This is determined through numerous trial and errors. The coefficients that are extracted from the wavelet decomposition process is the second level coefficients as the level two coefficients contain most of the correlated data of the voice signal. The data at higher levels contains very little amount of data deeming it unusable for the recognition phase. Hence for initial system implementation, the level two coefficients are used.

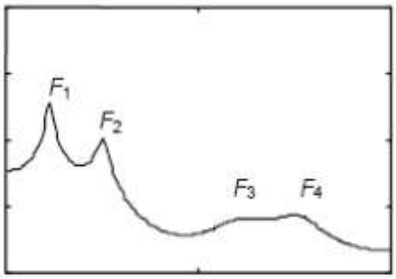


Fig. 7. Formant estimation

The coefficients are further threshold to remove the low correlation values, and using this coefficients statistical computation is carried out. The statistical computation of the coefficients is used in comparison of voice signal together with the formant estimation and the wavelet energy. All the extracted information acts like a 'fingerprint' for the voice signals. The percentage of verification is calculated by comparing the current values signal values against the registered voice signal values. The percentage of verification is given by:

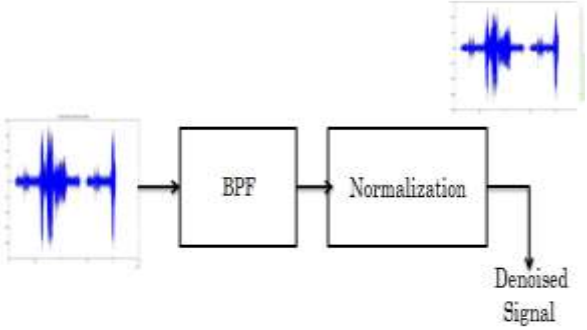


Fig. 8. Block diagram of RASTA process.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 3, No. 2, March : 2023

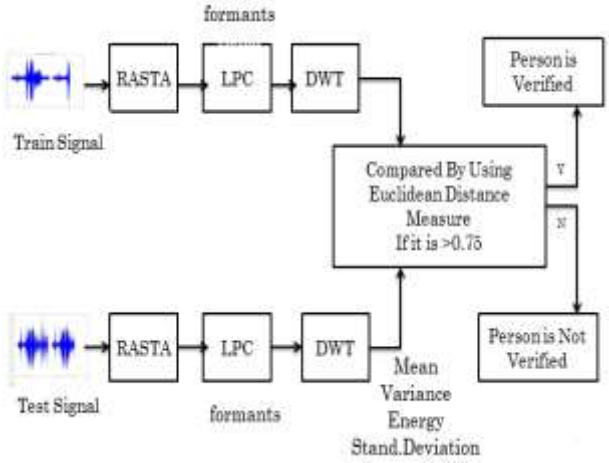


Fig. 9. Block diagram of proposed text dependent speaker identification system

Verification % = (Test value / Registered value) x100.

Between the tested and registered value, whichever value is higher is taken as the denominator and the lower value is taken as the numerator. Figure 9 shows the complete flowchart which includes all the important system components that are used in the voice verification program.

6. SIMULATION RESULTS

In this section, experimental results have been shown for various voice test signals with LPC and proposed algorithms. All the experiments have been done in MATLAB 2011a version with 4GB RAM and i3 processor for speed specifications.

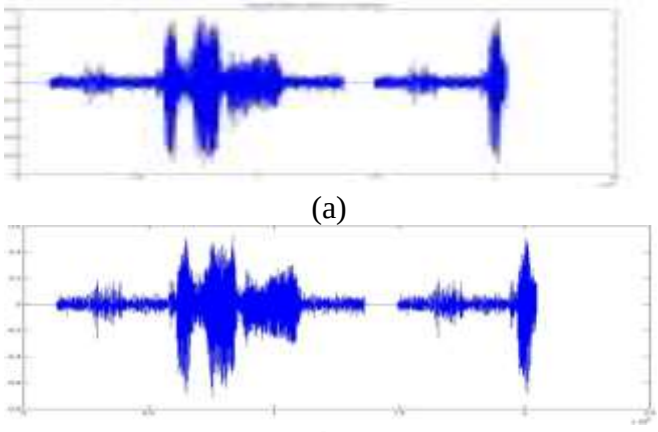
Mean = $\frac{1}{n}\sum_{i=1}^{n} x_i$

Std. deviation = $\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu)\wedge 2}$

Variance = $\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu)\wedge 2$

Energy = $T\sum_{i=1}^{n} x\wedge 2(i)$

Table 1 and Table 2 has shown the performance comparison of proposed and LPC in terms of recognition accuracy with statistical parameters. Finally, LPC achieved 66.66% accuracy where the proposed algorithm achieved almost 90% accuracy.

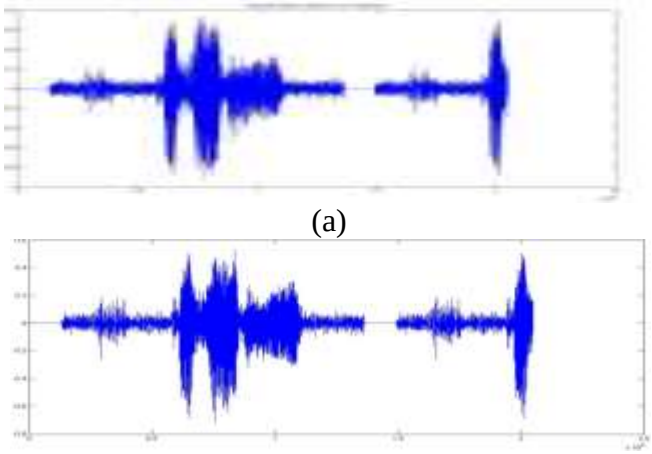


(a)

(b)

Fig. 10. (a) Original voice signal (b) De-noised signal for training

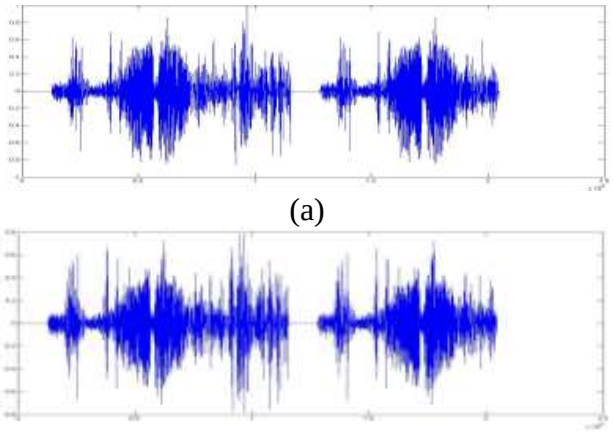
Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 3, No. 2, March : 2023



(a)

(b)

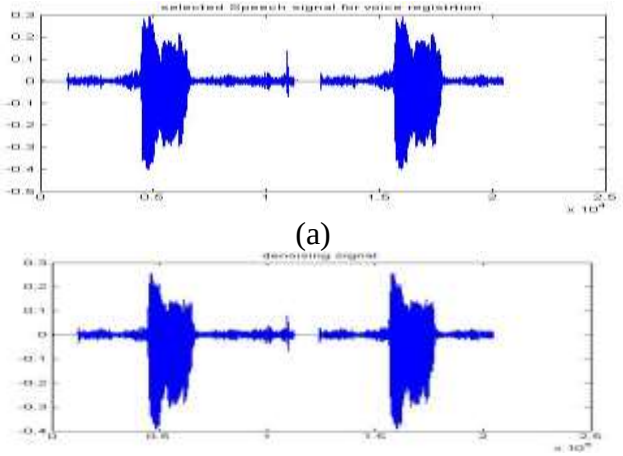
Fig. 11. (a) Original voice signal (b) De-noised signal for testing



(a)

(b)

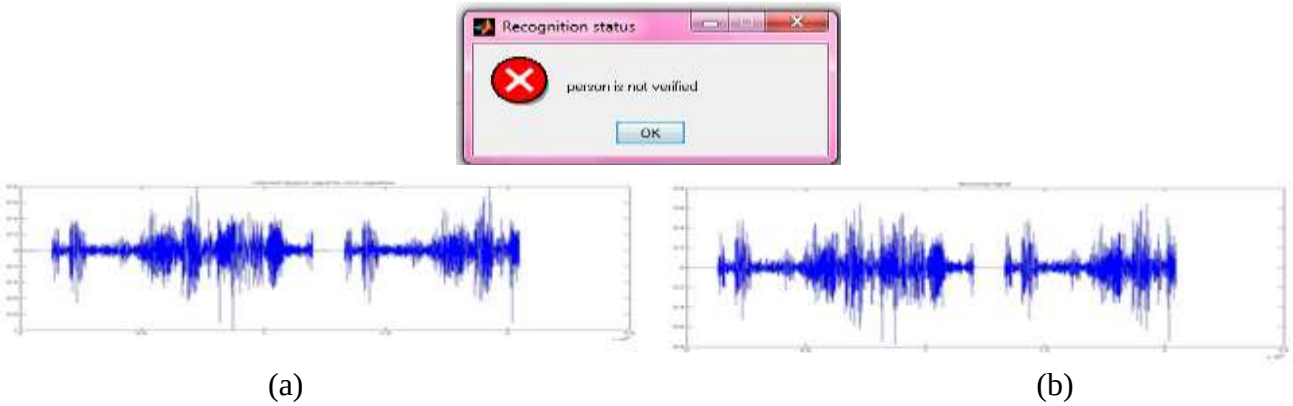
Fig. 12. (a) Original voice signal (b) De-noised signal for training



(a)

(b)

Fig. 13. (a) Original voice signal (b) De-noised signal for testing



(a) (b)

Fig14. (a) Original voice signal (b) De-noised signal for training

7. CONCLUSIONS

Text dependent Speaker Recognition system used to verify the identity of an individual based on their own speech signal using the statistical computation, formant estimation and wavelet energy. By using the fifty preloaded voice signals from six individuals, the verification tests have been carried and an accuracy rate of approximately 90 % has been achieved by proposed algorithm where the LPC has achieved only 66.66%. By observing the simulation results on various speech signals with different speaker we can conclude that the proposed algorithm accuracy has been improved when compared to LPC.

REFERENCES

[1] Soontorn Oraintara, Ying-Jui Chen Et.al. IEEE Transactions on Signal Processing, IFFT, Vol. 50,No. 3, March 2002
[2] Kelly Wong, Journal of Undergraduate Research, The Role of the Fourier Transform in Time-Scale Modification, University of Florida, Vol 2, Issue 11 - August 2011
[3] Bao Liu, Sherman Riemenschneider, An Adaptive Time Frequency Representation and Its Fast Implementation, Department of Mathematics, West Virginia University
[4] Viswanath Ganapathy, Ranjeet K. Patro, Chandrasekhara Thejaswi, Manik Raina, Subhas K. Ghosh, Signal Separation using Time Frequency Representation, Honeywell Technology Solutions Laboratory
[5] Amara Graps, An Introduction to Wavelets, Istituto di Fisica dello Spazio Interplanetario, CNR-ARTOV
[6] Brani Vidakovic and Peter Mueller, Wavelets For Kids – A Tutorial Introduction, Duke University
[7] O. Farooq and S. Datta, A Novel Wavelet Based Pre Processing For Robust Features In ASR
[8] Giuliano Antoniol, Vincenzo Fabio Rollo, Gabriele Venturi, IEEE Transactions on Software Engineering, LPC & Cepstrum coefficients for Mining Time Variant Information from Software Repositories, University Of Sannio, Italy
[9] Michael Unser, Thierry Blu, IEEE Transactions on Signal Processing, Wavelet Theory Demystified, Vol. 51,No. 2,Feb'13
[10] C. Valens, IEEE, A Really Friendly Guide to Wavelets, Vol.86, No. 11, Nov 2012,
[11] James M. Lewis, C. S Burrus, Approximate CWT with An Application To Noise Reduction, Rice University, Houston
[12] Ted Painter, Andreas Spanias, IE EE, Perceptual Coding of Digital Audio, ASU
[13] D P. W. Ellis, PLP,RASTA, MFCC & inversion Matlab, 2005
[14] Ram Singh, Proceedings of the NCC, Spectral Subtraction Speech Enhancement with RASTA Filtering IIT-B 2012
[15] NitinSawhney, Situational Awareness from Environmental Sounds, SIG, MIT Media Lab, June 13, 2013