AN IMPROVED SPEAKER VERIFICATION SYSTEM FOR ROBOTIC APPLICATIONS

Abdul Rahim[1], Kommera Shravya[2], Konda Reddy Lakshmi Prasanna[2], Kuthala Kavya[2], Komuravelli Varsha[2]

[1,2]Department of Electronics and Communication Engineering

[1,2]Malla Reddy Engineering College for Women (A), Maisammaguda, Medchal, Telangana.

ABSTRACT

Text Dependent Human Voice Recognition (TDHVR) systems are used to verify the identity of individuals based on their speech signals. This abstract presents a TDHVR system that utilizes statistical computation, formant estimation, and wavelet energy analysis to achieve accurate verification. The system is evaluated using fifty preloaded voice signals from six individuals, and the proposed algorithm achieves an accuracy rate of approximately 90%, surpassing the performance of Linear Predictive Coding (LPC), which achieves only 66.66% accuracy.

Through extensive simulation tests conducted on various speech signals from different speakers, it is observed that the proposed algorithm significantly improves the accuracy of the TDHVR system compared to LPC. The integration of statistical computation, formant estimation, and wavelet energy analysis enhances the system's ability to accurately verify the identity of individuals based on their speech signals.

This work contributes to the field of voice recognition by presenting a novel approach that outperforms existing techniques, such as LPC. The achieved accuracy rate of approximately 90% demonstrates the effectiveness and potential of the proposed algorithm in practical applications requiring reliable identity verification through speech signals.

Keywords: Text Dependent Human Voice Recognition (TDHVR), identity verification, speech signal, statistical computation.

1. INTRODUCTION

Speech Production

Speech is the acoustic product of voluntary and well-controlled movement of a vocal mechanism of a human. During the generation of speech, air is inhaled into the human lungs by expanding the rib cage and drawing it in via the nasal cavity, velum and trachea. It is then expelled back into the air by contracting the rib cage and increasing the lung pressure. During the expulsion of air, the air travels from the lungs and passes through vocal cords which are the two symmetric pieces of ligaments and muscles located in the larynx on the trachea.

Speech is produced by the vibration of the vocal cords. Before the expulsion of air, the larynx is initially closed. When the pressure produced by the expelled air is sufficient, the vocal cords are pushed apart, allowing air to pass through. The vocal cords close upon the decrease in air flow. This relaxation cycle is repeated with generation frequencies in the range of 80Hz – 300Hz. The generation of this frequency depends on the speaker’s age, sex, stress and emotions. This succession of the glottis openings and closure generates quasi-periodic pulses of air after the vocal cords.

The speech signal is a time varying signal whose signal characteristics represent the different speech sounds produced. There are three ways of labelling events in speech. First is the silence state in which

no speech is produced. The second state is the unvoiced state in which the vocal cords are not vibrating, thus the output speech waveform is aperiodic and random in nature.

Figure below shows the schematic view of the human speech apparatus.

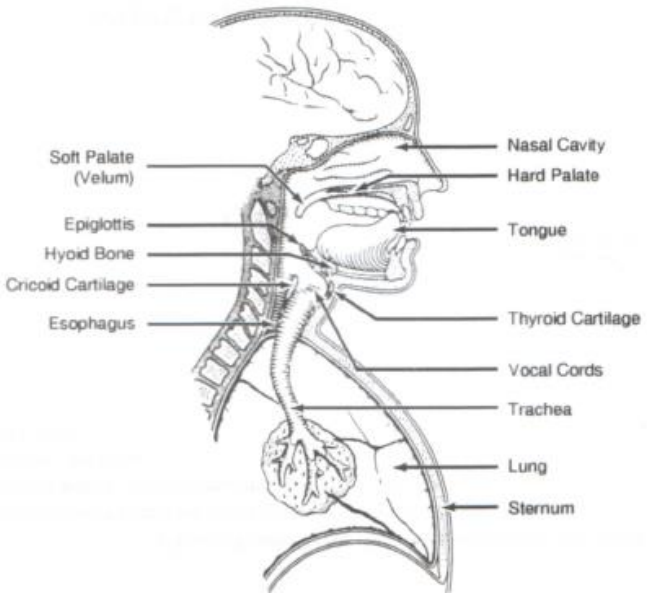


Fig. 1: Schematic view of human speech.

The last state is the voiced state in which the vocal cords vibrate periodically when air is expelled from the lungs. This results in the output speech being quasi-periodic. The type of sound produced depends on the shape of the vocal tract. The vocal tract starts from the opening of the vocal cords to the end of the lips. Its cross-sectional area depends on the position of the tongue, lips, jaw, and velum. Therefore, the tongue, lips, jaw, and velum play an important part in the production of speech.

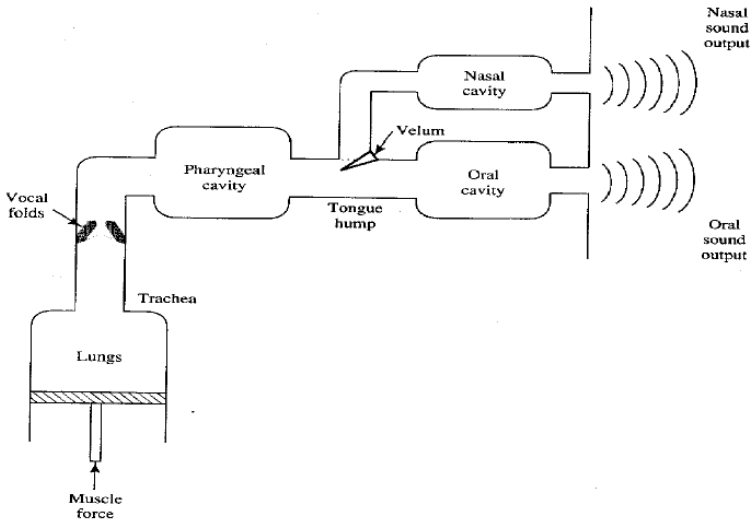


Fig. 2: Block Diagram (Engineering Model) of Human Speech Production System.

Factors associated with speech

Formants

It has been known from research that vocal tract and nasal tract are tubes with non-uniform cross-sectional area. As sound generated propagates through these the tubes, the frequency spectrum is shaped by the frequency selectivity of the tube. This effect is very similar to the resonance effects observed in organ pipes and wind instruments. In the context of speech production, the resonance frequencies of vocal tract are called formant frequencies or simply formants. In our engineered model the poles of the transfer function are called formants. Human Auditory system is much more sensitive to poles than zeros.

Phonemes

Phonemes can be defined as the "Symbols from which every sound can be classified or produced". Every Language has its particular phonemes which range from 30 – 50. English has 42 phonemes. For speech crude estimation of information rate considering physical limitations on articulatory motion is about 10 phonemes per second.

Types of Phonemes

Speech sounds can be classified in to 3 distinct classes according to the mode of excitation.

1. Plosive Sounds
2. Voiced Sounds
3. Unvoiced Sounds

Special Type of Voiced and Unvoiced Sounds

There are however some special types of voiced and unvoiced sounds which are briefly discussed here. The purpose of their discussion here is only to give the reader an idea about the further types of voiced and unvoiced speech.

Vowels

Vowels are produced by exciting a fixed vocal tract with quasi periodic pulses of air caused by vibration of the vocal cords. The way in which the cross-sectional area varies along the vocal tract determines the resonant frequencies of the tract (formants) and thus the sound that is produced. The dependence of cross-sectional area upon distance along the tract is called is called **area function** of the vocal tract. The area function of a particular vowel is determined primarily by the position of the tongue but the position of jaws and lips to a small extent also affect the resulting sound.

Examples a,e,i,o,u

Diphthongs

Although there is some ambiguity and disagreement as to what is and what is not a diphthongs, a reasonable definition is that a diphthongs is a gliding monosyllabic speech item that starts at or near the articulatory position for one vowel and moves to or toward the position for another. According to this definition, there are 6 diphthongs in American English.

Diphthongs are produced by varying the vocal tract smoothly between vowel configurations appropriate to the diphthong. Based on these data, the diphthongs can be characterized by a time varying vocal tract area function which varies between two vowel configurations.

Examples: Ei/ (as in b<u>ay</u>) , oU/ (as in b<u>oa</u>t) , aI/ (as in b<u>uy</u>) , aU/ (as in h<u>ow</u>)

Semivowels

The group of sound consisting of /w/, /l/ , /r/ ,/y/ is quite difficult to characterize. These sounds are called semivowels because of their vowel-like nature. They are generally characterized by a gliding transition in the vocal tract area function between adjacent phonemes. Thus the acoustic characteristics of these sounds are strongly influenced by the context in which they occur. For our purpose they just considered transitional vowel-like sounds and hence are similar in nature to vowels and diphthongs.

Voiced Fricatives

The voiced fricatives are /v/ , /th/ , /z/ and /zh/are the counterpart of the unvoiced fricatives /f/ ,/θ/ ,/s/ and /sh/ respectively, in that the place of constriction for each of the corresponding phonemes is essentially identical.

However, the voiced fricatives differ from their unvoiced counterparts in the manner that two excitation sources are involved in their production. The spectra of voiced fricatives can be expected to display two distinct components.

Voiced Stops

The voiced stops /b/, /d/, and/g/ are transient non-continuant sounds which are produced by building up pressure behind a total constriction somewhere in the oral tract, and suddenly releasing the pressure. For /b/ the constriction is at the lips; for /d/ the constriction is at the back of the teeth; and for /g/ it is near the velum. During the period there is a total constriction in the tract there is no sound radiated from the lips. Since the stop sounds are dynamical in nature, their properties are highly influenced by the vowel which follows the stop consonant.

Unvoiced Stops

The unvoiced stop consonants are /p/,/t/,and/k/ are similar to their voiced counterparts /b/ , /d/ and /g / with one major exception. During the period of the total closure of the tract, as the pressure builds up, the vocal cords do not vibrate. Thus, following the period of closure as the air pressure is released, there is a brief interval for friction (due to the sudden turbulence of the escaping air) followed by a period of aspiration (steady flow of air from glottis exciting the resonances of the vocal tract) before voiced excitation begins.

Hearing and Perception

Audible sounds are transmitted to the human ears through the vibration of the particles in the air. Human ears consist of three parts, the outer ear, the middle ear, and the inner ear. The function of the outer ear is to direct speech pressure variations toward the eardrum where the middle ear converts the pressure variations into mechanical motion. Mechanical motion is then transmitted to the inner ear, which transforms this motion into electrical potentials that passes through the auditory nerve, cortex and then to the brain. Figure below shows the schematic diagram of the human ear.

The Engineered Model

The speech mechanism can be modeled as a time varying filter (the vocal tract) excited by an oscillator (the vocal folds), with different outputs. When voiced sound is produced, the filter is excited by an impulse chain, in a range of frequencies (60-400 Hz). When unvoiced sound is produced, the filter is excited by random white noise, without any observed periodicity. These attributes can be observed when the speech signal is examined in the time domain.

2. Proposed Implementation

2.1 RASTA (Relative Spectral Algorithm)

RASTA or Relative Spectral Algorithm as it is known is a technique that is developed as the initial stage for voice recognition. This method works by applying a band-pass filter to the energy in each frequency sub-band 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. 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. The stationary noises that are identified are noises in the frequency range of 1Hz - 100Hz.

2.2 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. The formant of the signal can be obtained by analyzing the vocal tract frequency response. Figure 5.1 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 AP (Linear predictive coding) method is used. This method is derived from the word additive prediction; 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 additive function of previous samples.

2.3 Proposed TDHVR Implementation

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 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.

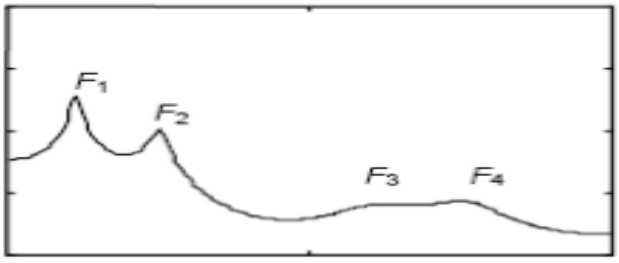


Fig. 3: Formant estimation.

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:

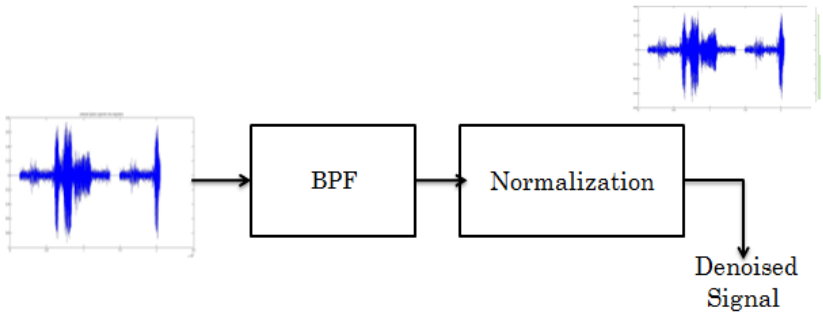


Fig. 4: Block diagram of RASTA process.

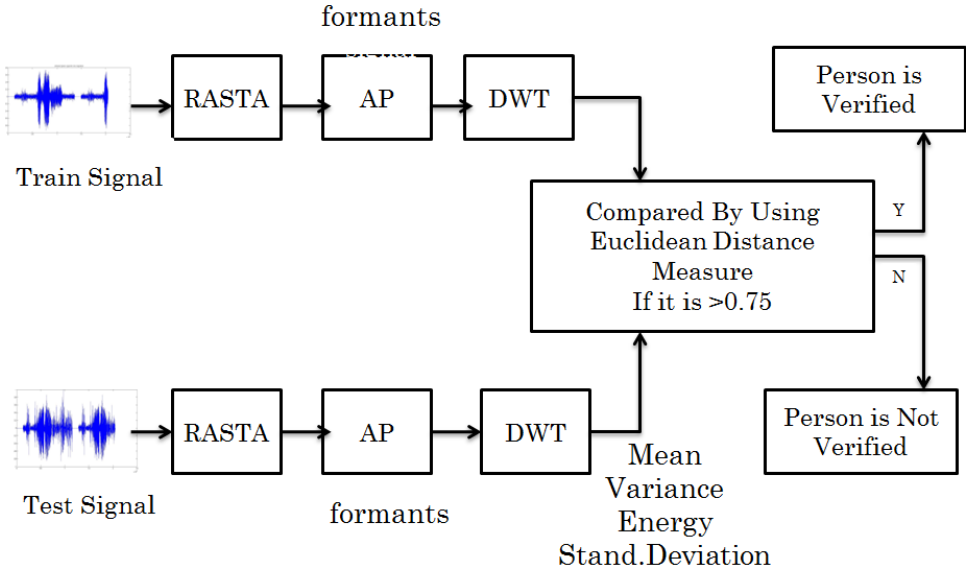


Fig. 5: Block diagram of proposed TDHVR system.

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.

3. SIMUALTION 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 = $\frac{1}{n}\sum_{i=1}^{n} x_i$

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

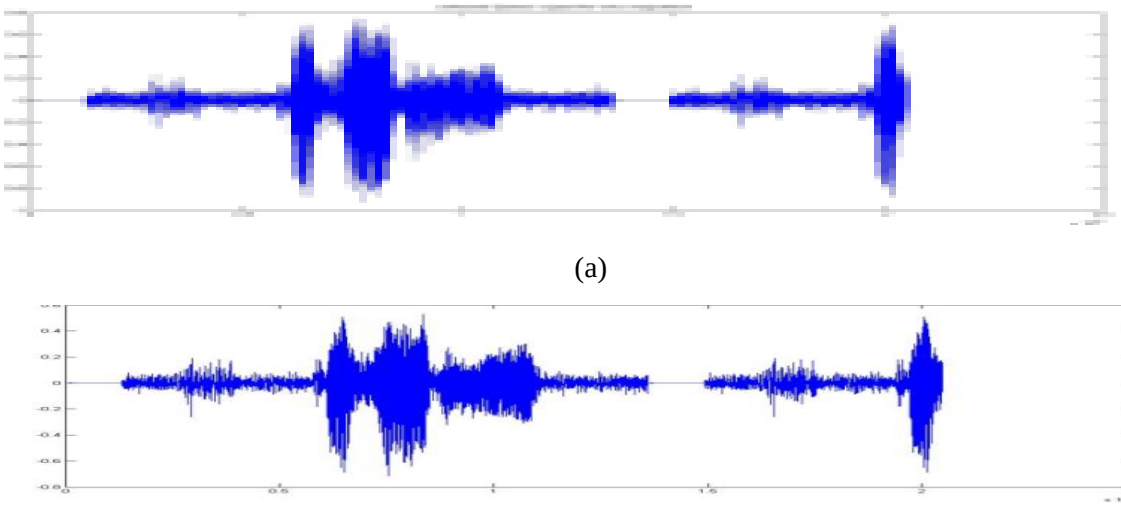
Std. deviation =

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 tabel2 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.

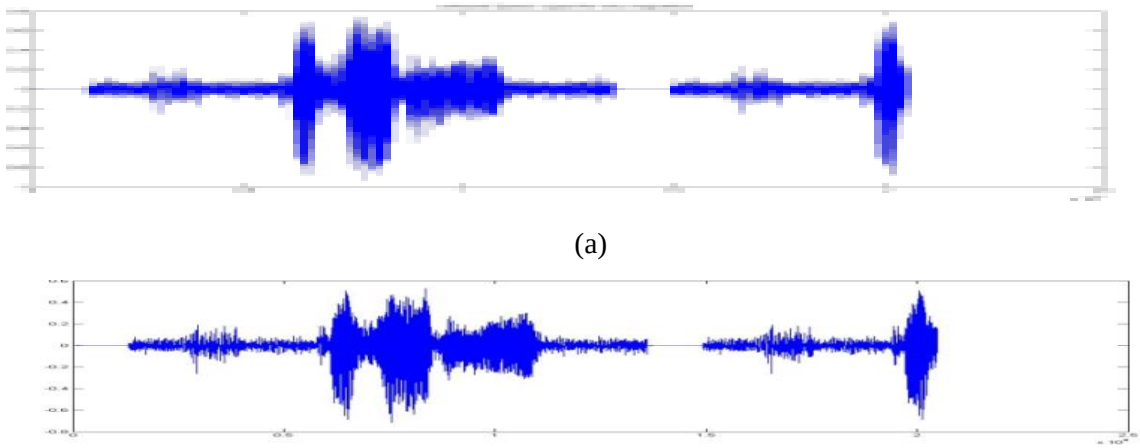
SAMPLE1



(a)

(b)

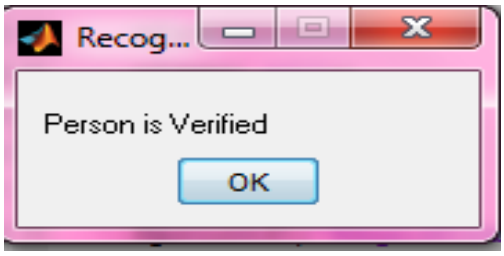
Fig. 6: (a) Original voice signal (b) De-noised signal for training.



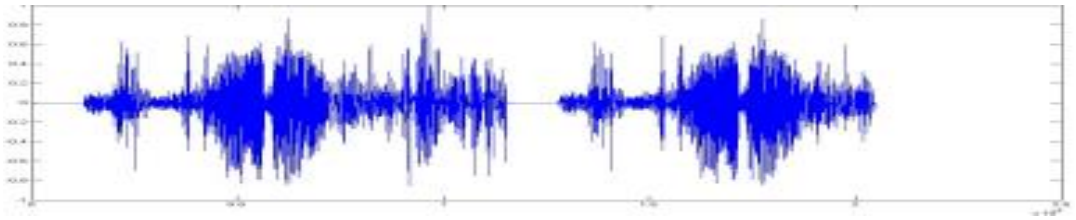
(a)

(b)

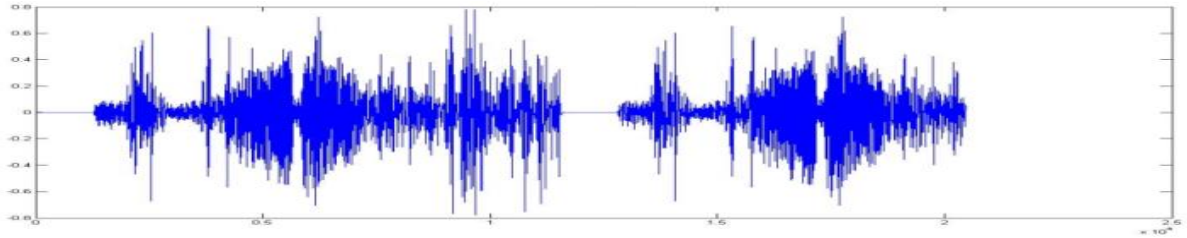
Fig. 7: (a) Original voice signal (b) De-noised signal for testing.



SAMPLE2

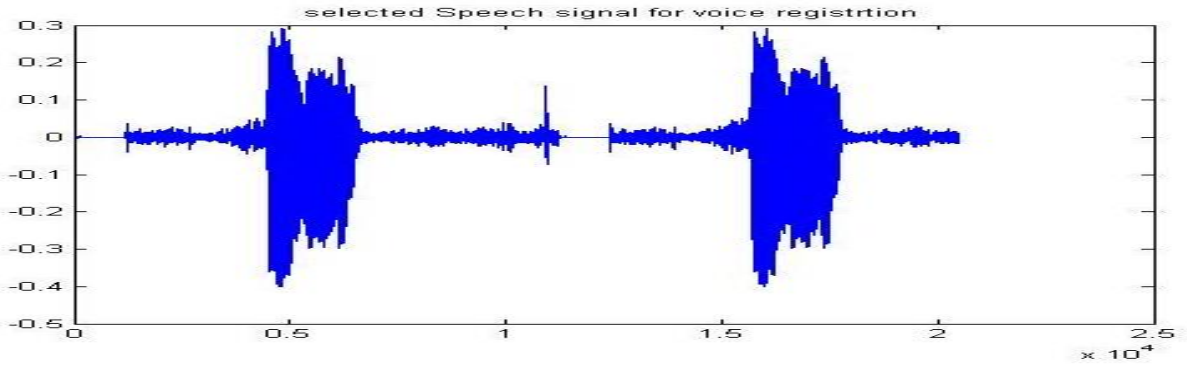


(a)

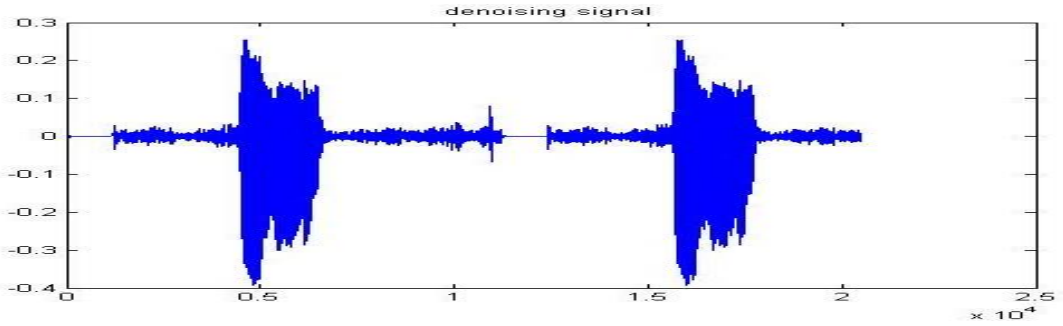


(b)

Fig. 8: (a) Original voice signal (b) De-noised signal for training.



(a)



(b)

Fig. 9: (a) Original voice signal (b) De-noised signal for testing.

3. CONLUSIONS

Text Dependent Human Voice Recognition (TDHVR) system used to verify the identity of an individual based on their own speech signal using statistical computation, formant estimation and wavelet energy. By using the fifty preloaded voice signals from six individuals, the verification tests have been carried out 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 speakers 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.

[16] Rami Al-Hmouz, Khaled and Ali, "Multimodal Biometrics Using Multiple Feature Representations to Speaker Identification System", *International Conference on Information and Communication Technology Research (ICICTR),* 2015