International Journal of Engineering & Technology, 7 (2.33) (2018) 145-150

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Prediction based person recognition using face and speech (multi modal) for improved performance

Dinesh Kumar. D. S [1] *, Dr. P. V. Rao [2]

[1] Ph.D. Research Scholor, VTU BelgaumAssociate Professor, KSSEM, Bangalore
[2] Prof., VBIT, Hyderabad & Research Supervisor, TJIT, Bengaluru-560083
**Corresponding author E-mail: pachararao@rediffmail.com*

Abstract

In recent World Technological Applications Person Recognition plays a major role in biometric security applications and it is a process of authenticating true identity of a speaker using speech or face image. This automatically recognizes the person speaking based on the speech information which includes the individual speech signals. It's one of applications is Bio-metric applications and in order to verify each person identity, the speaker's voice or face images have been used in the database. The importance of it becoming more popular now-a-days for security purpose and identification. In the existing work using visualization eye scan, impressions, expression scan, finger print, speech print, script for individuals of identifying the chances of theft and fraud are increasing. To address these issues, biometric voice recognition and real time face recognition system are proposed, the exclusive speaker physical appearance of a distinct can be identified. In general the individuals have different speed of speaking; therefore the sound should be adjusted, in order to match with the speed of the stored sounds templates in the memory of the proposed system. The proposed work is implemented and simulated using on Mat lab 2014A. The parallel hardware structure of the proposed work significantly reduces the time-consumption. The proposed research work provides maximum False Acceptance Rate (FAR) of 1.1765%, False Rejection Rate of 10% with an accuracy of 98.89%.

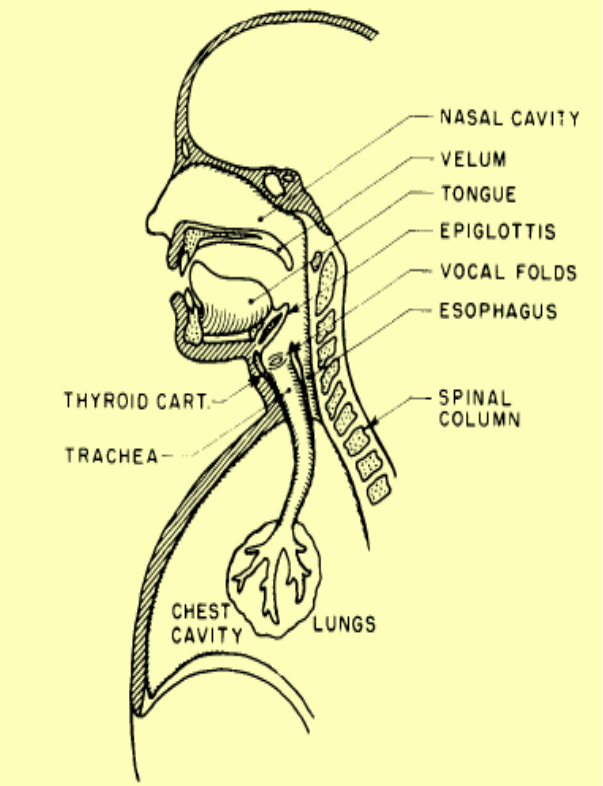
Keywords: *Speech; Human Communication; Feature Extraction; Front Ends; FRR; FAR*

1. Introduction

In current technologies, Facial expressions depend on the parametric such as kind of background noises and facial expressions. Principal Component Analysis (PCA) and Hidden Markov Model (HMM) are used for face recognition. Artificial Neural Networks (ANN) is used for classification. This method provides a recognition rate of 96% reduced FAR and FRR [1] Speech identification process recognizes the person identity by keeping individual's information from speech waves [2]. Speech recognition and face recognition forms a multimodal biometric system for better recognition rate and reduced FAR and FRR .In this histogram based method is used for Face feature extraction and Least Mean Square (LMS) algorithm is used for speech feature extraction. Adaptive filter is employed for removing noise in the speech signal. This method provides a recognition rate of 95% for 20 databases of both face and speech [2] [3].

Speech signals are mainly used for anthropological communication [22] [25].Human Speech represents the information with different levels of knowledge based information of sources like semantics, phonetic, acoustic and articulators. These parameters help in analysing speech processing with various fields [13].

Voice information is produced by combining the air with the vocal cords from the human body [13] [20]. The speech information produced by human is formed by vocal cavity. The speech source is one of the source within the speech system. In this processing, the features are extracted from raw input data which helps interpreting the speech signal [14] [15] [22] [27].



Copyright © 2018 Dinesh Kumar. D. S 1, Dr. P. V. Rao. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

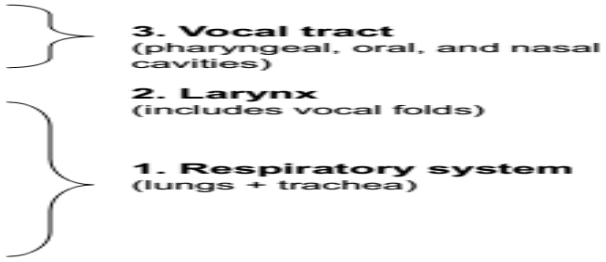


Fig. 1: Human Speech Production System.

Speech features are represented as prosodic features and acoustic features and time convolved signal [15] [20] [22]. Fig.1 represents the human speech production system. Front end means adaptation of speech wave form into parametric form. The speech processing are applied in most of the applications like; analysis, recognition and synthesis of speech to extract features as an important stage in speech processing [13] [22]. Speech analysis is the major step in processing the speech signal. There are several analysis methods available for speech. Cepstral analysis, MFCC, PLP, LPC, STFT and Wavelet transforms are used to identify the speech information.

Time domain analysis: In general human speech signals are non-stationary with short varying time amplitude. The speech is generally not smooth due to sudden transitions and is having irregular shapes at different instant. Analysing these irregular shapes of speech signals and obtaining desired signal transitions using ordinary transformations such as Discrete Cosine Transformation (DCT), Fast Fourier Transform (FFT) and conventional Discrete Wavelet Transform (DWT) is difficult. The DCT and FFT have more disadvantages and hence not suitable for high speed and large memory. The Short Time Fourier transform (STFT) is to be one of the suitable transform to analyze and for the represent the speech [20] [26].

2. Problem statement

For real time implementation of speech The Mel Frequency Cepstral Coefficient (MFCC) is not enough for synthesis for features extraction. MFCC is problematic due to its hard invertibility of the transform and MFCC are a destructive of speech signal particularly for fundamental frequency information and instantaneous phase. The person recognition using face image and audio signal of a particular person is main challenging task .hence prediction based coefficients are used to address the issues in real time. In the proposed work both real time captured face image and real time recorded speech signal are used as database for processing.

3. Methodology

Speaker or face recognition is broadly categorized into two stages; identification or verification and authentication [2]. Speaker identification is the process of linking unknown person identity to an individual with known identity. Authentication is the process of proving secure data against false identity. Fig. 2 represents the basic building blocks of speaker identification. In this proposed work both face and speech are combined to identify the person. Features for face and speech are obtained by using prediction based coefficients to find the recognition rate.

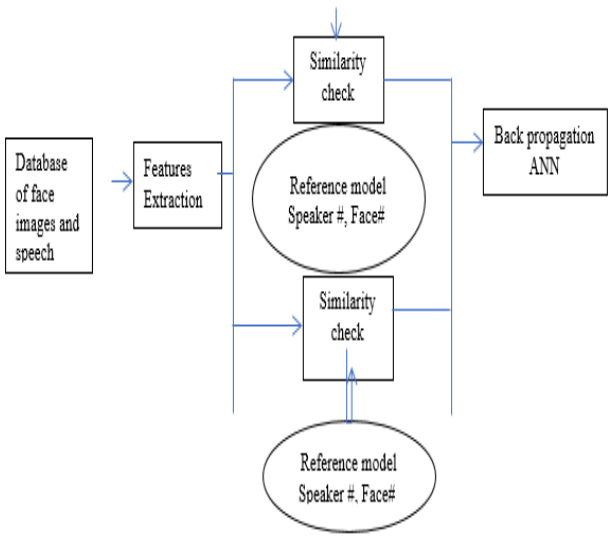


Fig. 2: Speaker Identification System.

The speaker validation is one process for person recognition and process of accommodating or rejection of the distinctiveness treat of a speaker. Fig. 3 represents the basic steps involved in feature extraction and matching is represented by speaker verification system. In this process, it extracts the required portion of the speech information with characteristics from the given voice database.

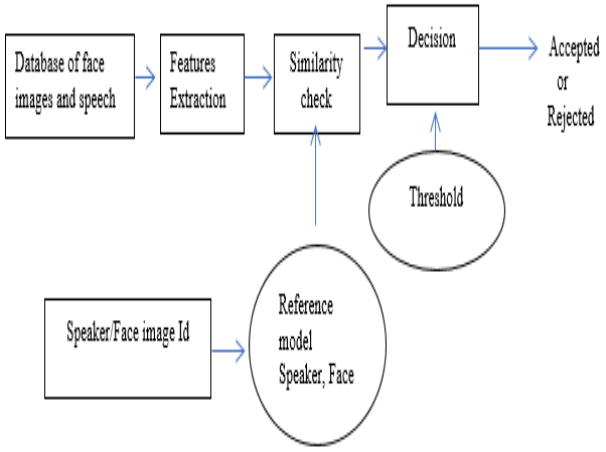


Fig. 3: Speaker Verification System.

STFT is mainly used as a tool for audio signal processing with more accuracy. Fundamental Frequency Information (FFI) and Instantaneous Phase (IP) features are extracted without any losses using STFT. In the proposed work, both STFT and Advanced Wavelet Transforms like; Symlet and Bi orthogonal sub band filters are used for face and speech feature extractions.

STFT uses small window of fixed length. The Fourier transform of the signal is a 2D signal represented using equation 1

$$S(\omega,\tau)=\int f(t)*g(t-\tau)\exp(-j\omega t)dt \quad (1)$$

Where, g (t) is the fixed width integral square short time window and is shifted along the time axis by a factor τ. f (t) is the Fourier transform of resultant signal.

For a discrete real time signal X (n), the energy E (n) is given by equation 2

$$E(n) = \sum_{n=-\infty}^{\infty} x^2(n) \quad (2)$$

The speech signals are non-stationary, whose energies are calculated by equation 3

$$E(n) = \sum_{m=0}^{N-1} w(m)x(n-m)^2 \quad (3)$$

Where, N= number of samples of x (m) selected through a weighing window w (m).
Normally unvoiced sounds represent lower energy level than voiced sounds. Hence, auto correlation function is used to estimate the speech.
The auto correlation function $\emptyset\ (m)$of a signal x (n) can be calculated using equation 4

$$\emptyset\ (m)\ = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=N}^{N} x\ (n).\ x(n+m) \tag{4}$$

Frequency Domain Analysis:
In quantum, physics to know arbitrary precision the position and momentum properties of Heisenberg uncertainty is applied. These properties of position and momentum can be applied for image processing to analyse precisely both time and frequency.The product of $\sigma_t \sigma_w$ must be greater than $\frac{1}{2}$, where σ_t is the deviations in time domain and σ_w is the deviation in frequency domain. The wavelet transform function is $\frac{1}{\sqrt{s}}\varphi(\frac{t-u}{s})$, where the parameter s is always inversely proportional to the frequency. Based on the value of s, low and high frequency signals can be extracted. The variation of s value represents increase in flexibility with respect to time-frequency analysis.
For 2D, the signal information is modified and represented by $e^{(j(W_1 t_1 + W_2 t_2)}$ instead of e^{iWt}. Now scaling and wavelet functions are represented as $\emptyset(x,y)\ and\ \psi(x,y)$. After applying separable theorem to scaling and wavelet functions, the new window functions are represented in equations 5 and

$$W_{\emptyset}(j,m,n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \emptyset_{j,m,n}\ (x,y) \tag{5}$$

$$W_{\psi}(j,m,n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \psi_{j,m,n}\ (x,y) \tag{6}$$

Where

$$\mathrm{f(x,y)} = \frac{1}{\sqrt{MN}} \sum_m \sum_n W_{\emptyset}(j,m,n)\ \emptyset_{j,m,n}(x,y)$$

$$\sum_{i=H,V,D} \sum_{j=0}^{\infty} \sum_m \sum_n W_{\emptyset}(j,m,n)\ \emptyset_{j,m,n}(x,y)_{j,m,n}(x,y)$$

Here wavelet transformation is used to detect low frequency components from wavelet function equation 7 is applied.

$$W_{f(s,u)} = \int_{-\infty}^{\infty} \frac{1}{\sqrt{s}}\ \psi\left(\frac{t-u}{s}\right) \mathrm{DT} \tag{7}$$

From equation 7, the low frequency signal can be determined at large of s while high frequency components can be located at small s and s inversely proportional to the frequency . Every non border pixel usually has eight adjacent neighbouring pixels. These eight neighbours are used to traverse the matrix. When the eight neighbors are moved bidirectional i.e. right-to-left and vice versa then the 2-D DWT coefficients remains constant. Hence four decomposition directions such as 0◦, 45◦, 90◦ and 135◦ are sufficient. The results obtained from two levels DWT are found to yield significant features. The output feature vectors are represented by equation 8.

$$\mathrm{Y\ (z)} = \frac{1}{2}\left\{(1+Z^{-1})(\frac{1-Z^{-1}}{2}) \pm (1-Z^{-1})(\frac{1+Z^{-1}}{2})\right\} \tag{8}$$

In order to reduce the $(1-Z^{-1})$ polynomial terms the Symlets and Biorthogonal wavelet sub-bands filters have been included in the proposed work.
DWT produces a spectrum for a speech signal [8]. The discrete short time spectrums of a given speech signal x (n) is represented by equation 9.

$$x_l\ (w) = \sum_{n=-}^{l} x(n).h(l-n).e^{-jwn} \tag{9}$$

Where, h (l-n) represents windowing function of filter represented by equation 10.

$$|X_l(w)|e^{j\theta_l(w)} = a_l(w) - jb_l(w) \tag{10}$$

The short term spectrum can be computed by the equation 11

$$X_l(w) = \sum_{n=0}^{N-1} x_l(n).w(n).e^{-jwn} \tag{11}$$

Where

$$x_l(n) = x(n+L), \quad n = 0 \text{ to } N-1\ \&\ L = 0, L, 2L.$$

The window length N specifies the resolution. The large window is selected for better estimation of acoustic pitch and vocal tract function [8] [13] [25].

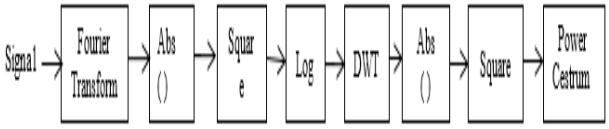


Fig. 4: Block Diagram for Power Cepstrum of Input Speech Signal.

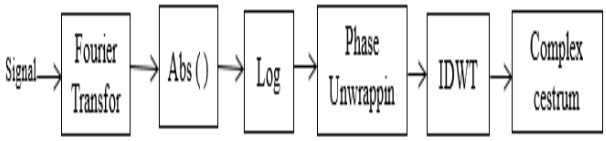


Fig. 5: Block Diagram for Complex Cepstrum of Input Speech Signal

Fig. 4 & Fig. 5 represent the block diagrams of input signal power and complex Cepstrum for a given speech signal. Fig. 6 represents the basic building blocks while computing for the Cepstral coefficients.

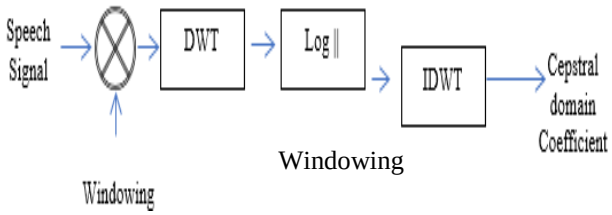


Fig. 6: Block Diagram for Computing the Cepstral Coefficients.

4. Cepstral analysis of speech:

The coefficients of Cepstral are obtained from the analysis and are used in speech recognition, speaker verification, etc. These coefficients are obtained represents Fourier Transform of the log spectrum signal. Cepstrum is categorized as Power and Complex Cestrum [10] [8] [20].

5. Mel frequency cepstral coefficients (MFCC)

The MFCC coefficients are obtained from speech signal spectrum [11] [12]. This work splits into two parts: i) design of band pass filters ii) measurement of Mel scale and bark scale .The Mel frequency and linear frequency scale is represented using equations 12 and 13.

Mel frequency=2595*log (1+linear frequency/700) (12)

=1127 Ln (linear frequency/700 +1) (13)

The conversion frequency (f) for a bark scale can be obtained by using equation 14

$$\Omega(w) = 6\ln\left[\frac{w}{120\pi} + \left[\left(\frac{w}{1200\pi}\right)^2 + 1\right]^{0.5}\right] \quad (14)$$

$\Omega \rightarrow$ Angular frequency in bark scale & $w \rightarrow$Angular frequency = $2\pi f$

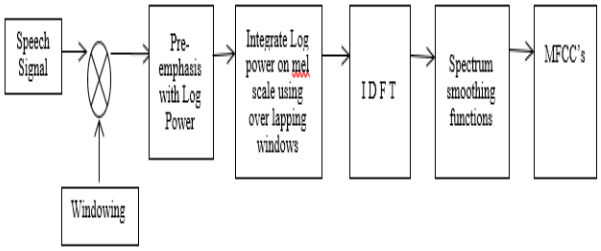


Fig.7: Block Diagram for Computation of MFCC.

6. Linear Prediction Coefficients (LPC)

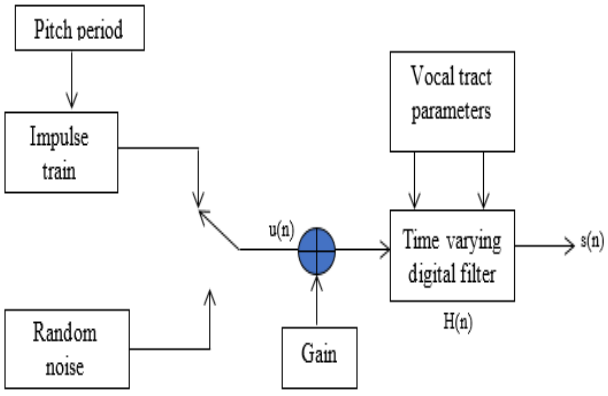


Fig. 8: Block Diagram Showing LPC Model.

Using IDFT the Cepstral coefficients are obtained and finally perform the spectral smoothing is done to obtain MFCC coefficients. Fig 7 represents the procedure for computation of MFCC.
The Cepstral coefficients can be determined by performing the DCT of Mel-bin log energies. The Cepstral coefficients are represented in equation 15.

$$C_i(t) = \sum_{b=1}^{B} \log m_b(t).\cos\left(\frac{i\,(b-0.5)\pi}{B}\right) \quad (15)$$

Where, B = number of triangular shaped filter bank functions.
S (n) H (n) Fig. 8 represents the block diagram of LPC model. The speech signal can be estimated as a linear combination of previous speaker sample signal is fundamental idea of LPC. The difference between speaker voice and their linearly predicted samples is the sum of the squared and these signal must be minimized to get obtain the coefficients of predictions [17] [12]. In speech signals the acoustic parameters represents intensity and pitch. The vocal tract system forms a structure characterized by formants.

7. Perceptual linear prediction (PLP) coefficients

PLP is a combined DFT and LP technique and adopts the perceptual property of the human ear. By using AR model, speech auditory spectrum is estimated [five] [24] [28].
Trapezoidal shaped filters are used in PLP technique in order to estimate the bark frequency on bark scale. The equal loudness pre emphasis is carried out using equation 15.

$$E(w) = \frac{(w^2+56.8\times10^6)w^4}{(w^2+6.3\times10^6)(w^2+0.38\times10^9)(w^6+9.58\times10^{26})} \quad (15)$$

Where w = angular linear frequency, E (w) = energy at a given frequency

8. Artificial neural network (ANN) for classification

The main parts of the neural network are synaptic, processing cell called perceptron and weights to process the input signal and decision of the output. Between neuron to neuron, the strength can increase or decrease so that subsequence neurons can excite from one neuron to another neuron. This process is used for the storage of information. The perceptron accepts the inputs in terms real value and then it calculates the combination inputs and produces the outputs based on the following conditions

$$output = \begin{cases} -1, & the\ result\ < threshold \\ 1\ , & the\ result > threshold \end{cases}$$

Let the real input value range is from x_i to x_n, the outputs are x_1, x2, xn. The inputs and outputs are computed by using processing cell when $w_0 + w_1x_1 + \ldots + w_nx_n > 0$. The number of inputs, hidden layer, learning rate, momentum parameters and output layer are decided by trial and error methods. To fix one parameter, remaining parameter must be fixed for some constant and vary the one parameter until to get expected output and similarly remaining parameters can be fixed. The main contributions of ANN architecture are data collection, training and testing the data separation, initialization of weight, data transformation and finally production of the output.

9. Results and discussion

A speech signal can be characterized as supernatural vectors. Speech spectrogram is a visual representation of log magnitude amplitude (dB) versus time and frequency as shown in Fig. 10(a b). From the spectrogram, the voiced and unvoiced regions of speech is distinguished which helps in detecting the acoustic parameters like formants of speech signal. The speech sounds can be detected by the formants and their transitions as shown in Fig.11 (a-b).
The spectrograms are calculated from the time signal using the STFT. Spectrogram represents spectra computed by the Discrete Wavelet Transform (DWT) displayed parallel to the Y-axis.

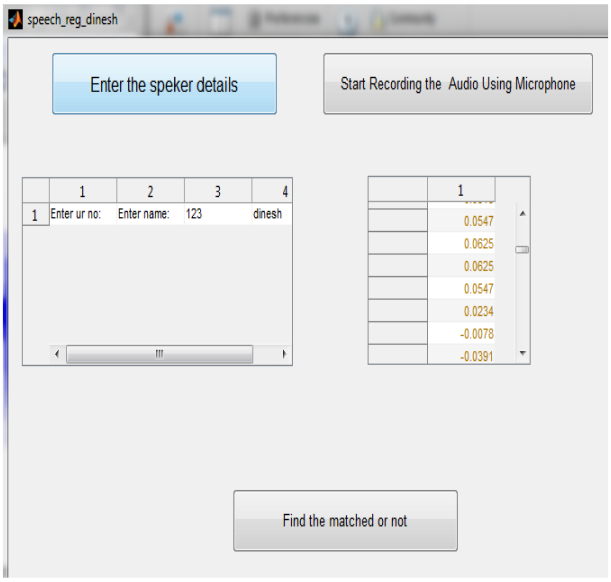


Fig. 10: (A) Enrolment of Speaker and Record the Voice Signal.

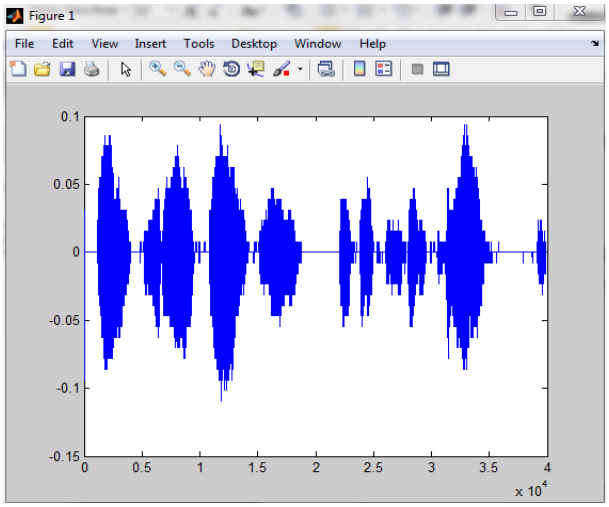


Fig.10: (B) Graphical Representation of Voice Signal.

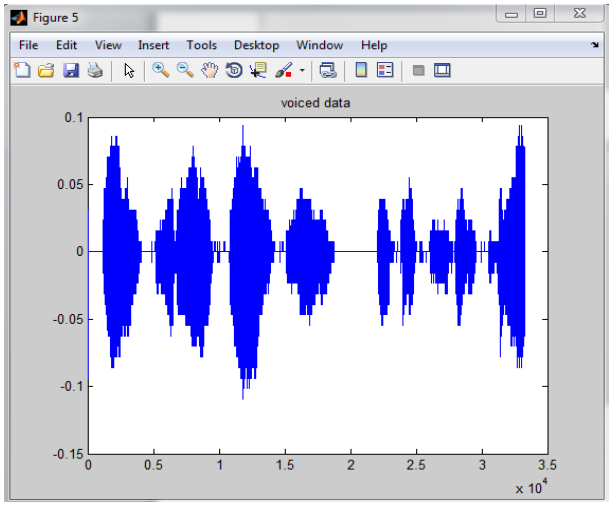


Fig. 11: (A) Reference Audio Signal Stored in the Database.

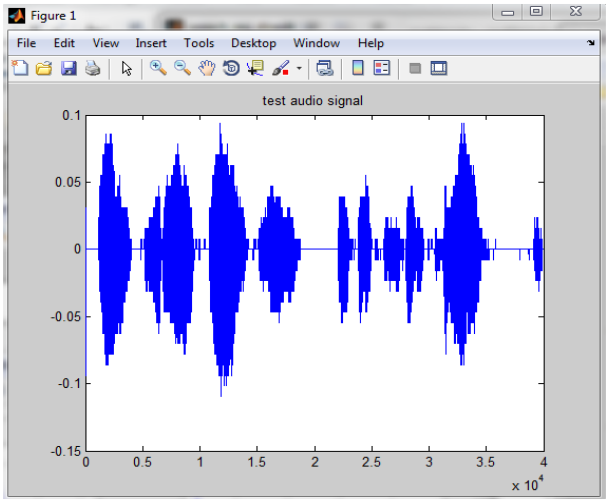
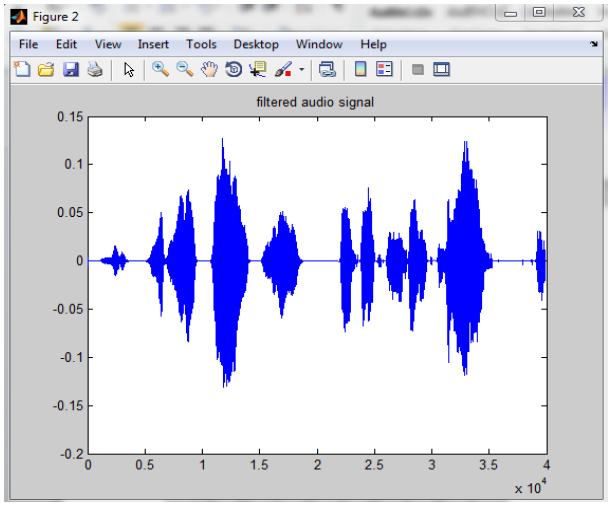


Fig. 11: (B) Output of Filtered Voice Signal.



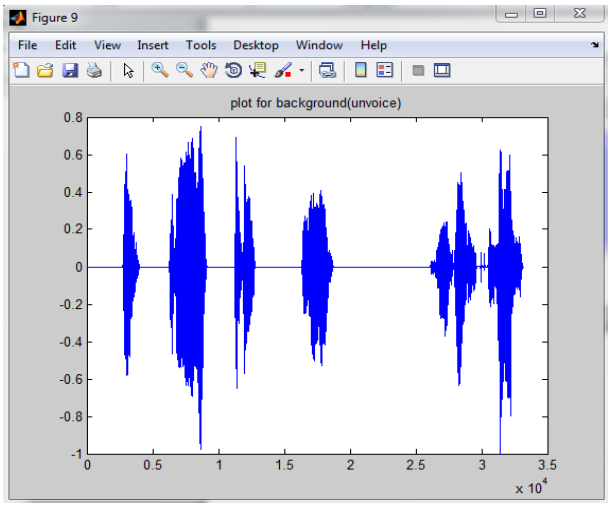


Fig.12: (A) Mel Frequency Voiced Data.Fig.12 (B): Background Unvoiced Data.

For analysis of speech, narrow band spectrogram is used which provides low frequency resolution as shown in Fig.12 (a-b). A narrow band spectrogram displays individual harmonics for analysis of pitch and vocal tract excitation. The filtered and its resonance frequency curves are shown in Fig.13 (a-b).finally training the data set with ANN architecture for both speech signal and database of face images produces the result as whether the person is identified or not as shown in Fig.14(a-b). Table.1 represents the comparison of FAR and FRR values with different databases.

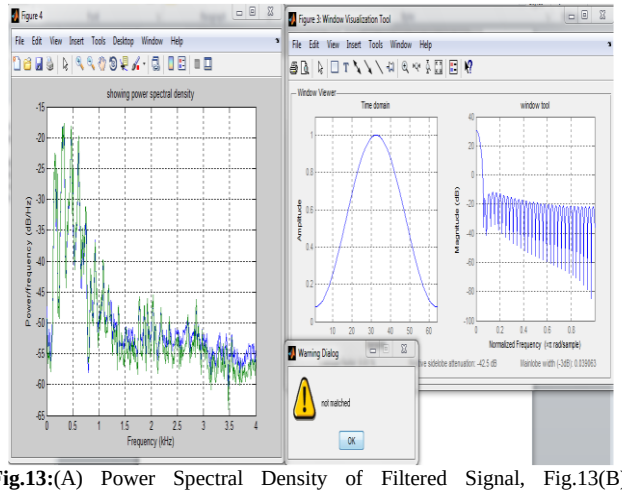


Fig.13:(A) Power Spectral Density of Filtered Signal, Fig.13(B): Frequency Signal in Window and Final Output.

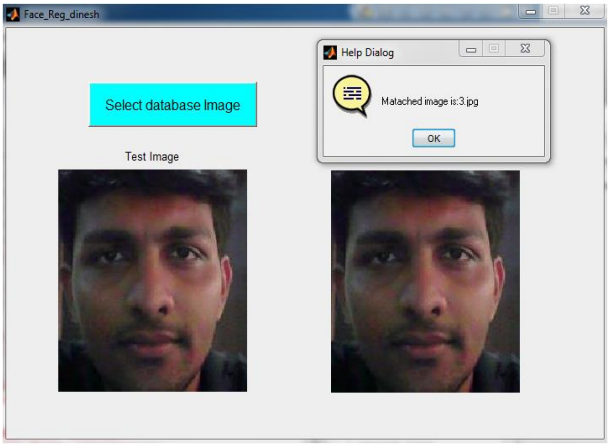


Fig. 14:(A) Load the Database Face Image And Finalized Decision of Matched Image of 3rd.

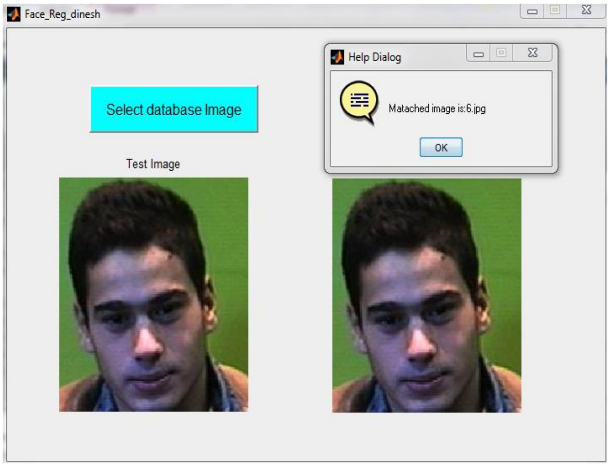


Fig. 14: (B) Load the Database Face Image and Finalized Decision of Matched Image of 6th

Table 1: Comparison of FAR and FRR Values with Different Databases

Number of Databases	FAR (%)	FRR (%)	Accuracy (%)
10	0.8772	5.2632	98.8920
20	0.7772	4.1632	97.120
30	0.399	9.4	98.4444
40	0.2299	10	96.4444
50	1.1765	10	95.5714

10. Conclusion

Multimodal biometrics is widely used to overcome the demerits of single mode trait. In the proposed method, both face and speech signals are used for person recognition. MFCC is not enough for feature extraction, prediction based analysis methods like; PLP, LPC are used along with MFCC for better recognition. Both face and speech features are extracted using MFCC, PLP and LPC methods and finally classification is done using artificial neural networks. Real time data bases of face and speech are used for verification. The proposed work is implemented and simulated using Mat lab 2014A tool. The parallel hardware structure of the proposed work can significantly reduce the time-consumption. The algorithm is tested for different data bases. The proposed method provides maximum False Acceptance Rate (FAR) of 1.1765%, False Rejection Rate (FRR) of 10% with an accuracy of 98.89%.

References

[1] Dinesh kumar D.S and Dr P V Rao, "Analysis and Design of Principal Component Analysis and Hidden Markov Model for face recognition" Procedia Materials Science 10 (2015) 616 – 625.

[2] Dinesh kumar D.S and Dr P V Rao, "Person identification using combined Face and Speech for the reduction of FAR and FRR" I J C T A, 10(9), 2017, pp. 891-898.

[3] Neethu Santhosh, Dominic Mathew, Abraham Thomas, "Person Verification Using Multimodal Biometric", International Conference on Computer Communication and

[4] Informatics (*ICCCI* -2017), Coimbatore, INDIA, 978-1-4673-8855-9/17/$31.00 ©2017 IEEE.

[5] Gilbert Strang and Truong Nguen, "Wavelets and Filter Banks", Wellesley-Cambridge Press, MA, 1997, pp. 174-220,365-382. [5] H. Hermansky, "Perceputal Linear Predictive (PLP) analysis of speech", Journal of Acoustics. Soc. Am., 87 (4): 1738-1752, 1990.

[6] H. Hermansky, N. Morgan, "Rasta Processing of Speech", IEEE Trans. on Speech and Audio Proc., Vol.2, No.4, 1994.

[7] M. A. Anusuya and S. K. Katti, "Mel-frequency discrete wavelet coefficients for Kannada Speech recognition using PCA", International conference on Advances in computer Science, Dec.21-22, 2010, Trivandrum, Kerala, India.