Speech Identification using GFCC, Additive White Gaussian Noise (AWGN) and Wavelet Filter

Sahil Arora
Student, M.Tech (CE) Department
UCOE,Punjabi University
Patiala, Punjab, India

Nirvair Neeru
Assistant Professor, M.Tech (CE) Department
UCOE,Punjabi University
Patiala, Punjab, India

ABSTRACT
This paper deals with the identification of speakers identity from the given set of values of speech from the database. The major problem during the identification of speech is noisy environment which degrades the system performance during its mismatch. So one can say identification using speaker recognition is the vital issue in research. This paper tells about the various used techniques like GFCC i.e. Gamma tone Frequency Cepstral Coefficients as its speech detection algorithm and Gaussian Mixture Model (GMM) to estimate the Gaussian model parameters. This paper basically focuses on improvement of speech identification in noisy environment using Wavelet filter which are added to de-noise the speech signals. These techniques are applied on store value of databases in Attendance system application. and the features of the speech are then matched from the database. Experiment are done 15 speech values saying phrases 'Present Mam','Present sir','Yes mam','Yes sir'with 4 types of utterance for each phase. This Experiment shows better results for stored database oriented system and gives 85% of the correct recognition rate i.e. CORR and 73% results are given when wavelet filter are not used .

Keywords
Gammatone Frequency Cepstral Coefficients (GFCC), Gaussian Mixture Model (GMM), Cepstral mean normalization (CMN), Robust Speaker Identification, Additive White Gaussian Noise (AWGN), Wavelet Filter, End detection of input signal.

1. INTRODUCTION
The word biometric comes from the word 'bio' which means biological and 'metric' means measurement. So one can say biometric means person's biological traits which can either be physiological or the behavioral one. Using either of these biometric trait it becomes very easy to identify the person's identity and the process of memorizing the password or record keeping also gets discarded. The Physiological characteristics are based on the shape of the body like fingerprints, iris recognition, DNA, Face recognition etc. And the Behavioral characteristics related with persons individuals behavior like his or her speech, word utterance, typing speed etc. Thus one can say the selection of trait depends upon user requirement and the characteristics of that particular trait.
Nowadays there are large number of biometric traits present but this paper discuss about the physiological trait i.e. speech as biometric attendance system
The Best way of communication among homo sapiens is speech. This biometric tool is most efficient way to identify the person's identity. The speech recognition depends on various features like accent, word utterance, spoken words etc. The Speech Identification (SI) is basically the authentication of person speech as his or her biometric. The SI system is categorized into two ways: Text dependent in which user knows what to speak whereas text independent in which user can freely speak anything without any given prescribed format.
The best thing to detect speech as biometric trait due to its low equipment cost, less time consuming as moreover there is no physical contact of user with the system. The main theme of this paper is to apply various techniques and to provide efficient match of speech from database even in the noisy environment.

1.1 Background study
In this modern scenario, attendance system is one of the key aspects in every field it can either be any government, private institutions. But even in this 21st century attendance monitoring in school and colleges are far behind. Every student individually marks his or her attendance manually on answer sheet which is quite time consuming. In addition to this the whole record of the sheets have to be built manually which is quite difficult and lengthy time consuming task and results into more paper wastage as well as lesser accuracy. To overcome this problem one has to ponder an automation system which can increase the efficiency and reliability of system.

Generally, this is the following parameters like Fingerprint [1], Thumb impression [2], Hand geometry [3], Iris Recognition [4], Facial Recognition [5] and Voice [6] on which the monitoring of attendance system depends. Along with these systems there are many others system also present which are based on the web like GSM/GPRS along with Radio Frequency Identification (RFID) to monitor attendance system [7]. But this kind of attendance monitoring system is quite expensive and has limited use. To overcome this drawback now we will discuss the speech/voice as biometric trait.

The speech identification mainly depends upon two parameters, one is the feature extraction and another is classification. Normally Mel Frequency Cepstral Coefficient (MFCC) is used for feature extraction but the efficiency, reliability of this algorithm is quite far behind as compared with Gammatone Frequency Cepstral Coefficient (GFCC) even in noisy environment.

When we compare the average accuracy of various MFCC methods based on SNR of test signal ranges from 0 to 40 dB it gives 50.05% efficiency whereas in case of GFCC extraction and CMN normalization of test signal the average accuracy is 55.43% .So one can say the GFCC approach for feature extraction is more reliable and efficient[11].

Now the second phase of speaker recognition is the classification phase in attendance monitoring [12].This classification will act as the main stage for recognition speech from speaker whether it is rejected or accepted. And the best identification technique is Gammatone Mixture Model [13] in comparison with Mel Frequency Cepstral Coefficient (MFCC). Here the correct identification in MFCC using first and second

derivative is up to 90%. Now, there are mainly three main techniques for normalization called MVN (Mean and Variance Normalization), CMN (Cepstral Mean Normalization) and PCA (Principal Component analysis)[15]. This paper compares the performance of these techniques and let us knows which method is best suitable in speech monitoring based on the result findings and one can say based on our result findings CMN gives the best result.

In addition to this for background noise and extraction of voiced portion generally Probability density function (PDF) and Linear Pattern Classifier are used.

This Paper is Categorized in such a way that Section III discuss abut speech recognition system design, Section IV tells us the methodologies applied at various stages in speech recognition, Section V represents the results and Performance and at last Section VI concludes the paper.

2. SPEECH RECOGNITION SYSTEM DESIGN

The method of speech recognition is decided into six different stages which includes Speech Acquisition, Pre Processing of speech signal, Feature Extraction with post processing, Normalization and classification as shown in Fig 1.

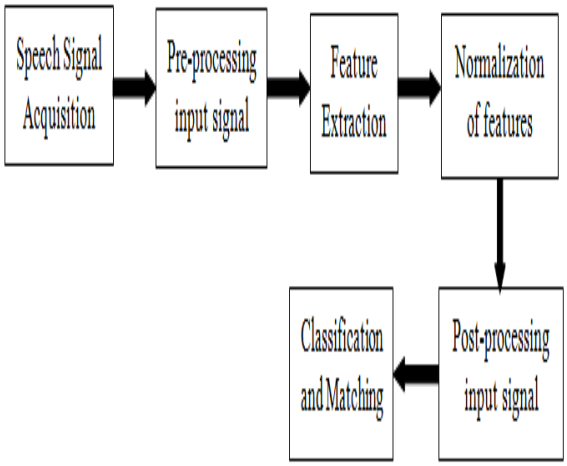

Fig 1:The process of speech recognition

The first most stage in speech recognition is Signal Acquisition is to replace speech signal as shown in Fig 3. The second phase deals with the extraction of speech signal in pre processing of signal. Further is the feature extraction which includes extracting the features of speech signal. Next Stage includes the feature post processing to enhance the results. Now to avoid the risk of greater values in feature vector, the Feature Normalization is used at this stage. And the last stage is the classification which includes making the input signal from stores speech signal in database.

3. METHODOLOGY

Fig2. Depicts the flowchart of speech identification system with different methods used at different stages. The speech of user is captured through system microphone and input signal is generated in first phase. Accordingly in next phase the features are extracted using Gammatone Frequency Cepstral Coefficient (GFCC).Further Normalization of signal is done using Cepstral Mean Normalization(CMN).then delta features are calculated .At last Modeling is done using Gaussian Mixture Model(GMM) and classification using log likelihood logic to monitor the speaker.

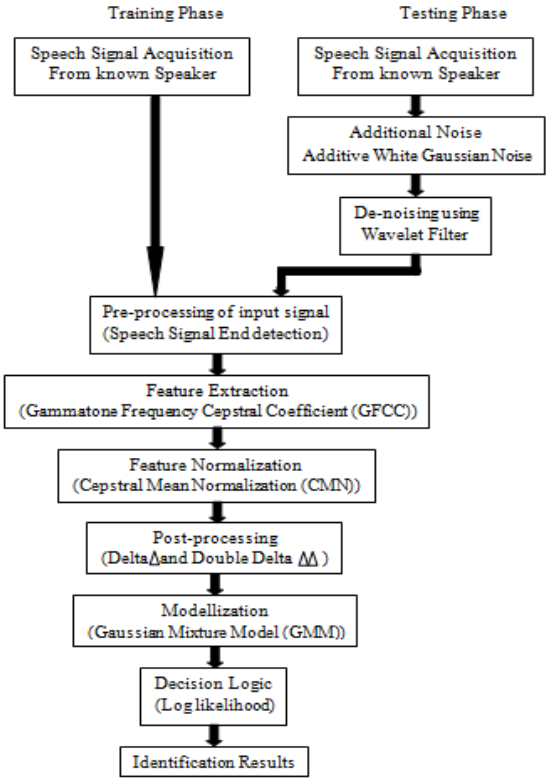

Fig 2: Architecture of robust speech identification

3.1 Speech signal acquisition

The speech of user is captured through system microphone (e.g. laptops or Personal Computer) and input signal is generated in first phase. This is the most Suitable Way for acquiring these biometric traits. Fig 3 shows the acquired signal using laptop microphone.

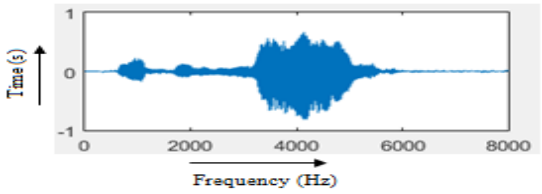

Fig 3: Acquired signal

Generally, signal generated in first phase is analog signal. So one can say we need to convert these analog signals into digital signals. As analog signals are continuous and are required to change into digital form. So in conclusion sampling rate is defined before acquisition of signal and the sample signals are samples at very much lower rate. Generally the speech signals are digitized at 8 kHz.

3.2 Additive White Gaussian Noise(AWGN)

Noise is generally regarded as useless signal. A noise contains the information regarding the source, environment in which it propagates. Sources of noise or distortions are of various types includes [18]:

1. **Electronic noise** which includes thermal noise and shot noise,
2. **Acoustic noise** which generally occurs vibrating or colliding source like by revolving machines, moving wind ,rain etc

3. **Electromagnetic noise** is basically the interfere between the reception of voice, image, data and its transmission, over the radio-frequency spectrum,
4. **Electrostatic noise** occurs due to the voltage
5. **Quantization noise** when packets are lost due to congestion over network.

When the various undesirable changes occur due to various different non-ideal characteristics like echo, reverberation, multiple reflections, etc called Signal distortion. Now on the basis of its time characteristics and its frequency spectrum it is further divided into various categories:

1. **White noise:** It has all frequencies with equal power on the basis of flat power spectrum and impulse autocorrelation.
2. **Band-limited white noise:** This noise is quite similar to white noise having flat power spectrum and having limited bandwidth and covers limited spectrum of devices. The autocorrelation of this noise is sinc-shaped.
3. **Narrowband noise***:* This kind of noise has very narrow bandwidth such as 50/60 Hz from the electricity supply.
4. ***Impulsive noise*:** It is a short duration of pulse having lesser time of occurrence and duration.
5. ***Transient noise pulses*:** These are quite long duration noise pulses like clicks, burst noise etc.
6. **White noise** is generally called uncorrelated random noise having equal power at all frequencies. It is assumed that voice is stationary in additive white Gaussian process. Below Fig.4 shows signal before adding noise and Fig. 5 shows signal after adding noise.

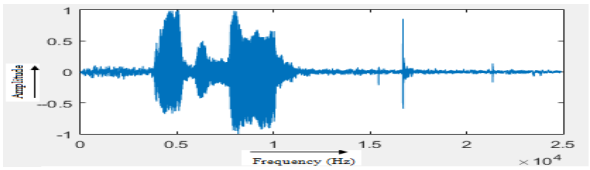

Fig 4: Signal before adding white noise

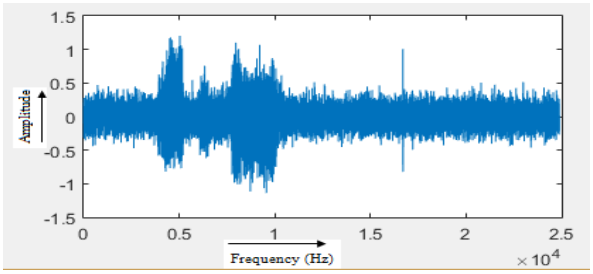

Fig 5:Signal after adding white noise

3.3 De-Noising with wavelet filter

Wavelets are used in wide variety of fields like in physics for the removal of noise from signals. Wavelets are characterized on the basis of position and its scaling

There are different ways to reduce noise in audio. Wavelets are characterized by scale and position, and are useful in analyzing variations in signals and images in terms of scale and position. Because of the fact that the wavelet size can vary, it has advantage over the classical signal processing transformations to simultaneously process time and frequency data. The general relationship between wavelet scales and frequency is to roughly match the scale. At low scale, compressed wavelets are used. They correspond to fast-changing details, that is, to a high frequency. At high scale, the wavelets are stretched. They correspond to slow changing features, that is, to a low frequency [18]. This paper implements coiflet wavelet for filtering. Following steps shows the filtering process.

Step1: Multilevel decomposition

For signals, low frequency content shows the identity of signal and high frequency content shows no important information. If high frequency components are removed, still the words are audible and can be recognized easily. However, if enough low frequency components are removed then signal is not clear. Fig.6 shows steps for signal decomposition.

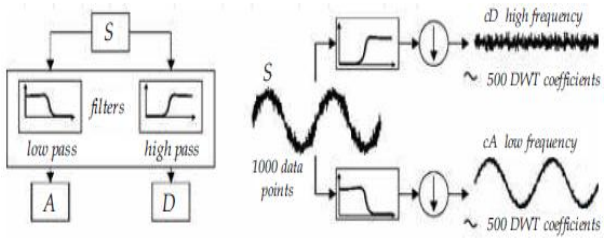

Fig 6: Steps for signal decomposition [18].

Decomposition is done in iterative succession. This paper worked on 10 levels of decomposition. Below fig.7 shows wavelet decomposition tree.

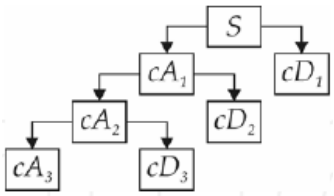

Fig 7: Wavelet decomposition tree [18].

Step2: Wavelet Thresholding

Determine noise threshold by using *wpbmpen* command that returns a global threshold THR for de-noising. THR obtained by a wavelet packet coefficients selection rule by means of using a penalization method provided by Birge-Massart.

Step3: Wavelet reconstruction

Reconstruction is the process of assembling those components back into the original signal without loss of information using threshold values obtained. While being this transformation, it is desirable to establish its investment, i.e. to return to the original signal from the output tree. The mathematical manipulation that affects reconstruction is called the inverse discrete wavelet transforms (IDWT). In order to reconstruct a signal by using Wavelet Toolbox software, reconstruct it from the wavelet coefficients and hard threshold.

3.4 Pre-processing of input speech

In this phase of speech signal the parts containing speech signals are extracted and their silence or unvoiced part is removed. This phase decreases the signal dimension but makes it valuable for next phases and also increases efficiency of system. There are wide varieties of algorithms and techniques present like Short Time Energy (STE) [16], Zero

Cross Rate (ZCR)[16].Now for background noise generally uses Probability Density function (PDF)and for extraction of voiced portion uses Linear Pattern Classifier[17].This algorithm is divided into five phases:

Step 1: Mean (μ) and Standard Deviation (σ) of first 1600 samples (if 8000 is sampling rate) is calculated. Usually, first 1600 samples of speech corresponds silence [17]. Background noise is characterised by Mean (μ) and Standard Deviation (σ). Mean and standard deviation can be written in the form of,

$$\mu = \frac{1}{1600}\sum_{i=1}^{1600} x(i) \quad (1)$$

$$\sigma = \sqrt{\frac{1}{1600}\sum_{i=1}^{1600}(x(i) - \mu)^2} \quad (2)$$

Step 2: For each sample of speech, check for one-dimensional Mahalanobis distance function condition. Analytically, If,

$$\frac{|x-\mu|}{\sigma} > 3 \quad (3)$$

Where x is an observation, then the sample is voiced, otherwise it is unvoiced or silence. Threshold value for voice sample is greater than 99.70 % as in Gaussian distribution and it generally rejects the samples up to 99.70 %.

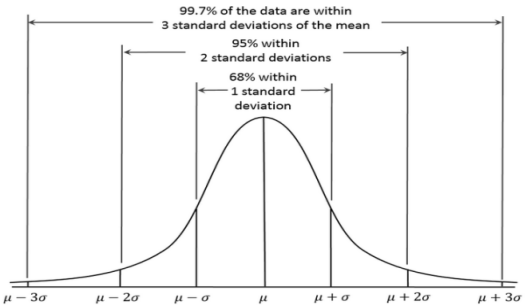

Fig.8 shows the Gaussian distribution for one-dimension.

The probability (P) is given as follows:

$$P(\mu - \sigma \le x \le \mu + \sigma) \approx 0.682 \quad (4)$$

$$P(\mu - 2\sigma \le x \le \mu + 2\sigma) \approx 0.9545 \quad (5)$$

$$P(\mu - 3\sigma \le x \le \mu + 3\sigma) \approx 0.9975 \quad (6)$$

Step 3: Mark the voiced sample as 1 and unvoiced as 0. Divide the whole speech into 5 milliseconds (ms) frames.

Step 4: If number of ones in a frame are greater than the number of zeros, denote frame as voiced; otherwise unvoiced.

Step 5: Collect all the voiced frames (labelled 1) and put in a new array.

Fig.9. Shows an example of speech input signal and Fig.10. Shows voiced portion after implementing the above algorithm.

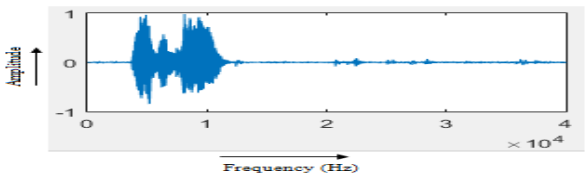

Fig 9: Input signal

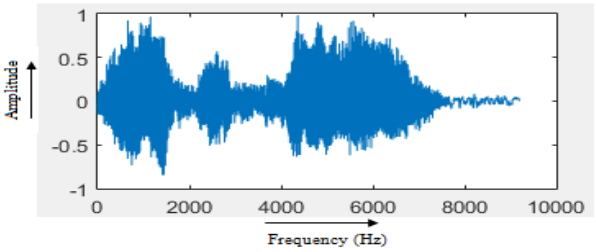

Fig 10:Voiced signal after application of algorithm

3.5 Feature Extraction

With the detection of suitable voiced signal, then the extraction of feature is done from the signal provided extract those features which gives maximum information related to the application and this stage is quite vital in speech recognition. Feature extraction normally extracts those features which are vital irrespective of any non vital information of signal depending upon user application

This stage is the most vital stage in speech recognition. Feature extraction is basically used to remove all the reluctant information of signal depending upon user application and retaining only necessary features which give us enough information to identify the speaker. There are many existing algorithms for feature extraction. In this paper, Gammatone Frequency Cepstral Coefficient (GFCC) is used which works more efficiently in noisy environment as compared to widely used Mel-Frequency Cepstral Coefficient (MFCC) [9].GFCC is Fast Fourier Transformation (FFT) based technique in speaker identification system.

In [9], [10] MFCC and GFCC algorithms are compared which shows that GFCC has a fine resolution at low frequency as compared to MFCC. MFCC works with a log while GFCC works with a cube root. The cube roots provide more robustness to GFCC as compared to that of logs in MFCC. Hence, in noisy environment GFCC is more robust than MFCC. The detailed process of GFCC extraction is listed as follows [9]:

1. Pass the input signal though a 64 channel gammatone filter bank.
2. At each channel, fully rectify the filter response (i.e. take absolute value) and decimate it to 100 Hz as a way of time windowing.
3. Then take the absolute value afterwards. This creates a Time-Frequency (T-F) representation which is a variant cochleagram.
4. Take the cube root of T-F representation.
5. Apply DCT to derive cepstral features

In this feature extraction stage, firstly a bank of gammatone filter is used for decomposing an input signal into T-F domain. These gammatone filters are standard model of cochlear filtering [9]. 64 filters are used with center frequency range [50, 4000]. Then sampling frequency is decimated to 100 Hz along the time dimension. Further, the magnitudes of the decimated outputs are then loudness-compresses by a cubic root operation. This results into matrix representation T-F decomposition of the input (which is a variant of cochleagram).The cochleagram provides finer frequency resolution at low frequencies than at high frequencies as compares to spectrogram. The following fig.11 shows cochleagram of an utterance.

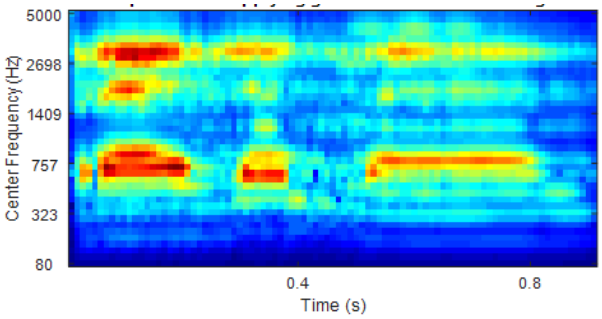

Fig 11: Cochleagram of input speech signal

Then apply a DCT to the GF featured matrix. Further, GFCC features are extracted using GF featured matrix.

3.6 Feature normalization

There are number of normalization techniques. But CMN is the most efficient method of normalization as this method generally reduces environmental distortion. In this method we emphasis on Cepstral Mean Normalization(CMN).but CMN takes the values of the vector having zero mean and one variance and ultimately it reduces to take the risk of taking larger vector values . The Mean of Cepstral Coefficient is invariant in CMN. So by subtracting the mean the irrelevant information is reduced.

In this paper, similar in [11], By subtracting the average of cepstral coefficient from each cepstral coefficient to characterize the speech signal and this is what called Cepstral Mean Subtraction (CMS).. and mean and variance are given as follows:

$$\mu = \frac{1}{N}\sum_{n=1}^{N} C\ (n) \tag{7}$$

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(C\ (n) - \mu)^2 \tag{8}$$

Where, C (n) is a feature vector of n^{th} frame and N is the total number of frames.

$$\hat{C}\ (n) = \frac{C\ (n) - \mu}{\sigma} \tag{9}$$

Where, $\hat{C}$ (n) is normalized feature vector.

3.7 Feature post processing

With the extraction of GFCC and double delta ΔΔ features time dynamics are collected and it was seen that the results of Delta + Double delta + GFCC are quite better if applied GFCC features alone. The first order derivative of input feature vector is computed using Delta function and the accordingly the second order derivative of delta function is calculated using double delta function.

3.8 Classification

With the feature post processing the signal is ready for its pattern matching called classification. The classification plays a very crucial role in speaker modeling. On the basis of this result one can know whether the speech is accepted or rejected. There are number of methods used for Classification of signals. The best method for speech identification is the Gammatone Mixture model.

1. Gaussian Model Description

A Gaussian mixture density is a weighted sum of M component densities, as depicted in Fig. 12 and is given by the equation

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \tag{10}$$

Where $\vec{x}$ is a D-dimensional random vector, $b_i(\vec{x})$, i = 1, 2. . . M, are the component densities and p_i, i = 1, 2. . . M, are the mixture weights. Each component density is a D-variate Gaussian function of the form

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x} - \vec{\mu}_i)'\ \Sigma_i^{-1}\ (\vec{x} - \vec{\mu}_i)\right\} \tag{11}$$

with mean vector $\vec{\mu}$ and covariance matrix $\sum_i$. The mixture weighs satisfy the constraint that $\sum_{i=1}^{M} p_i = 1$.

The complete Gaussian mixture density is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by the notation

$$\lambda = \{p_i, \vec{\mu}_i, \Sigma_i\}\ i = 1, \ldots, M. \tag{13}$$

For speaker identification, each speaker is represented y a GMM and is referred to his/her model λ.

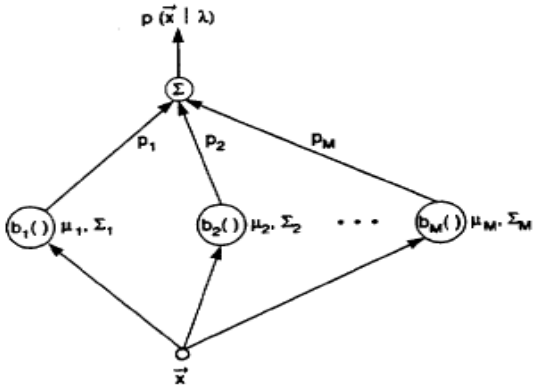

Fig 12: M component Gaussian mixture density [12].

2. Maximum Likelihood parameter Estimation

The main aim of speech mode training is to find out the parameters of the GMM, and further that matches with the distribution of training feature vectors. Nowadays there are number of ways for estimating parameters of a GMM and they are generally estimated using Expectation Maximization (EM) algorithm

In the expectation maximization algorithm the steps of guessing probability distribution with the completion of data as in E step and afterwards re-estimating the model parameters using this data called M step.

In the expectation step, the probability of each data pot is calculated using mean of vectors and its covariance matrices. and is analogous to cluster assignment step with k-means.

In maximization step, the cluster means and covariance are recalculated as earlier done in expectation step.and this step is anlogous to cluster movement step with k-means.

Given the training data X_1, X_2, X_3. . . X_n with the number of mixture M, the parameters λ are generally used in expectation maximization algorithm. During recognition, the input speech is again used to extract a sequence of features X_1, X_2, X_3. . ., X_n and the distance of the given sequence from the model is

calculated by computing the log likelihood of given data. The model with the highest likelihood score will be verified and called as the identity of that speech.

4. EXPERIMENTAL RESULTS

The Foremost part in training phase is to acquire the speech in normal clean environment condition and the database of 15 speech signals are generated with 16 utterances and each signal is used to generate features using GFCC and GMM as used in various phases for its identification. And generally these two experiments are done at testing stage.

In First phase, the signal is added with white gaussian noise d accordingly further steps are performed in training phase and at last phase minimum negative log likelihood is performed to identify the speech.

In Second phase, Speech signal is added with white Gaussian noise along and the wavelet filter and that wavelet filter is used to denoise the signal. And further steps are done in testing phase without using filters.

Here the accuracy of speech identification is measured by comparing the feature extraction features for speech i.e using GFCC. So one can say that the correct recognition rate (CORR) is given as:

$$\%age\ CORR = \frac{No.of\ speech\ sample\ correctly\ classified}{No.of\ total\ samples} * 100$$

In given figures the CORR for various person are tabulated. Fig 13 depicts the CORR rate without using wavelet filter of 10 signal to noise ratio (SNR). And Fig 14 shows the CORR ratio with using wavelet filter of 10 signal to noise ratio (SNR). And finally the fig 15 shows the CORR rate comparison of 10 Signal to noise ratio (SNR) ratio between GFCC along with wavelet filter and without wavelet filter.

Table 1 formulates the average identification rate for stored database and also for various stored database oriented experiments. While using wavelet filtering the database oriented identification rate is 84.50% and whereas with using wavelet filtering the average identification rate is lesser i.e. 73.33% and with real database oriented experiment using wavelet filter the average identification rate is 74% and when wavelet filter is not used the identification becomes quite lesser i.e. 45%.

So the above results clearly defines that GFCC algorithm works with more efficiency with using wavelet filter for speech recognition as compared with GFCC without filter. From the graph one can say that the wavelet filters works quite well even in noisy environment and filter out the added noise efficiently. In all categories of speech identification one can say the wavelet filter works in efficient way and gives far better results as compared without using wavelet filters.

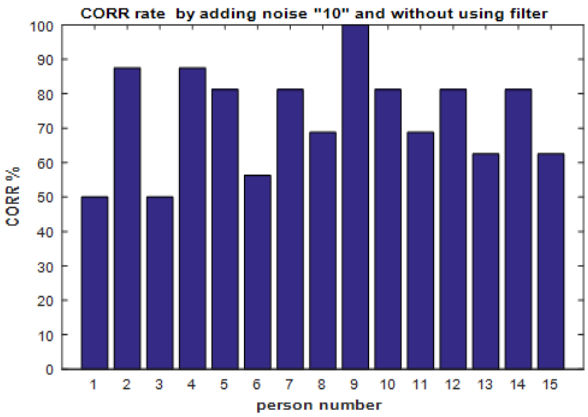

Fig 13: CORR rate by adding noise of 10 SNR without using wavelet filter.

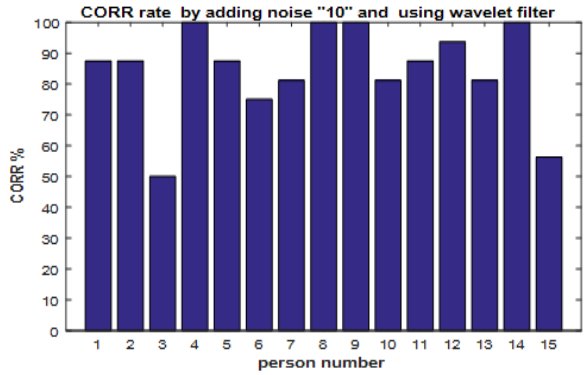

Fig 14: CORR rate by adding noise of 10 SNR and using wavelet filter.

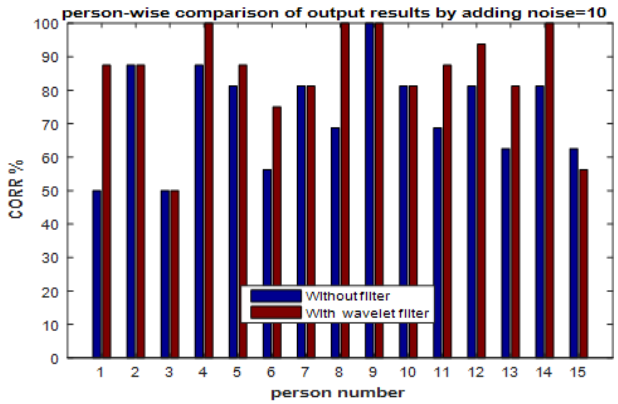

Fig 15: CORR rate comparison when wavelet filter used and when wavelet filter not used.

Table 1:Percentage of correctly recognized speakers in 10 SNR noise corresponding to feature extractor GFCC and classifier GMM along with sound detection, CMN, additive white gaussian noise (AWGN) and wavelet filter for denoising

Database	Average Correct Recognition rate		
	Noise (SNR)	***Algorithm without using filter for de-noising***	***Algorithm with wavelet filter for de-noising***
Stored database (15	10	73.33%	84.50%

Database	Average Correct Recognition rate		
	Noise (SNR)	***Algorithm without using filter for de-noising***	***Algorithm with wavelet filter for de-noising***
speakers with 16 utterances each)			
Real database (15 speakers with 16 utterances each)	10	45%	74%

5. CONCLUSION

In this paper, we have proposed a hybrid approach for speech identification at different phases. With four different phrases of two words pattern of fifteen people are collected like Present mam , present sir, yes mam and yes sir in this paper. And in next stage using gamma tone filter 13 GFCC features are extracted from these input signals. And accordingly normalization technique is applied i.e. CMN on GFCC features. With the help of various GFCC features, first order and second derivative features are also extracted. And accordingly all three matrices of 13 vectors are combined and are used as speech identification. And further the gaussian mixture model is generated using 8 gaussians. And this process is repeated for all values in the database and for 240 samples a total of 240 gmm models are generated. Further the input to the sample is done by adding white gaussian noise at 10 SNR and with the help of wavelet filters after adding noise or without using filter the testing sequence is generated. And in our experiment results it has formulated that the above used algorithm gives 85% CORR rate with using filter and 73% without using wavelet filter. with real database oriented experiment using wavelet filter the average identification rate is 74% and when wavelet filter is not used the identification becomes quite lesser i.e. 45%.In future this algorithm can be advanced for text dependent or in noisy conditions.

6. REFERENCES

[1] Nur Izzati Zainal, Khairul Azami Sidek, Teddy surya Gunawan, Hasmah Mansor, and Mire Kartiwi, “Design and development of portable classroom attendance system based on Arduino and fingerprint Biometric”, IEEE international conference on information and communication Technology, 2014.

[2] Engr. Imran Anwar Ujan and Dr. Imdad Ali Ismaili, “Biometric Attendance System”, IEEE International Conference on Complex Medical Engineering, 2011.

[3] Tsai-Cheng Li, Huan-Wen Wu, and Tiz-Shiang Wu1, “The study of Biometrics Technology Applied in Attendance Management System”, IEEE International Conference on Digital Manufacturing & Automation, pp. 943 – 947, 2012.

[4] Teh Wei Hsiung and Shahrizat Shaik Mohamed, “Performance of Iris Recognition using Low Resolution Iris Image for Attendance Monitoring”, IEEE International Conference on Computer Applications and Industrial Electronics, 2011.

[5] Mashhood Sajid, Rubab Hussain, and Muhammad Usman, “A Conceptual Model for Automated Attendance Marking System Using Facial Recognition”, IEEE International Conference on Digital Information Management, 2014.

[6] Subhadeep Dey, Sujit Barman, Ramesh K. Bhukya, Rohan K. Das, Haris B C, S. R. M. Prasanna, and R. Sinha, “Speech Biometric Based Attendance System”, IEEE National Conference on Communications, 2014.

[7] Aamir Nizam Ansari, Arundhati Navada, Sanchit Agarwal, Siddharth Patil, and Balwant A. Sonkamble, “Automation of Attendance System using RFID, Biometrics, GSM Modem with .Net Framework”, IEEE International Conference on Multimedia Technology, pp. 2976 – 2979, 2011.

[8] Balazs Benyo, Balint Sodor, Tibor Doktor, and Gergely Fordo, “Student attendance monitoring at the university using NFC”, IEEE, pp. 1 – 5, 2012.

[9] Zhao X., Shao Y., and Wang D.L., “CASA-based robust speaker identification”, IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, pp. 1608-1616, 2012.

[10] X Zhao, and DL Wang, “Analyzing noise robustness of MFCC and GFCC features in speaker identification”, IEEE International conference on acoustics, speech and signal processing, pp. 7204–7208, 2013.

[11] El Bachir TAZI, Abderrahim BENABBOU, Mostafa HARTl, "Efficient Text Independent Speaker Identification Based on GFCC and CMN Methods", IEEE International Conference on Multimedia Computing and Systems, pp. 90 – 95, 2012.

[12] Douglas A. Reynolds, and Richard C. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models”, IEEE Transaction Speech and Audio Processing, Vol. 3, pp 72–83, 1995.

[13] Liu Jiqing, Dong Yuan, Huang Jun, Zhao Xianyu, Wang Haila, "Sports audio classification based on MFCC and GMM", IEEE International Conference Broadband Network & Multimedia Technology, pp. 482 – 485, 2009.

[14] Md Jahangir Alam , Pierre Ouellet, Patrick Kenny, Douglas O’Shaughnessy, “Comparative Evaluation of Feature Normalization Techniques for Speaker Verification”, Nonlinear Speech Process., pp. 246–253, 2011

[15] Jelil S, Kachari G, and Joyprakash Singh, “Comparative evaluation of feature normalization techniques for voice password based speaker verification”, IEEE National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics, pp. 1-4, 2013.

[16] Madiha Jalil, Faran Awais Butt, and Ahmed Malik, “Short-Time Energy, Magnitude, Zero Crossing Rate and Autocorrelation Measurement for Discriminating Voiced and Unvoiced segments of Speech Signals”, IEEE International Conference on Electronics and Computer Engineering, pp. 208 – 212, 2013.

[17] G. Saha, Sandipan Chakroborty, and Suman Senapati, “A New Silence Removal and Endpoint Detection Algorithm for Speech and Speaker Recognition Applications”, Department of Electronics and Electrical Communication

Engineering Indian Institute of Technology, Kharagpur, Kharagpur-721 302, India.

[18] Adrian E. Villanueva- Luna, Alberto Jaramillo-Nuñez,Daniel Sanchez-Lucero, Carlos M. Ortiz-Lima, J. Gabriel Aguilar-Soto, Aaron Flores-Gil and Manuel May-Alarcon, "De-Noising Audio SignalsUsing MATLAB Wavelets Toolbox", www.intechopen.com.

[19] Anil K. Jain, Arun Ross, and Salil Prabhakar, "An Introduction to Biometric Recognition", IEEE Transactions on circuits and systems for video technology, vol. 14, 2004.

[20] Malcolm Slaney, "An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank", Advanced Technology Group, Apple Computer, 1993.