Enhancing Security via Speaker Recognition

Affaf Khan[1,*], Arqam Nawaz Babar[2], Ali Iqrash[3], Akbar Ali[3]

[1] Department of Biomedical Engineering and Sciences, National University of Sciences and Technology, Islamabad, Pakistan.
[2] Gina Cody School of Engineering, Concordia University, Montreal, Montreal, Canada.
[3] Department of Electrical Engineering, Pakistan Institute of Engineering and Technology, Pakistan.

* Corresponding author. Email: babaraffaf@gmail.com (A.K.)
Manuscript submitted September 22, 2023; accepted November 23, 2023; published December 8, 2023.
doi: 10.18178/jaai.2023.1.3.154-164

Abstract: In this era of globalization, networking, and the bulk of the information they demand reliable identity verification is quintessential. An efficient means to do this is using biometrics as a method of identification. The current biometric method, voice biometrics, may be an inexpensive and precise confirmation technology widely utilized in recent decades. Voiceprint, human biological characteristics, possess unique physiognomic features for every individual who is difficult to counterfeit, imitate, and replace. As a non-contact identification technology, the users are accepting Voice Recognition Technology and extensively deployed in authentication and assisting systems. The procedure of recognizing or refusing the individuality claim of a user is established on the individual's exclusive information present in the speech wave shape. Recently it received expanding interest in the past two decades, as an accessible, comprehensible way of substituting (supplementing) basic password-type matching.

Keywords: ASR, LBG, MATLAB, MFCC, VQ, LBGVQ

1. Introduction and Background

Humans have the innate ability to recognize familiar voices within seconds of hearing a person speak. Research in speaker verification, the computational task of validating a person's uniqueness based on their tone of voice, began in 1960 with a model based on the analysis of X-rays of individuals making specific phonemic sounds. Many researchers used observed patterns by HMM from the 1970s. With the advancements in technology over the past 50 years, robust and highly accurate systems have been developed with applications in automatic password reset capabilities, forensics, and home healthcare verification.

There are two phases in a speaker recognition system: an enrollment phase where speech models from different speakers are turned into models and the verification phase where a sample of speech is tested to determine if it matches the proposed speaker. It is assumed that each speech sample pertains to one speaker. A robust system would need to account for differences in the speech signals between the enrollment phase and the verification phase that are due to channels used to record the speech (Landline, mobile phone, handset recorder) and inconsistencies within a speaker (health, mood, effects of aging) which are referred to as channel variability and speaker dependent variability respectively.

Speaker recognition is a biometric method that utilizes the characteristics of a person's voice for recognition objectives. It should not be confused with "speech recognition" which is the science of translating the human voice into text or commands. Speaker recognition is one of the few biometrics that relies on both physiological characteristics i.e. the structure of the vocal tract and behavioral characteristics i.e. The

pronunciation of words. As mentioned in "Fig. 1".

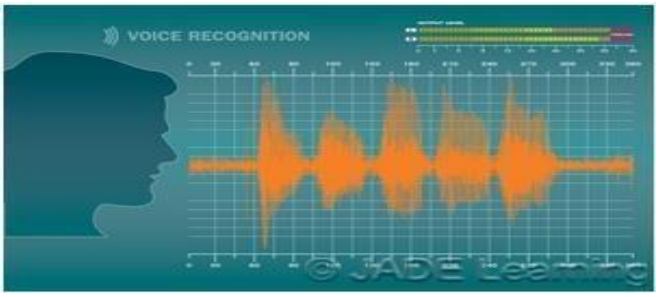

Fig. 1. Voice Recognition.

Employing biometry i.e., the distinctive attributes of a user, i.e., voice, vein patterns, and fingerprints. An effective method for enhancing safety. The biometric parameters can neither be erased nor replicated and are further tricky to manage in comparison to conventional security systems. A system that can identify a person using the distinctive acoustic traits of the user's voice is known as a voice recognition system. It has a huge sort of functions inside the subject of safety such as permitting system access, a safe place, and imparting offerings inclusive of digital banking, voice calling, and reconnaissance. [1]

The important thing in designing any recognizer is the extraction and selection of features of the input signal which represent that signal and will help in recognition. For this MFCC is proved as the best approach in which cepstrum coefficients are expressed on the Mel scale.

In the paper, the Speaker Verification applying Mel Frequency Cepstral coefficients/vector Quantization for the text "My voice is my password" is acquired. The experimental results are scrutinized using MATLAB and discussed in Table 1.

1.1. Literature Review

Vibha Tiwari designed a text-dependent identification system using the techniques Mel-frequency Cepstral Coefficient (MFCC) and Vector Quantization (VQ). Number of filters and windows is varied to produce the best combination. Accuracy of 85% was achieved by choosing 32 filters and a hanning window. The techniques were implemented on MATLAB using five users.

Arun Kumar Chaudhary, Jitendra Kumar Mishra proposed their method which is an MFCC centered speaker verification in a noisy atmosphere using VQ. The research evaluation was conducted on the speech signal in a daily noisy environment. They evaluated the techniques on MATLAB and got a False Rejection Rate (FR) of 15%, True Acceptance Rate (TA) of 96%, and False Acceptance Rate (FA) of 9%.

Maurya *et al*. used Hindi samples of speech to implement speaker identification with the techniques (MFCC- VQ) and (MFCC-GMM) for text-dependent and text- independent sentences. For a text independent system, the efficiencies are (MFCC-VQ) yields 77.64%, (MFCC-GMM) yields 86.27% and for a text dependent system the efficiencies are (MFCC-VQ) yields 85.49%, (MFCC-GMM) yields 94.12%. They implemented these systems on 15 users such as

10 males and 5 females. However, they concluded that according to psychophysical studies the voice of a person could change over a period of two or three years, so the system must be trained repeatedly.

Gaikwand *et al*. presented the paper on various techniques developed at every stage of speech recognition. For feature extraction few techniques are proposed such as Linear Discriminate Analysis, Principal Component Analysis, Filter bank analysis, Mel-frequency cepstrum, Wavelet and Kernel based feature extraction method. For feature modeling such techniques are proposed Pattern Recognition method, acoustic-phonetic, method based on templates, Dynamic timewarping, Approaches based on Knowledge and

Hidden Markov method. Whole word matching and Sub word matching are the techniques proposed for feature matching. The best speech recognition system among the explained techniques for the Marathi language are MFCC and HMM onthe basis of speed and accuracy.

Wali *et al*. presented a neural network-based speakerverification system using voice modality. MFCC-VQ technique is applied primarily then multilayer feed-forward neural network classifier using back propagation is used and the recognition accuracy for 10-50 users are Successful Acceptance Rate (92% to 70%) , False Rejection Rate (4% to 12%). Pass phrase used for the testing purpose is 'Basareshwar Engineering Collage'.

Chauhan *et al*. proposed an approach on MFCCprimarily based speaker classification in blaring atmosphere applying wiener filter. MFCC delivers efficient outcomes in a clear environment although cut out the results in loud environment. Using the recommended method in a noise mismatch situation with the classifier neural network and attained 88.57% accuracy in perceiving a person in a loudenvironment. It was examined that the speech signal is sharperthan the noise signals the Signal to Noise Ratio could be more than0dB so the proposed method provides efficient results for identification.

Fang *et al*. did a comparison of different implementations of MFCCs. It depends on the number of filters, the shape of the filters, the spacing of the filter, and the way power spectrum is wrapped. Several evaluation experiments are done to get the best implementation.

A Srinivasan aims to implement speaker identification/verification system using the techniques MFCCand VQ. The correspondence 'ZHA' (Tamil language) used as a pass phrase. 6 speakers (3 males and 3 females) took partin the research and experimented for different frequencies onMATLAB. Number of speakers can be extended to n using the above system.

Kumar *et al*. applied MFCC and VQ techniques to model a voice identification system. Spectral Subtraction was used to overcome the environmental noise. Intraining phase, the sampling frequency was 22 kHz and 8-bitencoding for the optimal clarity. 6 users (4 males, 2 females)were put to gain access to the system for 10 intervals each leading to 60 trials altogether. Efficiency in low noise environment was 80% and in moderate noise environment was 73.3%. Further explained that the system's correctness for classifying woman is barely lesser than men due to the inheritproblem of utilizing MFCCs.

Mohd *et al*. designed a biometric voice recognition system for the security of moderate level. MFCCand VQ techniques are used and are implemented on MATLAB and ARDUINO. MATLAB is used for signalprocessing and the latter is used for communication system i.e., LED control switch, display ON/OFF etc. The signal is sampled on 10 kHz with the duration period set as 1 second. In this paper, single pass phrase 'Hello' has been used for thetesting purpose. The system has been trained by a single person (admin) and evaluated by admin and other persons which gave the efficiency of 75%. Few recommendations to upgrade the precision of the system are to increase the complexity of the system, noise must be filtered completely, and more set of data can be obtained to gain accuracy.

Muda *et al*. explained voice recognition technique using MFCC/DTW. Purpose of DTW is to create warping function that reduces the overall distance between the respective points of the trained and testing signals. Several other feature matching methods such as LPC, HMM and ANN are under investigation to produce efficient systems.

Liu *et al*. presented their paper which has aunified framework that performs both speaker and utterance verification at the same time. The drawback of this techniqueis the performance of system degrades as the test/train data reduces. The system is developed using stack of long short- term memory networks (LSTM) and deep neural network (DNN) that learns both speaker and utterance models from thetraining data. Using Speaker and utterance verification (SUV) system, security levels can be altered as per requirement as it contains layered verification. The database of the system comprises of 300 speakers data from 143 female and

157 male speakers. The database has three different parts: part 1 has 30 fixed phrases utterance of 3-4 seconds, part 2 has 30 fixed short commands of 1-2 seconds and part 3 has random digit based fixed sequence. They compiled results and compared them with other system i-vector and HiLAM system and concluded that SUV gives better results.

Yang *et al*. proposed an advance method termed as sub band transformation. This technique captures artifacts more accurately in synthetic speech as compared to band transformation. An iso-sub band transform was proposed by them for constant sub band transformations. Studies have showed that sub band transformation function are more capable when compared with full-band transformation in both clean and noisy conditions. (Jichen Yang, 2019)

Wu *et al*. employed a technique called functional unification as they believed the differences between natural and synthetic speeches distribution is a significant discriminatory parameter. The employed technique performed authentication using convolutional neural network through the characteristics of natural voices. (Zhenzong Wu, 2020)

Park *et al*. proposed a highly secure user identification technique by employing deep learning technology. The user identification model employed MFCC feature (Mel-frequency Cepstral Coefficient) and after conducting various experiments under multiple environmental conditions, the model was found to efficiently distinguish each registered personnel. Along with this, a synthetic speech recognition technique was also used together to investigate the masquerading attack. (Hyun Park, 2021)

2. Speaker Recognition

Speaker/voice recognition is a biometric mode whichemploys the user's voice for recognition purposes. In other words, it is an automated method of authentication of identity of a person by his voice. The aforementioned process incorporates the physical structure of vocal tract and behavioral characteristics of the individual. There is a sharp deviation between speaker verification and recognition. The former refers to recognizing who is speaking or registering asound while the latter refers to recognizing the same sound and differentiate from unknown sounds. As described in “Fig.2”.

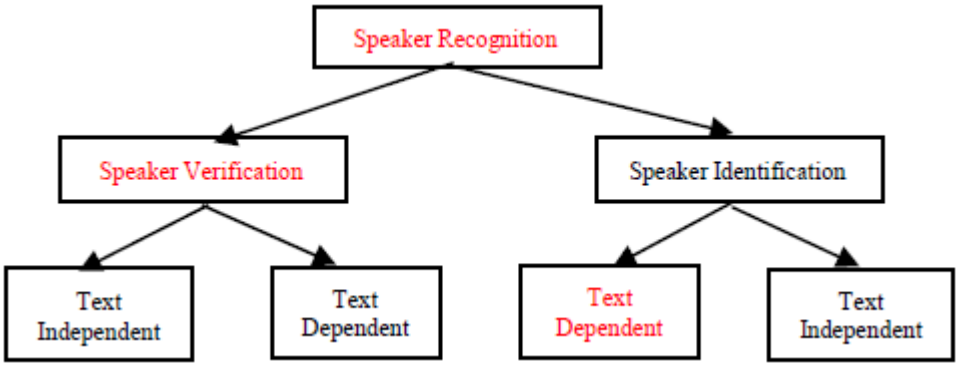

Fig. 2. The scope of speaker recognition.

2.1. Speaker Recognition and Verification

The task of verifying the identity of anonymous speaker issaid identification. Speaker identification is a type of identification where the voice is matched alongside N templates using the ratio 1:N.

The authentication or verification is to verify the claim ofidentity of certain individual i.e. the 1:1 match where the speaker’s voice is matched with the stored template model.

Speaker recognition system comprise of two types: Text-independent and Text-dependent.

Text-Dependent: If the model used for registration and authentication is text then it is termed as

text-dependentrecognition. Prompts in text dependent systems can also havedaily life communication among speakers such as a familiar passphrase or can be exclusive such as password/ pin.

Text- Independent: If the model used for registration and authentication is not text then it is called text-independent recognition and it requires very less effort from the speaker. The text in authentication and during test is different.

In a text-independent systems mutually acoustics and speech analysis techniques are used. This project intends to model a text-based voice detection system to be employed to ensure acertain user's system that the user can only retrieve it.

Speaker Recognition functions in two phases : Training and Testing mode. During Training mode, feature vector set of training voice is obtained from speaker's training waveforms, while in testing mode the speech sample given is matched with the remaining samples in a database. The system will give the consumer access after attaining a match.

2.2. Feature Extraction

Original signal of speech is characterized by feature vector's sequence which are calculated to give the feature extraction of sound. Mostly, it occurs in three stages.

The first step known as acoustic front-end/speech analysis produces the basic features defining power spectrum's wrap of short speech periods by performing the spectra-temporal analysis of speech signal. The main purpose of first step is to take out the features of speech in signal and extracts a frame every 16-32 msec and performs spectral analysis by updating every 8-16 msec. The acoustic front-end contains the following algorithmic blocks:

1) Fast Fourier Transformation (FFT),
2) Calculation of logarithm (LOG)
3) Discrete Cosine Transformation (DCT)
4) Linear Discriminate Analysis (LDA).

The second step comprises of the compilation of static and dynamic features collectively in extensive feature vector.

At last, the results from second step are optimized, compacted and then forwarded to the recognizer.

The speech extraction can be carried out by various algorithm which work efficiently with certain conditions such as PLP and RASTA-PLP (Relative Spectra Filtering of log domain coefficients) which are quite good in noisy conditions. Mostly the feature used for speech extraction are cepstral coefficients obtained by Linear Predictive Coding (LPC). Mel-frequency Cepstral Coefficients (MFCC), PLP and LPC are the most extensively used parameters in area of speech processing.

2.3. Feature Matching:

The vast topic of science and engineering, Pattern Recognition encompasses the problems of speaker recognition and brings the solution to them. The classification of objects of interest also known as patterns into number of categories or classes is one of the major goals of pattern recognition. The models or sequence of acoustic vectors thatare obtained using the abovementioned techniques from the input speech signal are used and the individual speakers are represented by classes. As the classification technique is utilized for feature extraction, it can also be known as FeatureMatching.

Similarly, if some patterns are previously known or enrolled, then they can have a difficulty in organized patternidentification. Training sets are constituted using classified algorithm derived by using such patterns. The rest of the patterns collectively referred to as test sets are applied to testthe classification algorithm. If the precise classes of certain patterns in test set are known, then the algorithm's performance can also be estimated.

Most common feature matching techniques used in speaker recognition are:

1) Artificial Neural Network (ANN)
2) Hidden Markov Modeling (HMM)
3) Dynamic Time Wraping (DTW)
4) Euclidean Distances (ED)
5) Vector Quantization (VQ)

3. Methodology

3.1. Mel-frequency Cepstral Coefficients (MFCC)

The Method meant for speech extraction in this project is**Mel-frequency Cepstral Coefficients (MFCC)**. It constitutes on human hearing observations which cannot perceive frequencies over 1KHz. Sequent, MFCC comprisesof identified variation of critical bandwidth with frequency ofhuman ear. It has two types of filters: one with linear spacingat frequency lower than 1000 Hz and one with logarithmic spacing at high frequency higher then 1000Hz. A skewed pitch is present on Mel Frequency Scale to describe crucial trait of phonic in speech.

A compact illustration would be given by a set of MFCC short-term energy spectrum resulted from cosine transform expressed on Mel-frequency scale. The performance of MFCC might be altered by the number of filters, the shape of filters, the way of filter spacing and the way of power spectrum warping. The conventional MFCC calculation excludes the 0th coefficient. As mentioned in "Fig. 3".

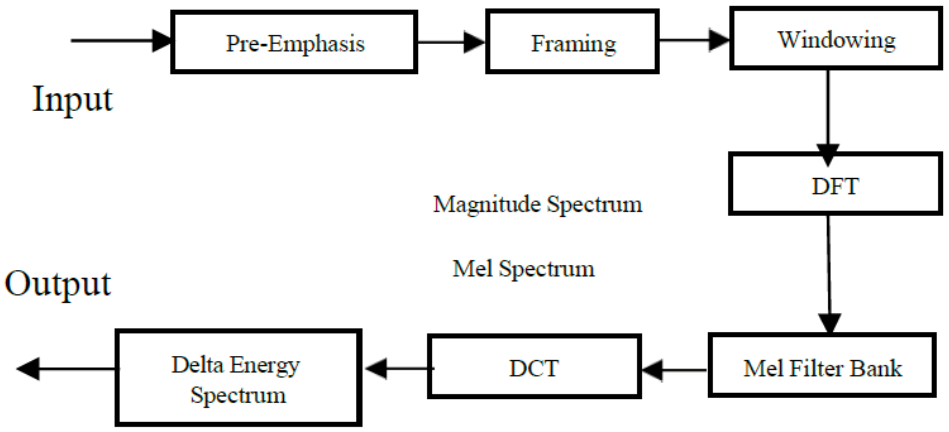

Fig. 3. Computational Steps of MFCC.

Purpose of each step and mathematical methodologies asconsidered in brief:

Pre-emphasis: Higher frequencies passing via filter and will raise the energy of those signals are described in the step.

$$Y[n] = X[n] - (0.95 \times X[n-1]) \quad (1)$$

Framing: Segmenting of the speech models acquiredfrom ADC into the length of 20-40 msecs with small frames. Voice signal is split into N samples of frame.M (M<N) is separating adjacent frames. M = 100 andN= 256 are typically used values.

Hamming Window: Window used to block subsequent frequencies in feature extraction process chain and integrates all the closest frequency lines. Mathematical derivation of hamming window is described as:

If the window is defined as W (n), $0 \leq n \leq N-1$ where N = Number of samples in each frame

Y[n] = Output signal

X (n) = Input signal

W (n) = Hamming window,
then the result of windowing signal is shown below: Y(n) =X(n) × W(n)

$$W(n) = 0.54 - 0.46 \cos\left[\frac{2\pi n}{N-1}\right] \quad 0 \leq n \leq N - 1 \tag{2}$$

"Fig. 4" is explained below.

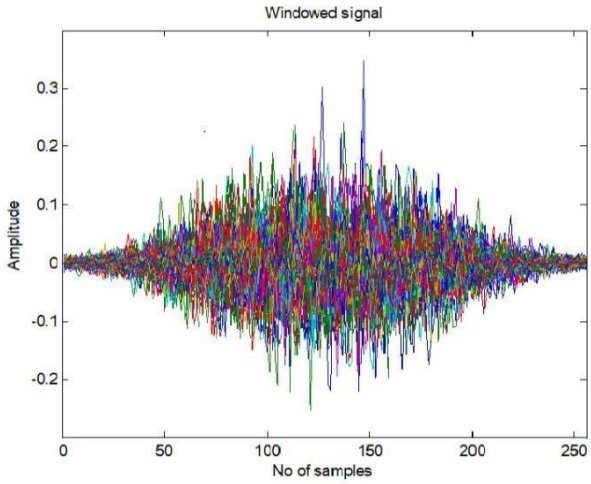

Fig. 4. Frames of input signal after hamming window.

Fast Fourier Transform: In frequency domain, eachframe of N samples is converted from time domain. The conversion of glottal pulse U[n] and the vocal tract impulse response H[n] in the time domain is termed as FFT. This equation fits the description of FFT.

$$Y(w) = FFT\,[h(t) \times X(t)] = H(w) \times X(w) \tag{3}$$

If X (w), H (w) and Y (w) are the Fourier Transformof X (t), H (t) and Y (t) respectively.

Mel Filter Bank Processing: In FFT, the signal of the voice does not follow the linear scale as thespectrum of frequency is wide. According to Mel scale, filter bank is performed and are indicated in figure. "Fig. 5" described as follows.

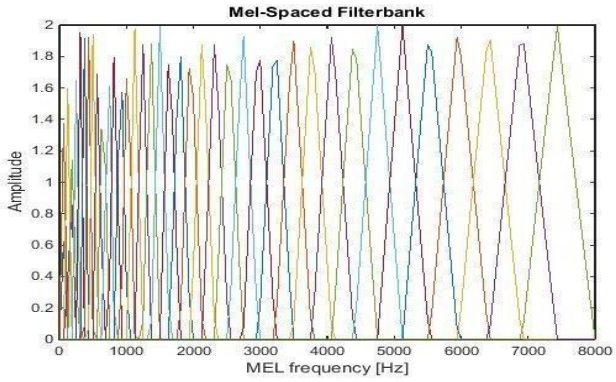

Fig. 5. Triangular Filters.

This picture displays a set of triangular filters which might be used to approximate the weighted sum of filter spectral components. The magnitude of frequency response is triangular at the center frequency and

equals to unity while a linear decline tozero at two adjacent filter's center frequency can be seen. The output represents the sum of filtered spectral components. Afterward, the subsequent equation is applied to compute the Mel for particularfrequency f in HZ:

$$F(Mel) = [2595 \times \log_{10}[1 + f/700] \quad (4)$$

Discrete Cosine Transform: Using DCT, log Mel spectrum is converted into time domain. Mel Frequency Cepstrum Coefficient are obtained as a result of this conversion. Each input utterance is therefore transformed into an acoustic vector's sequence and is known as acoustic vectors.

3.2. Vector Quantization

Feature matching technique used in this project is **Vector Quantization**. It is a procedure to represent vectors from a large vector space to a limited amount of regions in that space where each region called a cluster can be interpreted by its center known as a code word and the accumulation of all code words is said to be a codebook. The recognitionprocess is delineated by the diagram given below where two speakers and dimensions of acoustic space are shown. The acoustic vectors and result code words from speaker 1 and 2 are represented by circles, black circles and triangles, black triangles in Fig. 6 respectively. Using clustering algorithm, a speaker specific VQ codebook is produced for each speaker by gathering its training acoustic vectors in training phase. The VQ distortion is defined as the distance from a vector to the closest centroid of a code book. In the recognition phase, an input utterance of an unknown voice is "vector-quantized" using each trained codebook and the total VQ distortion is calculated. Speaker of original speech signal is identified by comparing the VQ codebook and the amount of total distortion. As mentioned in "Fig. 6".

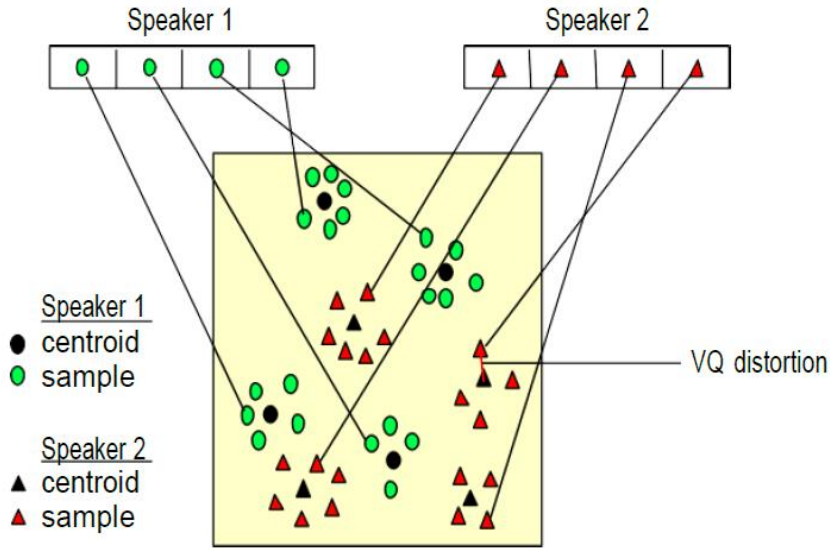

Fig. 6. Illustration of vector quantization codebook formation.

Location of centroids one person can be distinguished from another person.

Training vector‟s set for each user is taken by acoustic vectors obtained from input speech signal after the enrolment session. As stated above, there is a need to develop a user identifiable VQ codebook for each speaker using training vectors. A set of L training vectors are clustered into set of M codebook vectors using LBG algorithm.

The M-vector codebook in LBG algorithm is formed in stages. Primarily, it designs a 1 vector codebook which is further proceeded by using a splitting technique on code words to initiate the search for 2 vector codebook and extends the splitting till the desired M-vector codebook is received.

The aforementioned LBG algorithm is also portrayed by a flow diagram in Fig. 7. “Cluster vectors” is the procedure of nearest-neighbor search in which each training vector is assigned to a cluster associated with the closest code word. Upgradation of centroid is “Find centroids.” Sum of the distances determines the convergence of all training vectors in nearest-neighbor search i.e., “Compute D (distortion).” “Fig. 7” explained as follows.

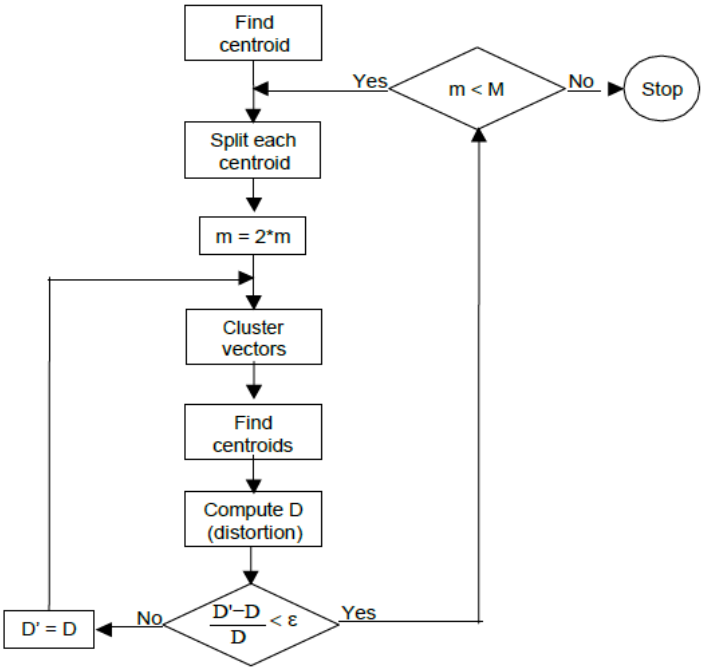

Fig. 7. Flow diagram of LBG algorithm.

4. Results

After implementation on MATLAB, we tested it on 60 students each with 10 voice samples. For training and testing,we used “My voice is my password” as a passphrase for the results there.

FA: False Acceptance, **FR**: False Rejection, **E:** Efficiency.

4.1. Results

Table 1. Results of 600 Samples

No. of samples	FA rate	FR rate	efficiency
200	2.5%	10.5%	87%
400	3.25%	8%	88.75%
600	3.34%	7.5%	98.16%

4.2. Comparative Analysis and Discussions

In recent years, voice recognition systems based on human unique biometrics have caught more attention due to high efficiency, reliability, and security. In our project, efficiency has been increased by using the same techniques with a little enhancement i.e., varying the number of centroids and filters hence found the best combination with higher efficiency and less computational time which has been applied on a higher number of users. The number of centroids is16, 24, and 32 with the combination of filters 20, 30, and 40. Altogether 9 combinations were produced, the best combination concerning computational speed was 16centroids and 30 filters but the efficiency was low. To improve efficiency computational speed was traded off and the

combination of 16 centroids and 40 filters was used.

In this section, a comparative analysis in Fig. 8 of different feature extraction techniques with their recognition rate has been discussed and a comparison has been made with the previous research papers that use the same techniques with a smaller number of users but have lower efficiency. The maximum number of users taken by Nair *et al*. [1] was 6 samples but this project has extended that approach and tested the signal on 60 samples.

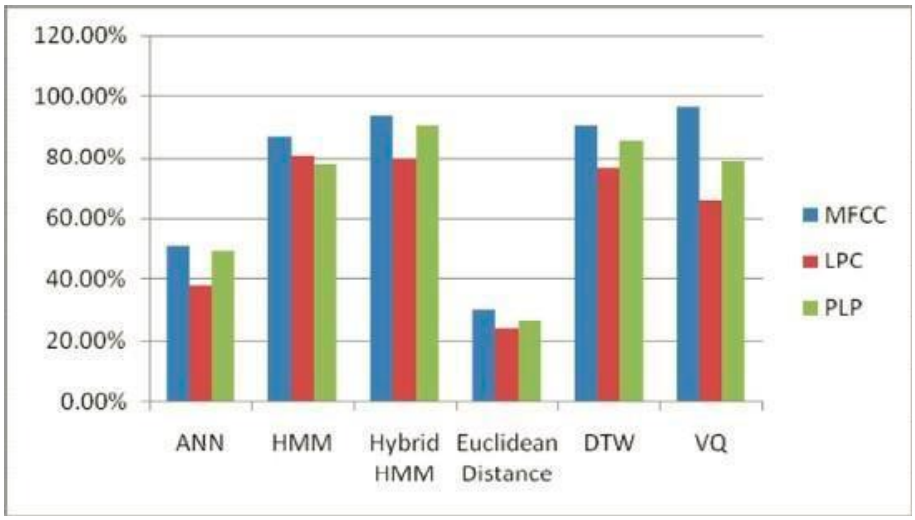

Fig. 8. Comparison of recognition rate with different feature extraction techniques.

According to their experimentation and results [1], the total accuracy of the system was only 73.3% along with a false acceptance rate of 11.67%. This means if the user claims an identity and the system grants false access. The false rejection rate is 15% which means if the user wants verification, then the system will reject the actual user at the rate of 15%.This system is therefore not an efficient one.

In a recent study by S. S. Wali and S. M. Hatture S. Nandya [2], used 50 users and implemented MFCC with BPNN and their efficiency was 78.4%.

In another study by Vibha Tiwari [5], she took 5 samples and implemented the same techniques with a little variation in the number of filters, but the efficiency was 65%.

Experimental results conducted by A. Srinivasan [7] using the MFCC with VQ techniques generate a moderate efficiency system. It used a text-dependent system and used the Tamil language word 'Zha' as a passphrase.

5. Conclusion

The project has deployed the effective algorithm MFCC and LBG VQ for speaker recognition. The MFCCs were calculated for each speaker and LGBVQ was used to form the vectors. The foundation of the speaker's verification was formed by VQ distortion between the MFCCs of an unidentified speaker and the sequent codebook. MFCC was chosen as they conform to the response of the human ear, but the execution was limited by a single coefficient bearing high VQ distortion with the representing codebook.

The usage of high-end audio devices in a noise-free environment can enhance the performance factor. The possibility of using the recorded speech in lieu of the original speaker has been overcome by employing MFCCs as the original signal and the recorded signal has different MFCCs. Psychophysical studies have demonstrated the chances of variance in human speech over the course of 2-3 years. Therefore, the user's codebook in the database is required to be updated via training sessions.

In the end, it is concluded that this project has certain restrictions, but its operation and efficiency have

outperformedthe limitations in a broad way.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Affaf Khan conducted the research, Ali Iqrash analyzed the data, Arqam Nawaz Babar wrote the paper and Akbar Ali supervised us with his scholarly skills. All authors approved of the final version.

References

[1] A. Kumar, A. Nair and S. Arumugam, "Text DependentVoice Recognition System using MFCC and VQ for Security Applications," presented at International Conference on Electronics, Communication and Aerospace Technology, 2017.

[2] S. M. Hatture and S. S. Wali, "MFCC Based Text- Dependent Speaker Identification Using BPNN," *International Journal of Signal Processing Systems*, vol.3, no. 1, June 2015.

[3] S. K. Gaikwad, B. W. Gawali, and P. Yannawar, "A review on speech recognition technique," *International Journal of Computer Applications*, vol. 10, no. 3, November 2010.

[4] I.-C. Yoo and D. Yook, *Robust Voice Activity Detection Using the Spectral Peaks of Vowel Sounds*.

[5] V. Tiwari, "MFCC and its applications in speaker recognition," *International Journal on Emerging Technologies*, pp. 19–22, 2010.

[6] P. M. Chauhan and N. P. Desai, *Mel Frequency Cepstral Coefficients (MFCC) Based Speaker Identification in Noisy Environment Using Wiener Filter*.

[7] A. Srinivasan, "Speaker identification and verification using vector quantization and mel frequency cepstral coefficients," *Research Journal of Applied Sciences , Engineering and Technology*, 2012.

[8] A. K. Choudhary and J. K. Mishra, "Mel Frequency Cepstral Coefficients (MFCC) based speaker identification in noisy environment using LBG vector quantization," *International Journal of Engineering Sciences and Research Technology.*

[9] M. F. Rashid, "Biometric voice recognition in security system," *Indian Journal ofScience and Technology*, pp. 104–112, 2014.

[10] J. Yang *et al*., *Significance of Subband Features for Synthetic Speech Detection*, 2019, pp. 2160-2170.

[11] Z. Wu e*t al*., *Light Convolutional Neural Network with Feature Genuinization for Detection of Synthetic Speech Attacks*, 2020.

[12] H. Park, T. J. S. Kim, and C. Networks, *User Authentication Method via Speaker Recognition and Speech Synthesis Detection,* 2022.