A Secure Voice Signature Based Lightweight Authentication Approach for Remote Access Control

Oladayo Olufemi Olakanmi
Electrical and Electronic Engineering
University of Ibadan
Ibadan, Nigeria
e-mail: Olakanmi.oladayo@ui.edu.ng

Aminat Shodipo
Embedded Systems and Security Research Group
University of Ibadan
Ibadan, Nigeria
e-mail: Olakanmi.oladayo@ui.edu.ng

***Abstract*—Crypto-authentication schemes become unreliable whenever the private key is compromised, making them unfit for any system or network that requires high level of confidentiality. Key compromise is inevitable due to the wider space of operability of key in most of the cryptography based authentication schemes. To improve the performance of any authentication system, there is need to narrow its key operability to the key owner, that is eliminating the influence of other parties in the key generation and operability. However, this is quite difficult especially for any network that is characterised with low computation and energy, and relied on the third party for key management. In this paper, we propose the adoption of Mel-Frequency Cepstral Coefficients (MFCC) based voice signature based authentication scheme with a taint of cryptography operation. The scheme extracts the user's voice signature as the hash of the MFCC parameters of the voice and unique passcode which is used for the authentication. We used exclusive or to filter off remnant of noise from MFCC values without incurring extra hardware cost. The performance evaluation results show authentication accuracy of 87.8% at low computation cost and communication overhead.**

Keywords-MFCC, Authentication, Cryptosystem, Coding, Security, Access Control

I. INTRODUCTION

The adoption of cryptography in security system has not only enhanced the integrity of data and confidentiality but obviously contribute to the acceptability of most of the new technologies. However, cryptography based security schemes may not proffer universal security solution to most systems or networks. This is due to the restrictive processing power and memory resources, wider key generation procedure and operability that characterised these class of cryptograph based security solutions. Therefore, they may not be effective for some systems. Besides, key distribution is another issue in cryptography based security solutions, although public-key infrastructure can be used to eliminate problems involved with key distribution, however it comes with a lot of overheads. Therefore, it is very important to find ways to reduce the overheads yet not sacrificing other aspects of security.

The major loophole of cryptography security solutions is key escrow. This is a cryptographic key exchange process in which a key is held in escrow, or stored, by a third party. The problem with this, is that a lost key or compromised by its original user(s) may be used to decrypt encrypted material. Key escrow is proactive through key disclosure laws. That is, it anticipates the need for access to keys. However, it also introduces new risks like loss of keys and legal issues such as involuntary self-incrimination which are likened to security weakness.

Recently, attention has been shifted to the adoption of biometric solution to mitigate some of the security loopholes of cryptography solution. Biometric based technique may limit key generation and operability only to the users. However, some biometric based techniques have shortcomings which hinder their adoptability in access control. For example, replication is one of the major vulnerabilities of finger print based authentication systems. A few methods for fingerprint replication such as the use of grease stains on the scanner and/or latent fingerprint, the use of live finger, which is forceful amputated from the owner.

MFCC is one of the popular algorithms for extracting features of speech in voice recognition. It is common to normalise their values when adopted in speech recognition systems. Efforts had been made to improve on its algorithm such as raising the log-mel-amplitudes to a suitable power before taking the DCT so as to nullify the effect of additive noise [9]. In this paper we used exclusive or (xor) to nullify the remnant additive noise in MFCC values and generates unique voice signature of the user in the authentication phase.

DOI: 10.21307/ijanmc-2019-049

II. RELATED WORK

Voice recognition system has become a sophisticate security tool and a technology with great potential in authentication systems. It is the only biometric based recognition system that processes acoustic information contrary to other forms of biometrics, such as fingerprint, DNA, retina etc., that are image-based. Each human has their unique characteristic in speech and voice that can be captured and analyzed. Voice recognition system can be divided into two phases; identification and verification phases. Voice identification is used to decide who an unknown voice belongs to amongst a set of known speakers while voice verification accepts or rejects the identity claim of a speaker.

Voice identification can also be sub-classified into text-dependent and text-independent identification. In text-dependent identification, the individual has to utter the same keyword both in the test and training phases. Meanwhile, text-independent identification properly identifies the speaker regardless of what is being said. Voice recognition system has a lot of applications, such as authentication in remote identification and verification, mobile banking, ATM transactions, and online transactions and reservations; information security in device logon, and application and database security; Law Enforcement such as forensic investigation, and surveillance applications.

Several works had been done on voice recognition and its application to solve different access control problems. For example, Sidiqs et al. [1], proposed a speech recognition system to translate human speech to an action in machine. They used MFCC by splitting the input signal amplitudes into frames which are processed using the mel-filterbak. The results are made into a codebook which is used as an input symbol to form a model of every word. Jagan and Rameshin [3] also developed a MFCC and Dynamic Time Wrapping (DTW) based speech recognition system for feature extraction and pattern matching respectively. Drisya and Anish [2] used Bessel features as an alternative to MFCC and LPCC to develop a text-independent speaker identification system. The quasi-stationary nature of speech signal was represented by damped sinusoidal based function, and a Bessel features derived from the speech signal was used to create Gaussian mixture models for text independent speaker identification. Meanwhile, work in [4] introduced a new algorithm for extracting MFCC for speech recognition. The results showed that the new algorithm reduced the computation power and accuracy compared to the conventional algorithm.

The ability to perform postural transitions such as sittostand is an accepted metric for functional independence. The number of transitions performed in real life situations provides useful clinical information for an individual recovering from lower extremity injury or surgery. Consequent to this, Sadra and Eric [5] proposed a new inertialsensor based approach to detect transitions using wavelet transform. Their approach is robust for supervised laboratory and ambient settings. Also, authors in [6] developed a speaker recognition system using a statistical model like Gaussian mixture model (GMM) to implement a recognizer. The features extracted from the speech signal were used to build a unique identity for each authorised user. Estimation and maximization algorithms were used for finding the maximum likelihood solution for a model with latent variables.

Laryngeal diseases and vocal fold pathologies have strong impacts on the quality of the voice recognition system. In [7] a user friendly approach was proposed to discriminate between normal and abnormal voice. The feature extraction technique was applied on the voice signal in the time domain and in the frequency domain. Another work on voice recognition is speaker identification system proposed in [8]. In the work, a speaker recognition system was implemented using a combination of MFCC and Kekre's Median Code book Generation Algorithm (KMCG). The MFCC algorithm was used for feature extraction while the KMCG algorithm plays important role in code book generation and feature matching.

However, most of these works were directed to voice recognition systems. Our work is directed to how the voice signature can be combined with cryptography operation to evolve an efficient access control scheme with narrow operability to mitigate key compromise.

III. VOICE SIGNATURE BASED AUTHENTICATION SCHEME FOR ACCESS CONTROL

The voice signature based authentication scheme involves two phases; registration and authentication phases. The registration phase takes voice samples and pass codes of all the authorised individuals, processed it in order to reduce the overall bulk and complexity, then extracts MFCC (Mel-Frequency Cepstral Coefficients) before generating the voice signature. This phase is sub-divided into six divisions; elimination of silent frame, framing, hamming windowing, MFCC generation, and voice signature generation. Meanwhile, the verification phase regenerates the user's voice signature and compare it with all the encrypted voice signature stored in the

memory of the access point's memory using the k-means algorithm and correlation coefficient.

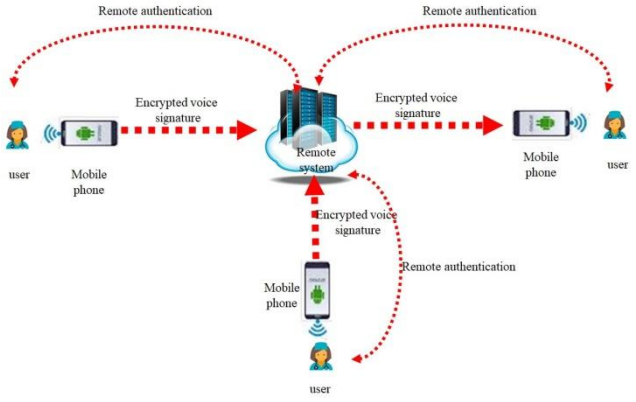

Figure 1. Model of the remote voice based access control

A. Registration Phase

In the registration phase, voice samples are collected from a number of authorised users by the remote system. These voice samples are then pre-processed and their voice signature are obtained, encrypted and stored in the memory of the access control unit of the remote system. This stage consists of six sub-stages which are described below. To generate the voice signature all the stages must be executed in that order.

1) Elimination of silent frames

It is pertinent to remove silent frames when processing speech in order to reduce the overall bulk and complexity of the speech signal. It is known that when humans are talking, it is very impossible not to have gaps or pauses in between words and sentences hence the importance of this phase. These gaps and pauses in the middle of speech increase the speech length or the number of frames to be processed, it therefore important to remove them. After proper studying of the spectrogram of speech, the amplitude for silent frames was pegged at 0.02 hence frames with amplitude less than 0.02 are removed.

2) Framing

Framing sub-stage divides the continuous speech signal into frames of B samples, with adjacent frames being separated by A where$A < B$. First frame consists of first B samples. Then, second frame begins with A samples after the first frame, and overlaps it by $B - A$ samples and so on. This continues until all the speech is accounted within one or more frames. Typically, B and A are chosen as 256 and 128 respectively. Figure 2 shows a speech signal after it has been framed.

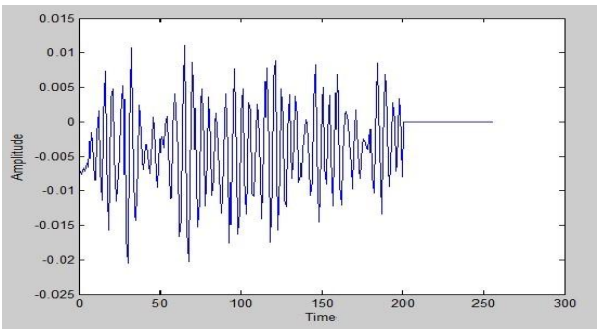

Figure 2. Speech signal after undergoing framing

3) Hamming Windowing

After framing, Hamming windowing, as shown in Figure 4, is now used to tape the voice signal to zero at the beginning and the end, thereby reducing discontinuities in the signal. This helps to focus on the information at the centre of the frame as shown in Figure 4. For example, if the window is defined as $w(n)$ and the speech signal as $x(n)$, then the resultant signal after windowing is the signal $y(n)$ defined as: $y(n) = x(n)\omega(n)$. In this work, we used a hamming window of the form:

$$\omega_n = 0.54 - 0.46\cos\frac{2\pi n}{N-1} \quad (1)$$

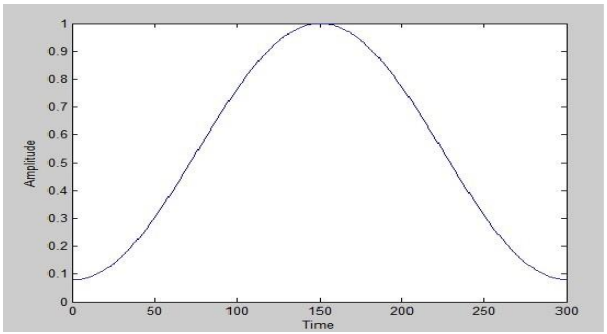

Figure 3. Hamming window before applying it on speech signal

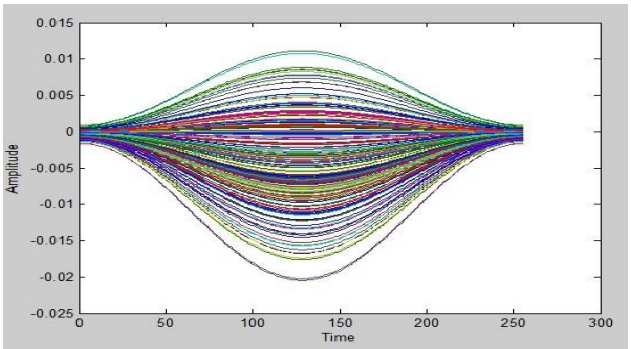

Figure 4. Speech signal after applying the Hamming Window

4) Fast Fourier Transform(FFT)

FFT is then used to transform the speech signal from time domain to frequency domain in order to easily obtain the MFCC. The FFT is a fast algorithm to implement the Discrete Fourier Transform (DFT), which is defined on the set of N samples $x[n]$. $x(k)$ gives the fast Fourier transform of frame $x(1:n)$. In general $x(k)$ are complex numbers and we only consider their absolute values (magnitudes) because considering the phase produces very skewed results as shown in Figure 5 and 6.

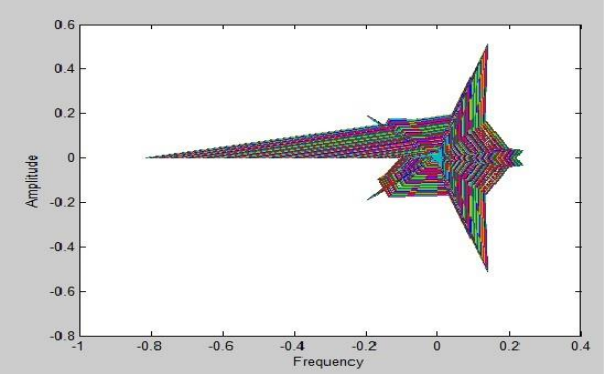

Figure 5. Audio signal with magnitude and phase

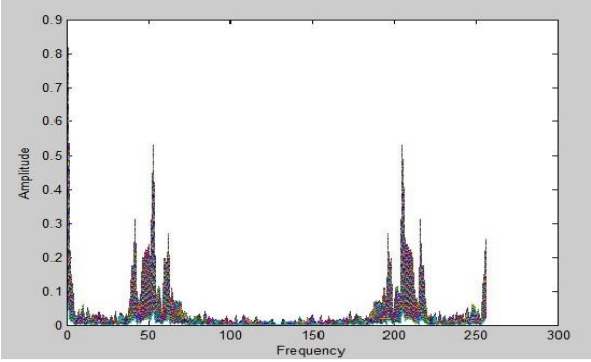

Figure 6. Audio signal considering the magnitude only

5) Mel-Frequency Cepstral Coefficient

Out of all the feature extraction algorithms that are available, MFCC is commonly used because of the way it closely mimics the natural human auditory system. The mel-frequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. The best approach to simulate the subjective spectrum is to use a filter bank, spaced uniformly on the spectral properties of the signal for the given frame analysis. The Mel spectrum coefficients are converted back to time domain using the Discrete Cosine Transform (DCT). The MFCC is calculated as:

$$\sum_1^k \cos[n(k-0.5)]\frac{\pi}{k};\ 0 \leqslant n \leqslant k-1 \qquad (2)$$

The first component is excluded from the DCT since it represents the mean value of the input signal and hence carries little speaker specific information [3].

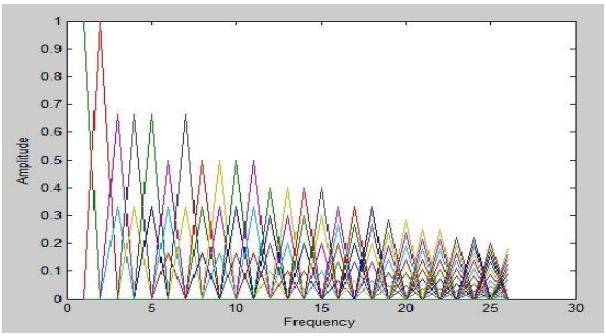

Figure 7. Mel-filter bank

6) Voice Signature generation

After obtaining the MFCC values, the access control unit then selects the maximum value $\vartheta 1$ and minimum value $\vartheta 2$ as the MFCC of the user, and calculate the user's voice signature as the encryption of hash of the MFCC and passcode of the user. That is, voice signature$\delta$is generated as:

$$\delta = H((\vartheta 1 \oplus \vartheta 2)||\text{passcode}) \oplus H(\text{passcode}) \quad (3)$$

The$\oplus$operator acts as a noise filter since the two MFCC values $\vartheta 1$ and $\vartheta 2$ includes the same additive noise and xor operator is mutually exclusive.

7) Authentication Phase

To access the system, the scheme requests for the user's voice signal through the mobile device in order to re-generate an access voice signature. It then compares the re-generated access voice signature with all the voice signatures in the memory in order to authenticate the user.

IV. PERFORMANCE EVALUATION

A. Experimental setup

One hundred and fifty two tests using 5 different users of mixed sex were used to test the efficiency of the voice based authentication scheme. Voice signatures of the five users (2 females and 3 males) saying the same sentence were extracted.

TABLE I. PERFORMANCE EVALUATION OF VOICE SIGNATURE BASED AUTHENTICATION APPROACH

Trial	No. of samples	No.of False rejection	No. of False acceptance	% Accuracy
User 1	30	3	1	87
User 2	30	3	2	83
User 3	30	2	0	93
User 4	30	1	1	93
User 5	32	4	2	83

Each person voice signature was matched with the rest of the voices signatures in the database. The accuracy of systems was determined in terms of false acceptances and false rejections. A false acceptance occurs when the system grants access to an unauthorised user while a false rejection occurs when the system denies the unauthorized user is granted access. The scheme was simulated on a simulation platform consisting of a mobile device (Samsung Galaxy S5 with a Quad-core 2.45 GHz processor, 2GB RAM, and Google Android 4.4.2 operating system, a PC with Intel(R) Core(TM) i5-7200U CPU @ 2.50 GHz processor as the remote access control system. We determine the computation and communication costs of the scheme.

B. Result and Discussion

The results of the performance of the scheme in terms of false acceptance and rejection ratios are shown in Table 1. It shows that the system is accurate with accuracy of 87.8%.

Also, computation cost, in terms of the execution time by the scheme for $N = 200$ points FFT, is obtained as shown in Figure 8. This shows that the computation cost increases as the number of users increases, and indicates that the scheme has low computation cost and can be easily adopted by any system that is characterised as computation and energy constraint system. Also, Figure 9 shows the energy consumption of the scheme in terms of number of cycles required. This also indicates that the proposed scheme is energy-aware since energy consumed by a processor is approximately proportional to number of cycles or frequencyf, and to the square of the processor voltage V [14].

Meanwhile, the communication overhead incurred for every authentication is 256 bits. This indicates that the scheme requires low bandwidth and there will never be congestion irrespective of the bandwidth of the communication channel.

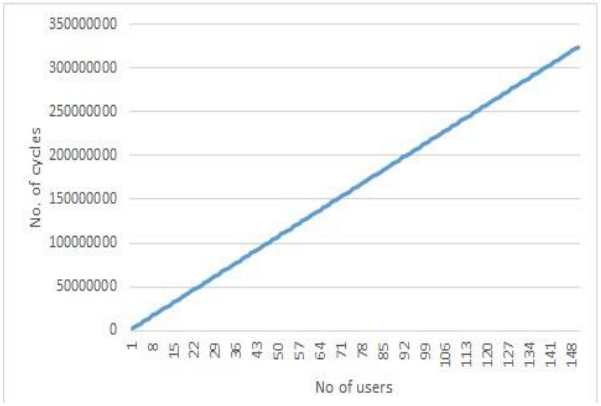

Figure 8. Computation cost (ms)

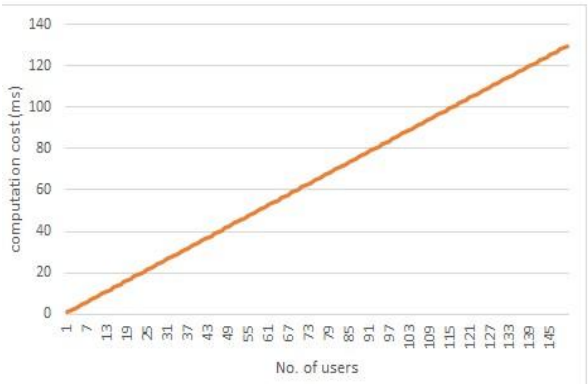

Figure 9. Energy cost in terms of cycles

V. CONCLUSION

In this work, we demonstrated how voice signature can be used to developed access control scheme for remote system. We solved the effect of the additive noise on MFCC using $\oplus$ to eliminate the congruent additive noise embedded in the maximum MFCC and minimum MFCC values. The MFCC, hash function and conventional pass-code are used generate voice signature from user's voice signal. This is used to solve the problem of wider operability of key in cryptography based

REFERENCES

[1] Muslim Sidiq, Tjokorda Agung Budi W, Siti Saadah (2015). Design and Implementation of Voice Command Using MFCC and HMMs Method. 3rd International Conference on Information and Communication Technology (ICoICT).

[2] Drisya Vasudev, Anish Babu (2014). Speaker identification using FBCC in Malayalam language. International Conference on Advances in Computing, Communications and Informatics (ICACCI).

[3] Jagan Mohan and Ramesh Babu(2014). Speech recognition using MFCC and DTW. 1st Int. Conference on Advances in Electrical Engineering, VIT, Vellore, India

[4] Wei Han, Cheong-Fat Chan, Chiu-Sing Choy, Kong-Pang Pun (2006). An efficient MFCC extraction method in speech recognition. 2006 ISCAS, Proceedings of IEEE International Symposium on Circuits and Systems.

[5] Sadra Hemmati, Eric Wade (2016). Detecting postural transitions: A robust wavelet-based approach. Proceeding of IEEE 38th Annual International Conference of the Engineering in Medicine and Biology Society (EMBC), pp. 3704-3707.

[6] S. G. Bagul, R.K.Shastri (2013). Text Independent Speaker Recognition System Using GMM. International Conference on Human Computer Interactions (ICHCI), pp 1 - 5.

[7] Manal Abdel Wahed (2014). Computer aided recognition of pathological voice. 31st National Radio Science Conference (NRSC), pp. 349 – 354.

[8] H B Kekre, V A Bharadi, A R Sawant, Onkar Kadam, Pushkar Lanke, Rohit Lodhiya (2012). Speaker recognition using Vector Quantization by MFCC and KMCG clustering algorithm. International Conference on Communication, Information and Computing Technology (ICCICT).

[9] Tyagi and C. Wellekens (2005). On desensitizing the Mel-Cepstrum to spurious spectral components for robust speech recognition. In proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 1, pp. 529-532.

[10] Saqui Z., Salam N., Nair N., Pandey N. (2011) Voiceprint recognition system for remote authentication survey. International Journal of Hybrid Information Technology, Vol. 4, No.2.

[11] Sunil A., Shruti A., Rama C. (2010) Prosodic Feature Based Text Dependent Speaker Recognition Using Machine Learning Algorithms. International Journal of Engineering Science and Technology, Vol. 2 No.10, Pp. 5150-5157.

[12] Kirti A., and Minakshee P., (2013). Speech and speaker identification for password verification system. International Journal of Advanced Research in Electrical,Electronic and Instrumentation, Vol. 2, Issue 6.

[13] Parrul, R., Dubey, B., (2012). Automatic speaker recognition system. International Journal of Advanced Computer Research, Vol. 2, No.4.

[14] Wikipedia. (2003). CPU power dissipation. (http://en.wikipedia.org/wiki/CPU-power-dissipation)