IDENTITY AUTHENTICATION USING VOICE BIOMETRICS TECHNIQUE

U A Kamalu[1], A. Raji[2], V.I. Nnebedum[3]

[1]*University of Port-Harcourt, Nigeria,* ***ugokamax@yahoo.com***
[2]*University of Port-Harcourt, Nigeria,* ***rajoys@yahoo.com***
[3]*University of Port-Harcourt, Nigeria,* ***nnebedum@yahoo.com***

Abstract

Identification of people using name, appearances, badges, tags and register may be effective may be in a small organization. However, as the size of the organization or society increases, these simple ways of identifying individual become ineffective. Therefore, it may be necessary to employ additional and more sophisticated means of authenticating the identity of people as the population increases. Voice Biometrics is a method by which individuals can be uniquely identified by evaluating one or more distinguishing biological traits associated with the voice of such individuals. In this paper, an unconstrained text-independent recognition system using the Gaussian Mixture Model was applied to match recorded voice to stored voice for the purpose of identification of individual. Recorded voices were processed and stored in the enrollment phase while probing voices were used for comparison in the verification/recognition phase of the system.

Keywords*: Model, Biometric, verification, enrollment, database, authentication, matching, identity.*

--***--

INTRODUCTION

As individual varies in sizes and shapes, so the voice varies also. There are many biometrics methods used for identification purpose. Fingerprint biometrics is perhaps the most commonly deployed biometrics for authenticating the identity of people but it has its demerits. Fingerprint is intrusive and requires the cooperation of the concerned individual to be effective. More so,injury and wounds to the finger can negatively affect fingerprint and render the system useless. In contrast, voice is not subject to injury, it is non-intrusive and in some cases may not require the cooperation of individual during verification and recognition. The voice of a speaker can be recognized and authenticated in several wayswith Speaker Recognition Systems (SRS) but the most common methods used are the Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM). Text-dependent voice recognition system uses the HMM while text-independent applicationsemploys the GMM. HMM uses fixed or prompt sentence of few words for testing authentication and is slightly intrusive and could be frustrating to some people. GMM do not use fixed or prompt sentence but allow any utterances for the identification of speaker because it uses the behavioral or physiological characteristics of individual (Yiyin, 2008). In this paper, GMM was used for the obvious reason of its flexibility and more friendly appeal.Biometrics techniques use physiological and/or behavioral characteristics for authentication and identification of people rather than the use of token. This technique uses personal characteristics that cannot be easily duplicated (Tracy et al, 2006). The biometric identifiers of various individuals are permanent features unique to them and as such more reliable than token or knowledge based authentication methods. The categories of biometrics used for authentication and identification of people are signature verification, hand geometry, Fingerprint identification, DNA identification, facial recognition, and voice recognition (Adewole et al, 2011).

The voice of a person is complex information-bearing signals that depend on physical and behavioral characteristics. In order to recognize a person based on voice, the complexity must be reduced while keeping sufficient information in the extracted voice feature vector. People’s voices can be distinguished based on the frequency of the speech, pitch and tone as well as the amplitude. The frequency of speech is measured based on the speed at which the individual pronounce a word or make a speech. The amplitude is based on the pitch or loudness of the spoken word. Any of the above mentioned characteristics can be used to distinguish voices of people from one another. The voices at different mood of an individual can vary andas such, the state of mood of an individual when the voice is recorded must be taken into consideration when comparing the voice to identify a person. To record the voice of a person, microphone and recorders can beused but recording can also bedone with software applications. The availability and characteristics such as sensitivity, signal-to-noise ratio and size were considered in the choice the recording unit.

METHODOLOGY

The system involves the enrollment and verification phase. During the enrollment, voices of individuals in the system were recorded and the features were extracted and stored as template in the database for each individual that enrolled. For the verification phase, the recorded probing voice sample collected was compared with those in the database to determine if a match is found to authenticate the identity of the new voice. Java Luna Eclipse software application tool was used as the integrated development environment (IDE) while SQL Server 2000 was used in the design of the database of the system. The system makes use of the voice of people to authenticate the identity of individuals that enrolled in the system. The verification and authentication of people was achieved from the one-to-one voice comparison of various individuals who enrolled and that of the assumed impostors in the system. Therefore, the entire system comprises of the extraction of features vector from the recorded voice, the modeling of the features and the comparison of the voices collected for verification to determine the identity of an individual in the system.

Voice Signal Recording

The voices of individuals were recorded both during enrollment and verification phase using the java swing sound recording tool since the IDE for the system is Java Luna eclipse. The recorded voices at the enrollment were then exported to the “house audio” folder of the system application for processing while the recorded voice during verification face were exported to the “match audio” folder of the system. The recording tool is shown in figure 1 below while a sample of one of the recorded voice is as shown in figure 2.

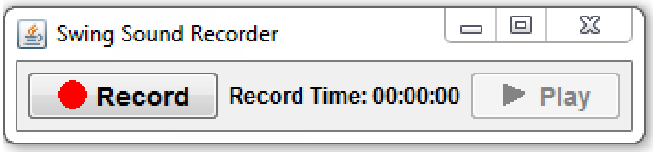


Figure 1: The java recording tool

Voice Features

There are different categories of voice features that can be used for physical interpretation of voice in a speaker recognition system. The common ones are Short-term spectral features, Spectra-temporal features, Voice source features, High level features and Prosodic features(Kinnunen and Li, 2009).Features extraction is a vital step in the process of any speaker recognition system. The acquired data (voice signal) from individual is processed in this section of the system either during the enrollment or during identification/verification. A set of feature vectors or parameters from the speech signal representing some speaker-specific information, which results from complex transformations occurring at different levels of the speech production are extracted and stored. (Mathur et al, 2013). It is good to adopt audio parameters which are related to voice characteristics that can distinguish speakers. Voice recognition depends upon both low level and high level information obtained from a person’s speech. The high level information includes values like dialect, accent, the talking style, the subject manner of context, phonetics, prosodic and lexical information (Shriberg, 2007). Low-level features comprise of data such as fundamental frequency, formant frequency, pitch, intensity, rhythm, tone, spectral magnitude and bandwidths of an individual’s voice. An ideal feature would possess the following characteristics:

- Robustness against noise and distortion
- Easy to measure and determine from the speech signal
- Difficult to mimic
- Naturally and frequently occurring in the speech
- Lower intra-speaker variability and high inter-speaker variability
- Less affected by speaker’s health or long term variations in voice

Voice Signal Processing and Feature Extraction

Thevoice signal that will appeal to the ear of man can range from 100Hz to about 4000Hz, using 6 bits per sample and based on Nyquist sampling criterion, the voice were sampled at least twice the maximum frequency (8000 samples per seconds). The bit rate was obtained using this formula:
Bit rate = Sampling rate x Number of bit per sample
Bit rate = 8000 x 6 = 48000 bits per seconds.
The recorded raw voice is in waveform format which can be represented as shown in figure 2. It comprises of several components of complex signal from the voice.

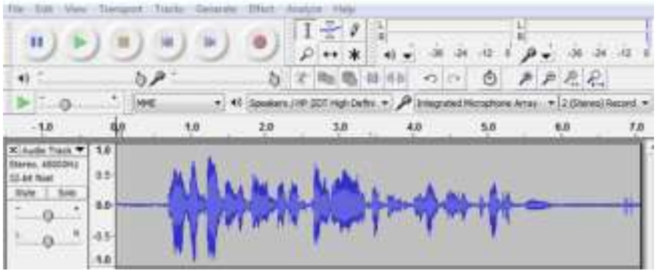


Figure 2: Sample of recorded voice

In the simplest form, a typical raw voice signal in waveform format is depicted in figure 3 with the amplitude of the voice signal shown on the vertical axis while horizontal axis represents the time.

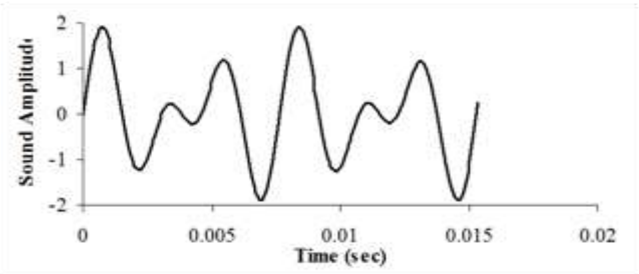


Figure 3: A Representation of simple waveform of audio signal X(t)

If the above is therecorded voice signal X(t) andsampled with Nyquist criterion, then figure 4 can be used to represent the sampled signal $X_s(t)$ where T_s is the duration of sampling interval.

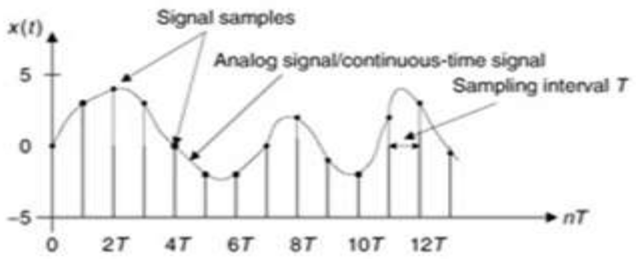


Figure 4: Simple Representation of a Sampled signal $X_s(t)$.

The signal depicted in figure 3 is sinusoidal and could be represented mathematically by the expression [Weiji, 2012].

$$\mathbf{X(t) = U(t) = a\cos(\omega\, t) + b\sin(\omega\, t)} \qquad (1)$$

The above equation can be transformed into complex form with Euler's formula in which case,

$$\mathbf{U(t) = ae^{j\omega\, t} + be^{-j\omega\, t}} \qquad (2)$$

The expression in equation (2) represent a simple single signal and for a complex ones like the voice signal comprising of several components of these simple signals, the above expression can be modified to cater for the summation of the various components of the simple signals and represented by equation (3) below.

$$\mathbf{U(t) = \sum (a_k e^{j\omega k\, t} + b_k e^{-j\omega k\, t})} \qquad (3)$$

For periodic signals, the fundamental frequency is given by Ω = 2Π/T such that equation (3) can be represented by the general equation below.

$$\mathbf{U(t) = \sum U_n e^{jn\Omega t}} \qquad (4)$$

Fourier analysis was used to get the value of U_n coefficients for each of the various simple waves that make up the voice signal. The decomposition of the signal into its fundamental sine wave using the Fourier transformation makes it possible to extract features from the voice.The mel-frequency cepstral coefficients (MFCCs) are common features in audio processing. It is a popular feature that can be extracted and used in speech processing and recognition. The MFCCs features are in frequency domain and more accurate compare to other features that are in time domain. It represents the real cepstral of a windowed of a short-time signal derived from the Fast Fourier Transform (FFT) of that signal. It extracts parameters from the voice signal which are similar to those used by human ear for hearing speech. MFCCs are computed with the aid of a psycho-acoustically motivated filterbank, followed by logarithmic compression and discrete cosine transforms (DCT). Denoting the outputs of an M-channel filterbank as Y(m), for m = 1 ……… M, the MFCCs are obtained as shown in the expression below (Kinnunen and Li, 2009).

$$C_n = \sum_{m=1}^{M} [\mathrm{Log}\, Y(m)] \cos\left[\frac{\pi n}{M}\left(m - \frac{1}{2}\right)\right]$$

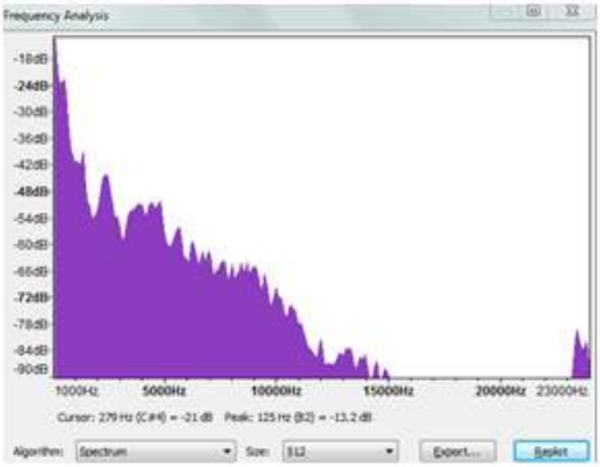
Where n is the index of the cepstral coefficient.
The Mel-scale is approximately linear up to 1 KHz and the relationship between frequency (f) in Hz of the signal and the Mel-scale is represented with the expression below [Yiyin, 2008].

$$\mathrm{Mel-Scale\ Frequency} = 2595\ \mathrm{Log}\left(1 + \frac{f(\mathrm{Hz})}{700}\right)$$

The algorithm used for extracting the spectra cepstrum features in each frame x[n] of the recorded audio signal is as indicated thus;

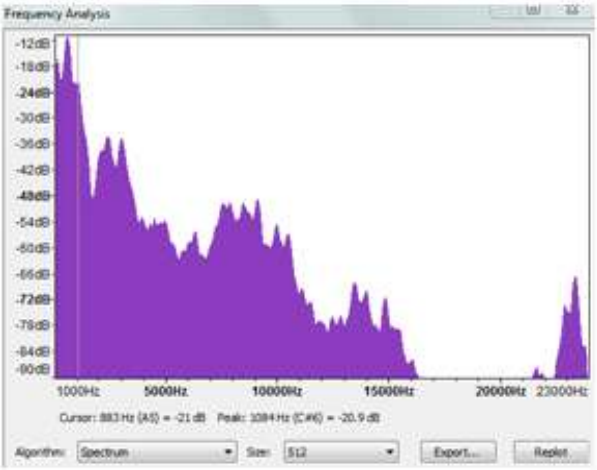
Step 1: Start
Step 2: Lunch SRS
Step 3: Perform DFT on x[n] (This computes the spectrum of recorded voice).
Step 4: Compute the Logarithm of the result obtained from step 3.
Step5: Perform the inverted DFT operation on the result obtained from step4.
Step 6: Store the result of Step 5 in the DB
Step 7: Stop

The result of the inverted DFT on the spectrum extracts two main frequencies which are the spectral envelope (low frequency) and the excitation signals (high frequency) which are stored and can be used for recognition during matching. Figure 5shows variation in frequency spectrum of four different people speaking the same sentence.



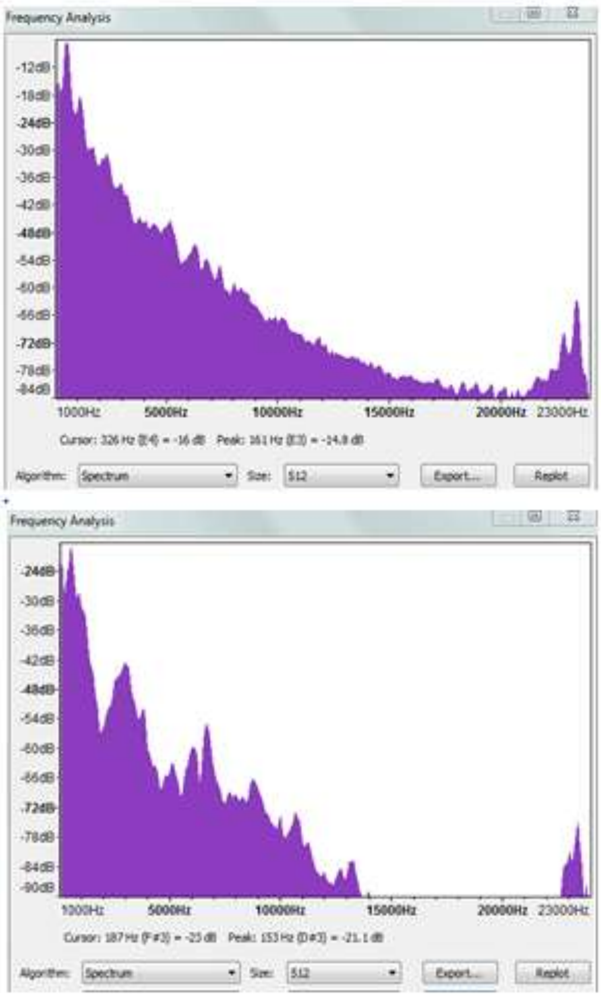
Cursor: 279 Hz (C#4) = -21 dB Peak: 125 Hz (B2) = -13.2 dB

Algorithm: Spectrum Size: 512



Cursor: 883 Hz (A5) = -21 dB Peak: 1084 Hz (C#6) = -20.9 dB

Algorithm: Spectrum Size: 512



Cursor: 326 Hz (E4) = -16 dB Peak: 161 Hz (E3) = -14.8 dB

Algorithm: Spectrum Size: 512

Cursor: 187 Hz (F#3) = -25 dB Peak: 153 Hz (D#3) = -21.1 dB

Algorithm: Spectrum Size: 512

Figure 5: Variation of Frequency Spectrum for four individuals speaking same thing.

Modeling and Matching

The Gaussian Mixture Models (GMMs) method uses the speaker feature vectors directly for recognition purpose in a text-independent system. Feature vectors of individuals from such system can be viewed as the coefficients extracted from the mathematical transformation of the voice signals using the DFT or FFT. The whole feature vectors would then represent a single point in the n-dimensional room for a person. These points represent speaker characteristics which can be used to compare speakers to each other by computing the distance between the points. When the result of the distance calculation between two points is small, it means the probability or degree of similarity of the speakers are high. GMMs give a statistical chance of belonging to a speaker to be identified. It is a weighted sum of M-component Gaussian probability densities functions [Reynolds, 2005]. The idea behind GMM is that a mixture of probability functions can be assumed as the combination of a number of separate probability functions as seen in Figure 6.

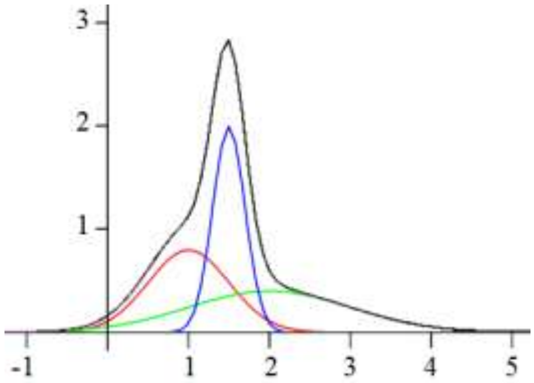


Figure 6: A simple example GMM in onedimension.

The black line represents the GMM, and the others represent its components.

Using GMM as the classification likelihood function has some merits which includes computationally inexpensive, the statistical model is well-understood and for text-independent tasks, it is insensitive to the temporal aspects of the speech and model only the underlying distribution of acoustic observations from a speaker.

The condition for testing if a feature vector belongs to a speaker voice model is to compute a similarity score when comparison is done. The similarity score is calculated based on the observed feature vector, the background model and the speaker voice model. The probability that a speaker model belongs to the observed feature vector can be computed using the equation below [Petter, 2012].

$$\mathrm{Log}[P(m /O)]=\mathrm{Log}[P(O / m)]-\mathrm{Log}[P(O /m_w)]$$

Where O is the observed feature vector, m is a speaker model and m_w is the background model. The score computed from the above equation is referred to as Log-Likelihood Ratio (LLR) score which solely dependent on comparisons between the speaker model and the background model. Any LLR score that is less than 0 indicates a rejection and a score greater than 0 indicates a possible match. A person whose voice is been verified will be rejected or accepted based on a fixed threshold value set for the value of Log [P(m /O)]. The set value should be surpassed to be convinced that a speaker model belongs to the observed feature vector. For this speaker verification system, the threshold for acceptance was determined through the mathematical computation and decision of the software tool application. The algorithm for recognizing a probe voice as part of the voices already stored in the DB or rejecting the voice is as shown below.

Step 1: Start

Step 2: Lunch SRS

Step 3: Perform DFT on the probe voice.

Step 4: Compute the Logarithm of the result obtained from step 3.

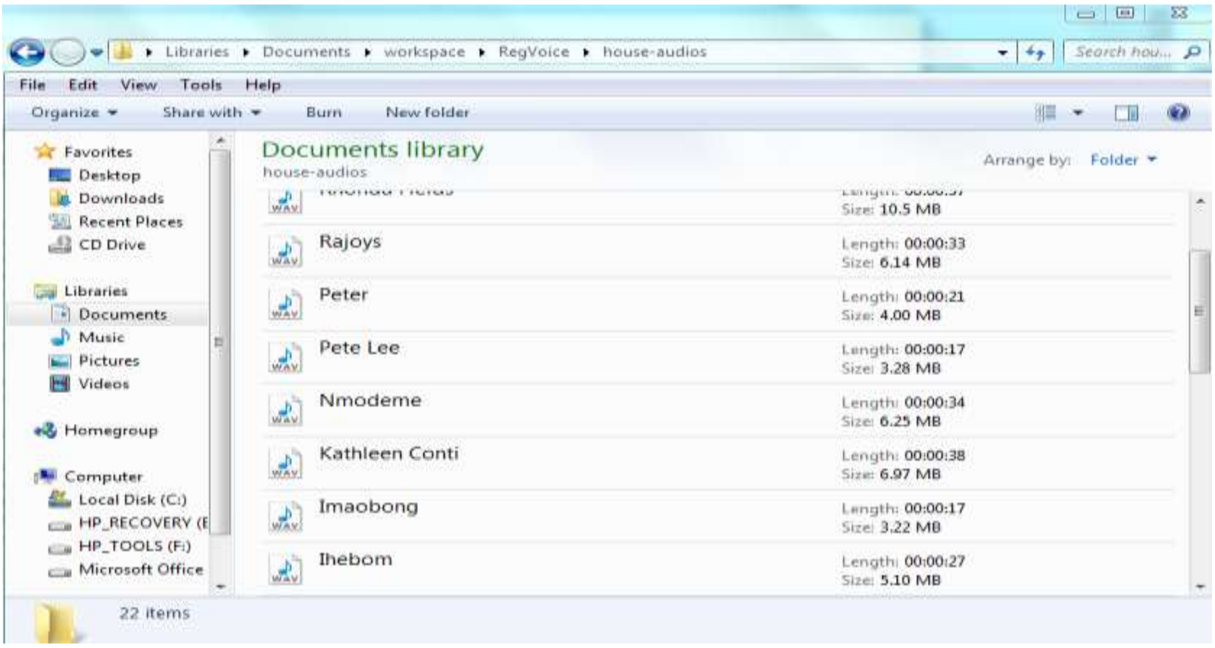
Step 5: Perform the inverted DFT operation on the result obtained from step 4.

Step 6: Compute LLR for the result of step 5.

```
    Step 7: Compare the result of step 6 with values of voices in DB
    Step 8: If the result of step 7≥ 0.95
        Then
            a. Identify the nearest stored voice
        Else
            b. Reject.
    Step 9: Stop.
```

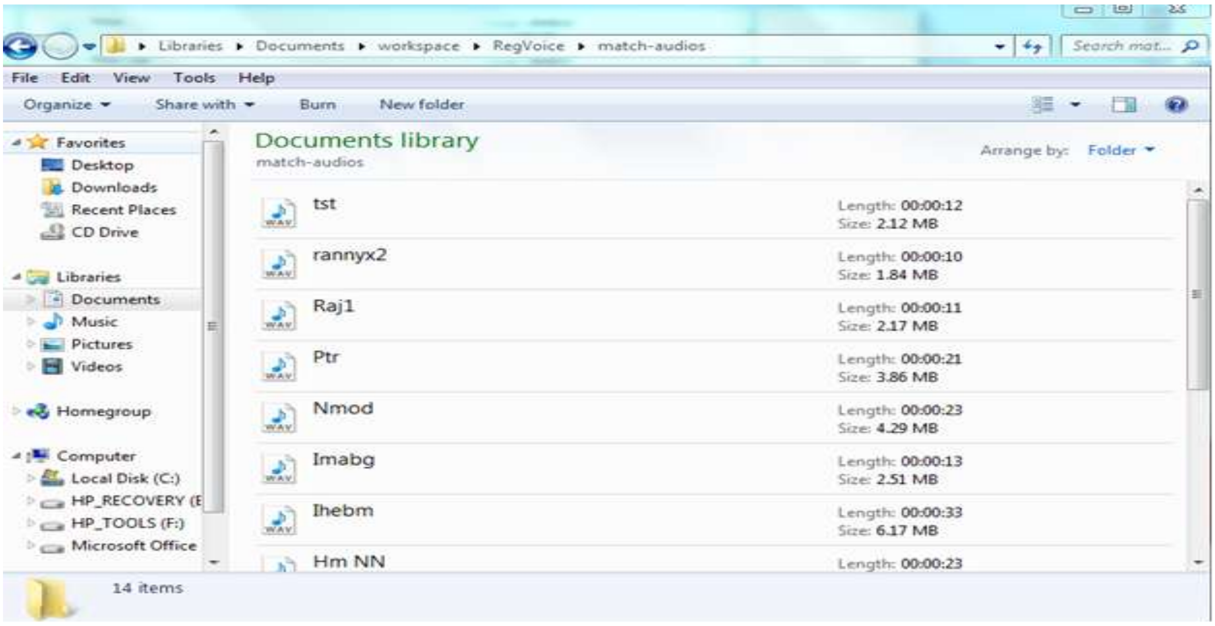
RESULTS AND DISCUSSION

The voice of 22 different peoplecomprises of male and female individuals were recorded at the enrollment phase of the system and some of the recorded voices are shown in the house-audio folder in figure 7 below. However, only 14 probing voices of people who were randomly chosen from the initial 22 people who participated during enrollment were later used for matching purpose in the recognition phase. The voices of these 14 individuals some of which are shown in figure 8 were recorded at the verification stage and compared with the earlier recorded 22 voices to determine if there is a match to identify the person.The voices of various individuals in the system were stored with an identification name both at the enrollment and verification phase for the purpose of recognition whenever there is a match after comparison. Since the work is on text-independent system that uses GMM, the individuals that participated were free to alter any sentence of their choice or sing for considerable time both during enrollment and verification phase.



Size: 10.5 MB
Rajoys Length: 00:00:33 Size: 6.14 MB
Peter Length: 00:00:21 Size: 4.00 MB
Pete Lee Length: 00:00:17 Size: 3.28 MB
Nmodeme Length: 00:00:34 Size: 6.25 MB
Kathleen Conti Length: 00:00:38 Size: 6.97 MB
Imaobong Length: 00:00:17 Size: 3.22 MB
Ihebom Length: 00:00:27 Size: 5.10 MB
22 items

Figure 7: Sample of Some Recorded voices during enrollment.



tst Length: 00:00:12 Size: 2.12 MB
rannyx2 Length: 00:00:10 Size: 1.84 MB
Raj1 Length: 00:00:11 Size: 2.17 MB
Ptr Length: 00:00:21 Size: 3.86 MB
Nmod Length: 00:00:23 Size: 4.29 MB
Imabg Length: 00:00:13 Size: 2.51 MB
Ihebm Length: 00:00:33 Size: 6.17 MB
Hm NN Length: 00:00:23
14 items

Figure 8: Sample of probing voices recorded for matching purpose

The voice matching code do a sequential one to one comparison based on frequency of the voices recorded and stored during enrollment and the probing voice.It compares the probing voice with each of the 22 voices and returns a value and the name of the most likely voice from 22 people that enrolled. The value is the likelihood ratio of how closed the probing voice is to any of the enrolled voices stored in the database. Table 1 shows the representation of the probability value of the likelihood rate that was return for the matching of eight authentic individuals and that of two other assumed impostors for each of the authentic person. It shows value obtained for eight individuals whose voices were recorded twice, one for enrollment and the second for matching during recognition. The table also shows the value obtained when 2 other people voices are used as impostors on the eight people. The name of real owner of the voice and the abbreviated name for their matching for each person are shown. Similarly, the names of the assumed impostor are also shown.

Table 1: Probability value obtained for matching of genuine voice and Impostor voices.

S/N	Names of Genuine voice	Genuine value	Impostor1 name	Impostor1 value	Impostor2 name	Impostor2 value
1	Imabong and Imbg	1	Rajoys	0.73	Ihebom	0.69
2	Peter and Ptr	1	Ayo Raji	0.67	Susan	0.27
3	Akerele and Akere	0.92	Ekpetiama	0.95	Ihebom	0.92
4	Ekpetiama&Ekpe	0.98	Nmodeme	0.45	Rhonda	0.36
5	Nmodeme&Nmod	1	Akerele	0.62	Pete lee	0.26
6	Ihebom and Ihbm	0.99	Akpada	0.69	Peter	0.42
7	Akpada and Akpa	0.99	Nmodeme	0.72	Ayo Raji	0.53
8	H-MAN and HM	1	Akerele	0.45	kathleen	0.22

The second column in the above table indicates the names and short form of the names of the genuine person whose voice was recorded. The full name was used at the recording of the voice during the enrolment phase while the short abbreviated name was used to record the voice of the same person during the recognition phase. The result of the matching of two genuine voices from the same person recorded at different time is reflected directly in beside each name in the third column. The fourth and the sixth column contain the names of the various assumed impostors whose voices are to be compared with the genuine voice of a person during recognition. The results of the matching between the assumed impostor's voices and the genuine voice are also reflected beside each name of the impostors in column 5 and column 7 respectively. From the table, the value for matching of the voices of a genuine person is as high as 1 for some people but for others, it varies from 0.92 to 0.99. Hence, the criteria for accepting a voice to be genuine in this work were put at 0.97 and above. The impostors also score far less than this range of acceptance from the table. The only exception to this is the case of the impostors in row 3 in the table where an impostor value is even higher than the value of the genuine owner of the voice while the second impostor has exactly the same value of 0.92 with the genuine owner of the voice. This was taken as error initially when this was noticed during testing, but the process was repeated again and similar abnormal results were observed for the genuine owner of voice and the impostors in row 3. This is a case of false acceptance and false rejection in the algorithm for the matching of the voices.

CONCLUSION

Voice signal from people like other biometrics method such as fingerprint can be used for authenticating the identity of a person. The identity authentication of people using voice has some merits. It is easy to use and less intrusive into personal life especially in forensic applications. It is also not affected by physical injury and can be deploy for remote authentication of people with or without their knowledge. These advantages will make the technology to continue to gain more popularity in years to come and researches in these areas seems promising too.

Nevertheless, this technology of authenticating identity of people from their voices has some downside. These include the variation in people voice when they are sick, angry or change in mood. All these conditions must be properly accounted for both during enrollment and verification phases to minimize the error rate in the system. In this paper, voice signal was presented as another biometrics parameter that can be measured to serves as the basis for authenticating the identity of an individual. The result presented for 8 people above was fairly good but cannot be consider of being perfect authentication means. It may work for forensic application by narrowing down the number of suspects but there is need to take other measure to ensure a reliable authentication especially when security matter of high importance and huge financial transaction is of interest. Therefore, the systems cannot be seen as an ultimate security tool or the perfect solution but rather as attempted approach to security. Majority of programmes and initiatives going on in the current Information Technology driven world involve collection of biometrics data from the citizen. It can be leveraged on to have a robust biometric database for the society using voice of individual in the world that is now a global village.

REFERENCES

[1] Adewole, KS.,Abdulsalam, SO., Jimoh, RG (2011). Application of Voice Biometrics as an Ecological and Inexpensive Method of Authentication.

[2] Kinnunen T, Li H (2009) An overview of text-independent speaker recognition: from features to supervectors. Speech communication 52: 12-40.

[3] Mathur S, Choudhary SK, Vyas JM (2013) Speaker Recognition System and its Forensic Implications.

[4] Petter, L. (2012).VoiceR, a System for Speaker Recognition using Gaussian Mixture Models. Degree Project in Computer Technology and Software

Development, First Level Royal Institute of Technology School of Technology and Health 136 40 Handen, Sweden.

[5] Reynolds, DA., (2005). Gaussian Mixture Models.MIT Lincoln Laboratory, Lexington, MA USA.

[6] Shriberg E (2007) Higher-level features in speaker recognition. Speaker classification I, Springer Berlin Heidelberg 4343:241-259.

[7] Weiji, W. 2012, Introduction to Digital Signal and System Analysis. Ventus Publishing Aps.

[8] Yiyin, Z. (2008). “Speaker Identification: E6820 Spring Final Project Report.A Project Research work in Columbia University.