Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 564904, 21 pages
http://dx.doi.org/10.1155/2015/564904

Research Article

Experimentation Using Short-Term Spectral Features for Secure Mobile Internet Voting Authentication

Surendra Thakur, Emmanuel Adetiba, Oludayo O. Olugbara, and Richard Millham

ICT and Society Research Group, Durban University of Technology, P.O. Box 1334, Durban 4000, South Africa

Correspondence should be addressed to Oludayo O. Olugbara; oludayoo@dut.ac.za

Received 1 April 2015; Revised 1 July 2015; Accepted 2 July 2015

Academic Editor: Ivanka Stamova

We propose a secure mobile Internet voting architecture based on the Sensus reference architecture and report the experiments carried out using short-term spectral features for realizing the voice biometric based authentication module of the architecture being proposed. The short-term spectral features investigated are Mel-Frequency Cepstral Coefficients (MFCCs), Mel-Frequency Discrete Wavelet Coefficients (MFDWC), Linear Predictive Cepstral Coefficients (LPCC), and Spectral Histogram of Oriented Gradients (SHOGs). The MFCC, MFDWC, and LPCC usually have higher dimensions that oftentimes lead to high computational complexity of the pattern matching algorithms in automatic speaker recognition systems. In this study, higher dimensions of each of the short-term features were reduced to an 81-element feature vector per Speaker using Histogram of Oriented Gradients (HOG) algorithm while neural network ensemble was utilized as the pattern matching algorithm. Out of the four short-term spectral features investigated, the LPCC-HOG gave the best statistical results with R statistic of 0.9127 and mean square error of 0.0407. These compact LPCC-HOG features are highly promising for implementing the authentication module of the secure mobile Internet voting architecture we are proposing in this paper.

1. Introduction

Election is the process by which voters in a political entity elect leaders among competing candidates by casting of votes either on a ballot paper or electronically in order to actualize desired changes in their society. The electoral processes are very vital because they empower citizens to have an influence on the future policies of their governments and consequently on their own futures. Electoral processes involve voter registration, voter validation, voting, tallying, transmission, tabulation, and result publication [1].

There have been series of evolution in voting over years. In 4BC voting in Athens, for instance, the use of vocal votes known as *viva voce* was prevalent and some later evidence of democratic voting practices in this era encompassed the *showing of hands* by the electorates to indicate their choice of candidates [2, 3]. Other election practices by the Athenian Greek voters involved voting by inscribing their choices on discarded pieces of pottery called *ostraka*, which was placed in an *urn* and tabulated. During the Renaissance period, some voting practices included the use of *white balls* for acceptance and *black balls* for rejection of candidates. The balls, called *ballotta*, are the origin of the term *ballot*, which is now an essential component of conventional paper-based elections [4, 5]. Elections in India are traceable to 920AD in which the voters wrote candidate names on the palm leaf (known as *Panai olai*) and dropped the *Panai olai* in the pot from which counting was done for each of the participating candidates. After the counting, the candidate with the highest number of votes was elected [6]. The *ballotta*, the *ostraka*, and the *Panai olai* are examples of elections where available resources were used to record votes in economical ways. These electoral scenarios also represent environmentally friendly election administration methodologies in the early eras. Although these methods were seemingly crude, they were able to achieve vote secrecy and adequate counting and auditing was made possible through vote recounting. These older voting practices were adjudged transparent until when coercion

and buying of votes started interfering regrettably with the sanctity of the process and consequently prevented voters from voting with their conscience [2–4, 7].

As populations began to increase across different countries, larger-scale elections became unwieldy to administer and with the increasing availability of paper, pen, and ink the use of paper ballot was birthed [8]. The paper-based voting process involved the votes being recorded by officials with citizen input or *viva voce*. However, it later became evident that paper ballot elections have a lot of logistical and administrative challenges such as high cost, slow tabulation, ballot misinterpretation, vote miscounts, possibility of voter coercion, and vote buying. Shamos and Yasinsac [9] are unequivocal in asserting that every form of paper ballot that has been devised can and has been manipulated with considerable ease. As the number of voters grew, faster and more accurate tabulation of votes and speed in result compilation became desirable. This led to the mechanization of voting mostly in the United States of America (USA) with the introduction of lever machines. Punch card system which employs cards and a small clipboard-sized device for recording votes were also some early approaches adopted for election mechanization. In using this system, voters punched holes in the cards in the opposite position to the choice of their candidates. Punch card systems were for a while perceived to reintroduce transparency and auditability into electoral processes [4]. The system beneficially introduced tallying speed and automatic ballot count, removed voter monitoring, mediated irregularities such as ballot stuffing, ballot interpretation, chain voting and finally prevented over voting. New challenges were however introduced, which include logistic problems, equipment failure, and training requirements. Furthermore, the events of the year 2000 regarding *Bush versus Gore* election in Florida, USA, challenged the persistent use of the punch machines for voting. Here, many deployed punch systems did not punch holes clearly on the ballot and this led to hanging portions that are now famously called hanging, dimpled, or pregnant chads [5].

The rapid advancement in electronic technologies has led to the development of electronic voting systems. An electronic voting system often abbreviated as *e-voting* is an integrated system that uses electronic components to perform electoral functions. E-voting introduced technology to electoral process so as to leverage on possible benefits such as more efficiency, transparency, auditability, speedy release of results, and ease of voting and to ultimately enhance the trust of the electorate in the management of election and referendum. The Americans and the Dutch are the forerunners in the development and deployment of computer-based e-voting terminals referred to as Digital Recording Electronic (DRE) equipment. A DRE is a programmed device that operates as a vote capturing terminal. DREs were also adopted and implemented in other political climes but with different nomenclatures. They are called Electronic Voting Machine (EVM) in India; the Brazilians refer to them as urnas while the Filipinos call them Precinct Optical Scan (PCOS) [8]. An immediate problem with the first generation DRE/EVM was that the votes were captured and a black box result was produced by the lack of a paper trail. This led to Mercuri [10] proposing a Voter Verifiable Audit Trail (VVAT), which is a printed equivalent of the computer choice for voters. EVMs with VVAT are now being implemented in some parts of India, many parts of the USA, and across the world. Besides DRE/EVM, other typologies of e-voting systems include Optical Mark Recognition (OMR) systems and Electronic Ballot Printers (EBP), which are similar to DRE.

The progression of desktop computing to web-based applications has led to further innovations for creating two fundamental streams of e-voting systems. These are controlled voting, sometimes called poll-site e-voting, and uncontrolled voting, also referred to as remote e-voting. A controlled voting environment is a secure area that the Electoral Management Board (EMB) temporarily sets up, by installing equipment and implementing a clearly defined process flow [5, 11]. On the other hand, uncontrolled voting refers to a situation in which a voter remotely accesses a system from his/her own locality (home, office, or mobile) and cast a vote. Example of uncontrolled e-voting system is Internet Voting (also called i-voting, online voting, or online ballots). Oostveen [12] define i-voting as "an election system that uses encryption to allow a voter to transmit her secure and secret ballot over the Internet." The surge in Internet subscriptions and increasing availability of access points such as computers, mobiles, and iDTV have made i-voting increasingly attractive [13]. I-voting is currently not just a research topic as Krimmer et al. [14] inform of 104 Internet elections worldwide with 40% being binding elections.

Norway and Estonia are two countries that have incorporated i-voting into their electoral processes [8, 15]. While e-voting in its various forms provides several opportunities to solve many old electoral problems such as human errors, coercion, and inaccessibility, it reopens new problems. Most e-voting solutions are fully comprehensible to only a fraction of experts, which makes the integrity of the voting process rest substantially in the hands of few system operators rather than in thousands of electoral commission officials [8, 16]. Moreover, despite the huge investment in e-voting systems to enhance electoral processes in some political climes, it has been reported that voter turnout is dropping for reasons such as apathy, contentment, anger, boycott, disengagement, disinterest, or fear. For example, it is increasingly observed that youths are participating less in elections than other demographic groups [8]. Allen [17] however asserts that "for a democracy to command respect, it must operate in the same way as people do everything else in their lives." A survey of 1,200 Canadians by Goodman at Carlton University [18] also established that young citizens would vote online, if provided with Internet option. Similarly, the world is said to have surpassed the *mobile moment* when the number of mobile devices equals the number of people on the planet [19].

The ubiquity of the Internet and the global attainment of the *mobile moment* serve as a motivation to propose a *secure mobile Internet voting* architecture in this study. This voting scheme being proposed will be a paradigm shift in e-voting and it can potentially mediate some of the highlighted challenges facing the traditional electoral processes and the conventional e-voting systems. One of

the primary objectives of this research work is to report the proposed secure mobile Internet voting architecture, which we evolved from the well-known Sensus e-voting protocol [20, 21]. The second paramount objective of this study is to report experiments carried out to determine the most appropriate short-term spectral features for implementing the voice biometric authentication aspect of the voting scheme. The spectral features we examined are Mel-Frequency Cepstral Coefficients (MFCCs), Mel-Frequency Discrete Wavelet Coefficients (MFDWC), Linear Predictive Cepstral Coefficients (LPCC), and Spectral Histogram of Oriented Gradients (SHOGs). We utilized the Histogram of Oriented Gradients (HOG) algorithm which is reputed to be a good descriptor of spectral shape and appearance in [22] to both capture the discriminative content and dimensionally reduce the image spectrograms of the MFCC, MFDWC, and LPCC spectral features. The resulting dimensionally reduced Speaker discriminatory features were labelled, MFCC-HOG, MFDWC-HOG, and LPCC-HOG. SHOG features, which were utilized to classify persons into male and female gender, based on their speech input in [23], were also investigated. The neural network ensemble was generally utilized as the pattern matching algorithm in the four experimental models that were set up to investigate the short-term spectral features in this study.

2. Mobile Internet Voting

Mobile Internet voting (MI-voting) refers to the use of mobile devices to ubiquitously access wireless voting services on the Internet anytime and anywhere. Distributing the processing of votes over multiple web servers installed with a tamper-resistant provides an environment that can improve the security requirements of elections. This security process is made possible using the Smart Card Web Server (SCWS) on a mobile phone Subscriber Identity Module (SIM) [24]. In general, mobile Internet technology is the result of the convergence of networks of traditional Internet technology, broadband mobile networks, and mobile terminals [25]. Mobile Internet can play an important role by taking advantages of a large user base, surging sales of smartphones, tablets and 3G data, and exploration of other mobile electronic devices [26].

Mobile devices have become an important aspect of the modern society because they allow people to move from one place to another whilst they remain connected to others. The mobile technological revolution in the Information Communication Technology (ICT) industry has led to the creation of sophisticated mobile services such as m-television, m-payment, m-health, m-government, and m-banking to mention just a few. This is because, in most parts of the world, mobile devices have become highly ubiquitous because of their portability and affordability. The International Telecommunication Union (ITU) 2014 ICT statistics indicates that there are 6.9 billion mobile subscriptions in the world, out of which 4.5 billion are unique [27]. Amazon, Wikipedia, and Facebook also informed that about 20% of their traffic originates from mobile-only users. According to a Pew Internet report in 2012, 45% of young adults of the age between 18 and 29 claim that they browse the Internet mostly with a mobile device [28]. In South Africa also, 81.9% of households use mobile phones, out of which 30.8% use only mobile to access the Internet [29]. For the purpose of this paper, a mobile device is a communication system with the following desirable features [30, 31]:

(a) A small form factor, which refers to the size, shape, style, and layout of the device.

(b) At least one wireless network interface for data communications such as Wi-Fi and GSM/GPRS.

(c) Local built-in and nonremovable data storage.

(d) An operating system such as Android, iOS, Blackberry OS, or Windows.

(e) Applications available as a web browser or mobile apps acquired and installed from third parties.

(f) One or more wireless personal area network interfaces, such as Bluetooth or Near-Field Communications (NFC).

(g) One or more wireless network interfaces for voice communications, such as cellular phones.

(h) Global Positioning System (GPS), which enables location services.

(i) Battery powered.

Popular examples of mobile devices include cellular or smart phones, personal digital assistants, tablets, netbooks, and laptops [32].

Given the foregoing discussion, a secure mobile Internet voting architecture is creatively evolved in this work to take advantage of the aforementioned state-of-the-art technologies on the Internet and the mobile platforms. Subsequently, in this paper, the secure mobile Internet voting architecture is referred to as SMIV for convenience. The SMIV represents a voting paradigm, which ameliorates both the traditional electoral process and the conventional e-voting systems in innovative ways by leveraging on both mobile and Internet technologies. The SMIV architecture can potentially offer the following benefits [8, 33]:

(a) Probable decrease in costs of printing and transporting of paper ballots across the country.

(b) Mobile device ubiquity and frugality that position it as an alluring channel to realistically involve digital natives, rural populace, youths, healthcare workers, elderly, diplomats, soldiers, nomads, and the Diasporas who are unable to make it to poll-sites for casting of their votes because of their peculiar circumstances.

(c) Facilitation of voters mobility because mobile devices offer a platform for ubiquitous voting at any time and at one's convenience.

(d) Design customization of the interface to assist the totally or partially disabled members of the populace to cast their votes from their own habitations.

(e) Capability of multilingual instructions on the interfaces without an accompanying increase in the cost

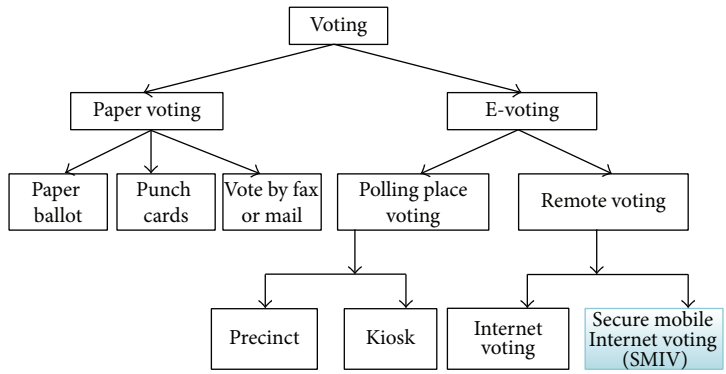

Figure 1: Different types of voting showing the placement of SMIV in the taxonomy [1].

of printing, which consequently reduces language bigotry.

(f) Expediting the concept of Bring-Your-Own-Device (BYOD), which serves to reduce operating expenses.

(g) Mitigating hacking attempts such as tempest attacks which is the electronic monitoring of radiation of voting screens to capture image and therefore monitors the vote, one of the reasons, the Dutch stopped e-voting.

The subsequent section of this paper describes the architecture of the proposed e-voting paradigm.

3. Secure Mobile Internet Voting Architecture

The secure mobile Internet voting (SMIV) system architecture we are proposing in this work is inherently a type of e-voting scheme as illustrated in Figure 1. Although there are different varieties of voting protocols that have been reported in the literature, the basic procedures for elections are generally standardized. These voting procedures often implement four specific sets of tasks, which are registration, collection, validation, and tallying [20]. Registration is comprised of the compilation of the list of eligible voters. Validation involves checking the credentials of someone that makes an attempt to vote and only allowing the eligible voters that have not yet voted to proceed. Collection job involves collecting the voted ballots while tallying has to do with the counting of the votes. In e-voting system design, ensuring that these tasks are carried out electronically is a primary objective. In order to realize this objective, no opportunity must be created for fraudulent practices that may breach the sanctity of the electoral process and thereby impair the trust the electorates have in the process.

Several sets of requirements have been established in the literature that e-voting schemes must satisfy in order to realize the aforementioned universal voting tasks [20, 34]. These requirements are grouped into two categories, which include generic and extended requirements. The generic requirements are as follows[20, 35–37]:

(a) *Accuracy*. An e-voting system is said to be accurate if it is impossible for a valid vote to be excluded, altered or to include an invalid vote in the final counting.

(b) *Privacy*. This involves the inability to link a voter to the vote he or she cast (i.e., anonymity) and the inability of the voter to prove the manner in which the vote was cast.

(c) *Verifiability*. An e-voting system is adjudged verifiable if a voter, an observer, or anyone can autonomously verify that all the cast votes were correctly tallied.

(d) *Eligibility*. Some authors refer to this requirement as invulnerability [20, 34] or democracy [36]. These all imply that an e-voting system permits only eligible voters to vote only once and nobody can vote more than once or vote for others.

Some other requirements categorized as extended requirements have also been established in the literature as necessary for the viability of e-voting systems. The examples in this category are as follows [35–37]:

(e) *Convenience*. An e-voting system should enable voters to vote easily and quickly and with minimal equipment and with no special expertise.

(f) *Mobility*. There should be no geographical restriction with respect to where voters decide to cast their vote. This requirement also implies that e-voting system is available and accessible during the voting phase regardless of where the voter decides to cast his or her vote.

(g) *Flexibility*. An e-voting system can be said to be flexible if it allows diversities of ballot question formats including open-ended questions. This is a very vital requirement for utilizing Short Message Services (SMS) on regular mobile devices and write-in candidate voting option.

(h) *Incoercibility*. An e-voting system is expected to be coercion resistant. Coercion takes place when an entity makes efforts to manipulate the manner in which a vote is cast, influences a voter to abstain, and/or represent a valid voter by obtaining the voter's credentials.

A voting system can be said to be secured if all the above stated requirements are satisfied [38]. Even though

all these security requirements are desirable for any type of voting system, apparently they are not all achievable by the conventional voting systems. Fulfilling the convenience, mobility, and flexibility requirements, for instance, is very vital for ensuring a high participation of voters in elections, although with a need to make provisions for sustaining the security requirements of privacy, eligibility, and incoercibility. Quite a number of voting schemes have been proposed in the academic literature over the past 25 years. These proposed schemes were targeted at fulfilling some of the generic requirements for e-voting and minimizing electoral frauds. Examples of these voting schemes include absentee balloting, vote by mail balloting, cryptographic protocols, two-agency protocol, one-agency protocol, FOO voting scheme, Sensus, SEAS, and EVOX [20, 21, 34, 38].

The SMIV architecture being proposed in this research work is based on the reference architecture of Sensus, a well-known security conscious Internet polling system [20, 21, 34], whose underlying foundation is the FOO, a practical secret voting scheme for large scale elections [38]. The security requirements fulfilled by Sensus are the first seven requirements stated earlier, although the authors of Sensus posit that the architecture does not fulfill the second aspect of privacy requirements, that is, the inability of the voter to prove the manner in which the vote was cast. In this current work, all the highlighted eight security requirements above are realizable in the SMIV architecture. However, security requirements such as eligibility, convenience, and mobility are realized differently from the Sensus approach. This is motivated by our desire to leverage on the recent advances in mobile, Internet, Global Positioning System (GPS), Near-Field Communication (NFC), and voice biometric technologies. The voter identification number and secret token were used to implement eligibility in Sensus [20] while, in this work, identification numbers, voice biometric, and GPS locations were utilized. The basis for the proposition of voice biometrics for realizing SMIV eligibility requirement is explained in detail in the latter part of this section. The security requirement of convenience implemented in Sensus with familiar devices and casting of vote in one or two sessions was enhanced in this work using NFC tag attached to the voter's ID card in addendum to the Sensus approach. Incoercibility which is a security requirement that must go hand in hand with the satisfaction of the mobility requirement was realized in our architecture using GPS service. Using the GPS service will help in controlling the maximum number of voters that can vote from a fixed location. This is an emulation of the polling booth or kiosk that is used for conventional electoral processes. GPS service can also help to mitigate hacking by either agenda-driven or disinterested foreign hackers and enhance the fulfillment of the accuracy requirements better than the implementation in Sensus.

The use of modern technologies to satisfy the requirements of eligibility, convenience, and accuracy as proposed in this study is an important contribution to e-voting research. Furthermore, the incoercibility security requirement that is satisfied with this work was not implemented in the Sensus architecture [20] and this also provides another important contribution of this work to the Sensus reference architecture and to e-voting research in general. Sensus is realized through three different modules described as follows [21]:

(i) *Pollster*. This is a module that serves as an agent for voters to anonymously, privately, and securely cast their ballot.

(ii) *Validator*. This is a server that first checks the eligibility of the pollster and the uniqueness of its submission. Once the pollster passes the eligibility criteria and the vote being submitted is unique, it validates the submitted vote.

(iii) *Tallier*. This is a server that collects and counts all the validated votes. The tallier confirms the authenticity of the validation and verifies that the encrypted ballot is unique. The tallier issues a signed receipt to the pollster once the ballot is valid and unique.

A preliminary phase of the Sensus architecture requires the participation of another entity named the registrar. An identifier, a secret token, and a public key are sent to the registrar to initiate the registration process. Accordingly, the registrar checks the validity of the token and updates a Registered Voters List (RVL) with the voter identifier and its public key. The RVL contains a validation field for each voter that is set to 0 before each election and changed to 1 by the validator after a voter's ballot is validated [20]. Cranor and Cytron [20], the authors of Sensus, posit that some election administrator may choose not to automate the registration process, that is, the registrar entity, for an election.

The functionalities of the registrar, the pollster, and the validator are extended in our proposed SMIV architecture so as to realize the earlier stated enhancements to the reference Sensus architecture. In codicil to the voters parameters required in Sensus for registration, the SMIV architecture's registrar expects the voter to also submit other information such as the voice biometric short-term features and the GPS location of the intended place for voting. This implies that the registrar in SMIV must be automated for voice biometric based voter's identification to be feasible. Since SMIV is targeted primarily at large scale elections, the SMIV registrar is to be implemented as a server and should not reside on the same machine as either validator or tallier to further enhance the security requirement of privacy. This increases the numbers of servers in SMIV to three instead of the two servers that were implemented in Sensus. Apart from the RVL that is generated by the registrar in SMIV, it also generates a machine learning based on the training ID entity for voter identification. The communication links between all the component modules in SMIV are assumed to be anonymous similar to the stipulation in Sensus.

The pollster in SMIV is incorporated with the capability to render the ballot using conventional software standards such XML/HTML, WML, and plain SMS so as to ensure compatibility with diverse devices and thoroughly satisfy the security requirement of convenience. The pollster also incorporates the capability of being loadable using the voter's ID card tagged with NFC, which is also a contributory factor in the realization of convenience. The validator in SMIV architecture incorporates functions that will make it capable

Table 1: The reference Sensus [20] and the proposed SMIV architecture's fulfillment of the eight e-voting security requirements.

Serial number	Sensus	The proposed SMIV
1	Accuracy Only one of several identical encrypted ballots got counted	Accuracy (i) Only one of several identical encrypted ballots got counted (ii) GPS location of place of voting to prevent agenda-driven or disinterested foreign hackers
2	Privacy (i) Blind signature and data encryption using the RSAREF encryption library (ii) Different servers run validator and tallier (iii) Pollster does not run on a machine that runs either validator or tallier (iv) Installation of personal copy of pollster on trusted machine by voter (v) Anonymous channel	Privacy (i) Blind signature and data encryption using the RSAREF encryption library (ii) Different servers run registrar, validator, and tallier (iii) Pollster does not run on a machine that runs either registrar, validator, or tallier (iv) Installation of personal copy of pollster on trusted machine by voter (v) Anonymous channel
3	Verifiability (i) Publishing of a list of encrypted ballot, decryption keys, and decrypted ballots (ii) Only voters can verify that their votes were counted correctly and correct any mistake anonymously	Verifiability (i) Publishing of a list of encrypted ballot, decryption keys, and decrypted ballots (ii) Only voters can verify that their votes were counted correctly and correct any mistake anonymously
4	Eligibility (i) Voters ID number (ii) Secret token (iii) Blinded validation certificate and signed receipt to certify uniqueness of vote	Eligibility (i) Voters ID number (ii) Voice biometrics (short-term spectral features) (iii) GPS location (iv) Blinded validation certificate and signed receipt to certify uniqueness of vote
5	Convenience (i) Familiar devices and user interfaces (ii) Casting of vote in one or two sessions	Convenience (i) Familiar devices and user interfaces (ii) Casting of vote in one or two sessions (iii) Voter ID card affixed with NFC tags for autoloading of voting application
6	Mobility Internet enabled computers	Mobility (i) Internet enabled computers (ii) Mobile devices
7	Flexibility Ballot description language (BDL)	Flexibility (i) XML/HTML (ii) WML (iii) Plain SMS
8	Incoercibility Not implemented	Incoercibility GPS location

of validating the pollster using the trained ID entity from the registrar. The trained ID entity inherently contains the identification numbers, voice biometric short-term features, and the GPS locations for all the voters. Table 1 illustrates the evolution of the proposed SMIV architecture from the Sensus reference architecture and a summary of the realization of the eight different security requirements that are satisfied by SMIV. The sequence diagram of the SMIV architecture is also shown in Figure 2.

Remote or precinct voter authentication has been touted as a nontrivial challenge even in established democracies that have adopted e-voting systems and it is also an essential component of e-voting system eligibility security requirements [16]. Several approaches have been proposed for its implementation in a secure manner and one of the chief propositions is the use of biometrics. These are automatic methods of identifying or verifying the identity of a person based on behavioral and physiological physiognomies. Examples of human features that are used in biometric systems include fingerprint, voice, iris, face, signature, DNA, hand, keystroke, gait, handwriting, and finger shape. Biometric authentication involves the comparison of an enrolled biometric template against a freshly taken biometric sample [39, 40]. This authentication mechanism can be used in verification or identification modes. Verification mode involves the validation of a person's identity by comparing the freshly captured sample with its own template and the system conducts a one-to-one comparison to determine if the claim is true or not. Furthermore, verification helps to achieve positive recognition in which the goal is to inhibit multiple individuals from using the same identity. In identification mode, the biometric system distinguishes a person by searching

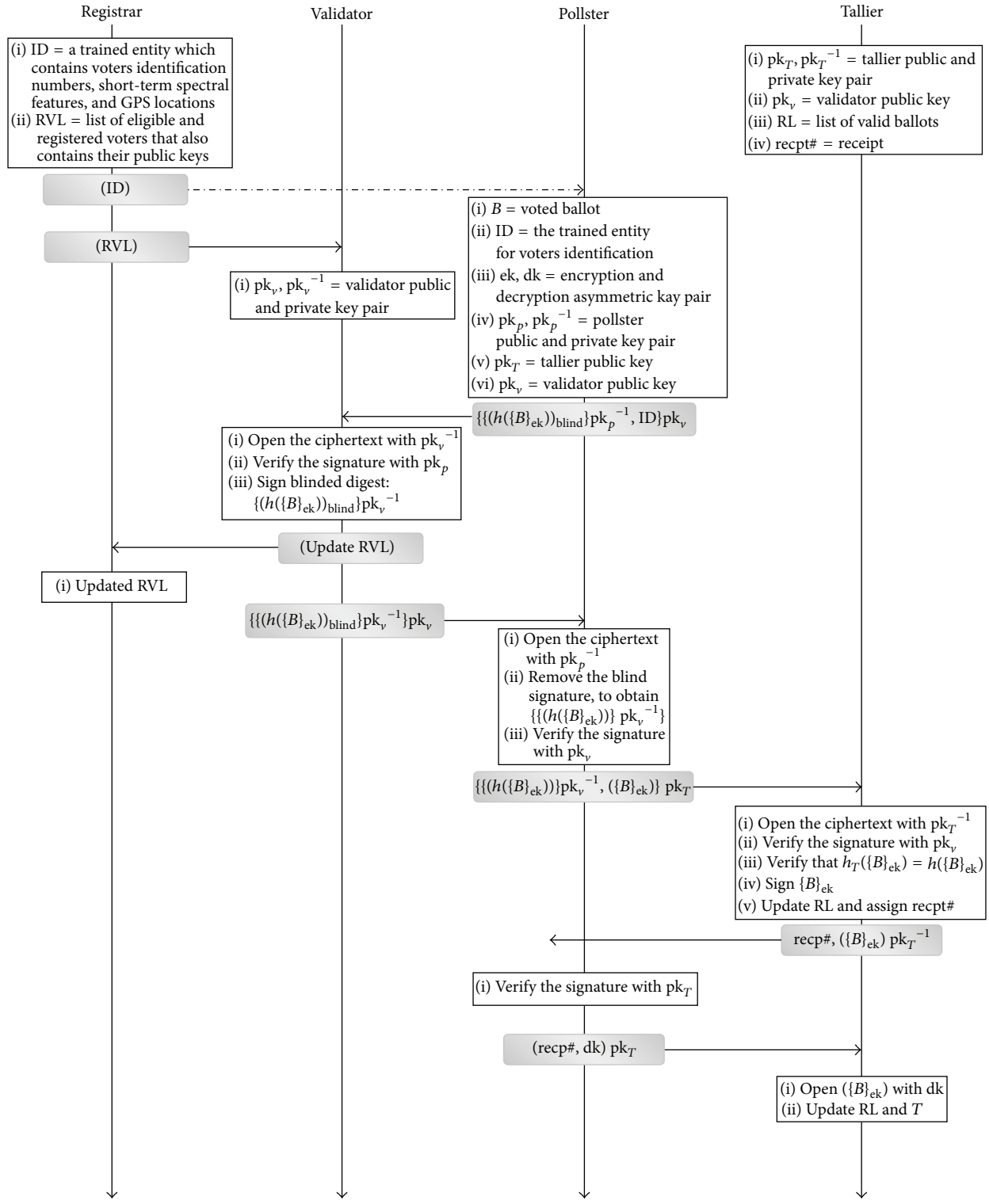

Figure 2: Sequence diagram of SMIV architecture; an extension of the Sensus reference architecture adapted from [21].

the templates of all the users for a tie by conducting a one-to-many search. Identification is often employed for negative recognition in which the system establishes whether an individual is who she/he disagrees to be so as to prevent a single individual from using multiple identities [41]. Older methods such as PIN, password, tokens, and keys may be engaged successfully for positive recognition, but, for establishing negative recognition, only biometrics will suffice. Worthy of note is the fact that both verification and identification are generically called recognition.

An ideal biometric feature must therefore fulfill essential requirements such as robustness, uniqueness, universality, permanence, collectability, performance, and acceptability, which are explicitly defined in [42]. Fulfilling these requirements is strongly dependent on the application domain, the population, and the hardware or software systems in use. The performance of one application cannot be predicted from tests carried out on another application. However, voice biometric has been acclaimed in the academic literature as a good choice for phone based applications and other applications that require remote authentication [42–46]. Voice biometrics is a combination of physiological and behavioral characteristics. Although the physiological features (i.e., vocal tract structures, which are also known as *short-term features*) of human speech are invariant for each person, the behavioral features vary over time as a result of age, medical conditions, and emotional state [40]. Voice biometric recognition systems can be either text-dependent or text-independent. Text-independent voice recognition systems are based on the utterance of a fixed word or phrase while text-independent system distinguishes an individual regardless of the uttered word or phrase. A text-independent voice recognition even though very challenging to design offers more protection against scams [40]. Voice biometrics is also referred to as Speaker recognition, which is the process of automatically recognizing who is speaking on the basis of the information inherent in the speech waves. Speaker recognition is a different technology compared to speech recognition in which computer algorithms extract features of the spoken utterance to determine the word that is spoken [47].

Given the foregoing discussion in the last two sections, a *text-independent Speaker identification* approach using short-term spectral features is nominated for fulfilling the authentication requirement of the proposed SMIV architecture. As earlier established, this choice satisfies the biometric *negative recognition* paradigm, which prevents a single individual from using multiple identities [41]. Consequently, this choice will also satisfactorily enhance the realization of both mobility and eligibility security requirements better than the method used in Sensus. The block diagram for the enrollment and identification of a biometric authentication system is shown in Figure 3. For convenience, recognition is subsequently used in place of identification in this paper.

Experimentations were carried out in this work to determine the best combinations of algorithms for realizing the *text-independent Speaker recognition* authentication of the SMIV architecture. Meanwhile, the subsequent section details the theoretical foundation of short-term feature-based Speaker recognition systems.

4. Theoretical Foundation for Speaker Recognition Authentication

A Speaker recognition authentication system is analogous to the diagram in Figure 3 and typically is comprised of the digitization of the analogue speech, preprocessing of the digitized speech, extraction of discriminating speech features, dimension reduction of the extracted features, training of pattern matching model, and recognition of Speakers through pattern matching with the trained model. The features for Speaker recognition are divided into short-term spectral features, voice source features, spectral-temporal features, prosodic features, and high-level features. The short-term features have hitherto been the dominant features in Speaker recognition systems because of their stability, ease of extraction, requirement of a small amount of data, text and language independence, and less computational requirements. The most prominent short-term features in the literature include Mel-Frequency Cepstral Coefficients (MFCCs), Mel-Frequency Discrete Wavelet Coefficients (MFDWC), and Linear Predictive Cepstral Coefficients (LPCC) [48, 49].

Mel-frequency is the measure of the human perception of the frequency content of speech signals on the "Mel scale." Mel-Frequency Cepstrum (MFC) stands for the short-term power spectrum of the speech, based on a linear cosine transform of a log power spectrum, computed on the nonlinear Mel-frequency. The MFCCs are, therefore, the coefficients that collectively make up the MFC. The frequency bands in the MFC are equally spaced and from research findings in the psychophysical field it has been established that the Mel scale approximates the auditory system of humans better than linearly spaced frequency bands. The computational components of the MFCC algorithm are captured in Figure 4 [50].

Assuming that $x[n]$ is the digitized version of the input speech signal with sampling frequency f_s it is divided into P frames each of length N samples with an overlap of $N/2$ samples. x_p denotes the pth frame of the speech signal $x[n]$ and to compute the MFCC of the pth frame x_p is multiplied with hamming window. The hamming window is given as

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad n = 0, 1, \ldots, N-1. \tag{1}$$

The windowing function is purposely for the smoothening of the signal for the computation of the Discrete Fourier Transform (DFT). The DFT is used for computing the frequency response of each frame to generate the spectrogram of the speech signal as

$$X[k] = \sum_{n=0}^{N-1} x[n]\, e^{-j2\pi kn/N}, \quad k \in Z. \tag{2}$$

The relationship between the Mel-frequency and linear frequency is

$$\text{mel}(f) = \begin{cases} 2595\log_{10}\left(1 + \dfrac{f}{700}\right) & \text{if } f > 1\,\text{kHz} \\ f & \text{if } f < 1\,\text{kHz}, \end{cases} \tag{3}$$

where $\text{mel}(f)$ is the Mel-frequency scale and f is the linear frequency. The Mel-filter bank filters the magnitude spectrum that is passed to it to give an array output called Mel-spectrum. Each of the values in the Mel-spectrum array corresponds to the result of the filtered magnitude spectrum

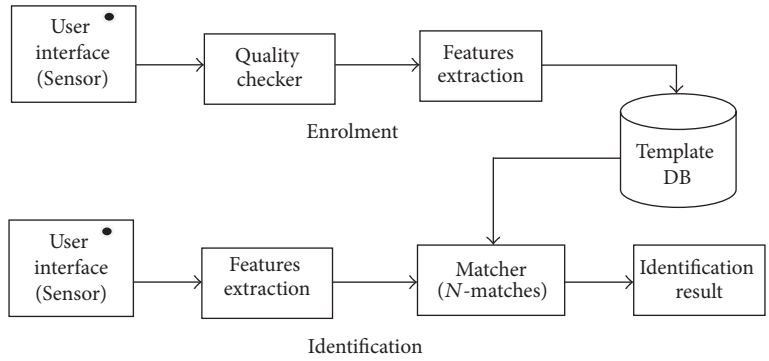

Figure 3: Generic block diagram for enrollment and identification of biometric authentication system.

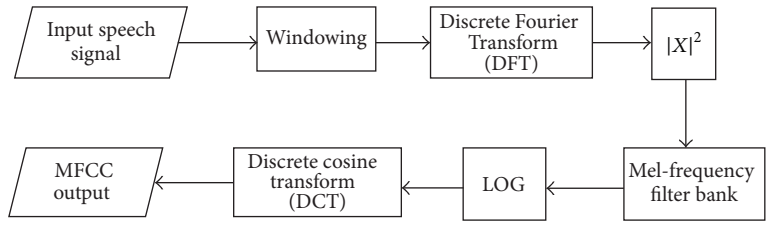

Figure 4: MFCC computational components.

through the individual Mel-filters. The Mel-spectrum is given as

$$Y(n) = \sum_{k=0}^{N/2} |X[k]| * \text{Mel-Weight}[n][k], \quad 0 < n < M, \tag{4}$$

where M represents the number of filters. The MFCC features are computed by taking the log of the Mel-spectrum and computing the DCT as follows:

$$C_n = \sum_{n=1}^{N} [\log Y(n)] \cos\left[k\left(n - \frac{1}{2}\right)\frac{\pi}{N}\right] \quad \forall k = 1, \ldots, M. \tag{5}$$

The C_0 is omitted from the DCT computation because it represents the mean value of the input speech that contains little Speaker unique information but rather contains information on the microphone used for recording the speech signal.

The MFCC feature vector is obtained per Speaker by retaining about 12–15 lowest DCT components [51, 52]. MFDWC features are calculated using similar procedures to the computation of MFCC features. However, the DCT computation in the last step is substituted with the Discrete Wavelet Transform (DWT). DWT is acclaimed to allow better localization in both time and frequency domains and based on this the MFDWC has been shown to give better performance in noisy environments [53]. Linear Predictive Coding (LPC) is an alternative spectrum estimation method to DFT. It has a good intuitive interpretation in both frequency and time domains. Given a signal, $s[n]$ in the discrete time domain, the LPC prediction error is given as

$$e[n] = s[n] - \sum_{k=1}^{p} a_k s[n-k], \tag{6}$$

where a_k are the coefficients of the predictor. Assume that $s[n]$ is the speech signal and $e[n]$ is the voice source (or glottal pulses) [53]. Equation (6) is transformed to

$$E[z] = S[z]\left(1 - \sum_{k=1}^{p} a_k Z^{-k}\right). \tag{7}$$

The spectral model is therefore given as

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k Z^{-k}}, \tag{8}$$

where a_k are the predictor coefficients that are often computed by minimizing the residual energy using the Levinson-Durbin algorithm [54] and $H(z)$ is the spectral model. However, these predictor coefficients are infrequently used as features; rather, they are transformed to the more robust LPCC features using a recursive algorithm proposed by Rabiner and Juang [55]. Unlike MFCC features, the LPCC features are not based on the auditory perceptual frequency scale.

The output of the computed short-term features (i.e., MFCC, MFDWC, and LPCC) is essentially 2-dimensional matrices, which can be described analogously as 2-dimensional digital image signals [57]. More so, short-term spectral features have been described as the acoustic correlate

of the "color" of sound in [58]. These matrices therefore can be processed further using digital image processing algorithms to achieve dimension reduction, because one of the expected characteristics of an ideal feature in the Speaker recognition is that the number of features should be relatively low to prevent the curse of dimensionality in which the required training samples grow exponentially with the number of features. Another important benefit of compact features is the reduction in the computational complexity of the pattern matching models in Speaker recognition systems [58, 59]. Examples of dimension reduction methods in signal processing, image processing, and statistics include Principal Component Analysis (PCA), Factor Analysis (FA), Independent Component Analysis (ICA), projection pursuit, random projections, vector quantization, and Histogram of Oriented Gradient (HOG). The HOG is a recent descriptor developed by Dalal and Triggs in 2005 [22] that can effectively capture the local appearance and shape information by encoding the spectral gradient orientation from the output of the short-term features as histograms. The algorithm has been reputed to be successful in recent applications such as pedestrian recognition, activity recognition, and speech processing [22, 60, 61]. This algorithm is adopted for the dimension reduction task in this work.

Furthermore, Artificial Neural Network (ANN) ensemble is selected as the pattern matching method. ANNs have been used for several years in both speech and Speaker recognition systems because of their high accuracy, noise tolerance, and nonlinear property [47]. Meanwhile, ensemble learning improves ANNs performance by giving better accuracy than a single ANN [56, 62]. In machine learning, the idea of ensemble learning is to engage multiple learners and combine their predictions. Ensemble of learning models is known to enhance the performance of single models by giving better accuracy than the individual members of the ensemble. One of the most effective methods used for constructing ensembles is the manipulation of the training samples to generate multiple models [63]. In this method, the learning algorithm is run in several iterations with a different subset of the training samples at each iteration. This method is known to work efficiently with unstable learning algorithms such as decision tree and neural network. Examples of different algorithms used for manipulating the training datasets are bagging, cross-validated committees, and AdaBoost [64]. Bagging was developed in 1996 and it means bootstrap aggregation. It is reputed as the first effective method of ensemble learning and is one of the simplest methods [65]. The method creates multiple versions of a training set by sampling with replacement and each of the resampled datasets is used to train a different model. The output of the model is often combined by averaging or voting depending on the nature of the problem. Bagging is adopted for this work to leverage on its benefits.

5. Experimental Results

5.1. Data Collection. For the experimentation aspect of this work, speech signals of selected Speakers were recorded at 11025 Hz sampling rate as ".wav" files using a Logitech microphone. The recording was done in MATLAB R2012a at 16 bits per sample for 10-second duration. The phrases that were uttered sequentially by the Speakers for each instance of recording are as follows:

(i) *"Hello Hello Hello . . .,"*

(ii) *"1 2 3 . . .,"*

(iii) *"A, B, C . . .,"*

(iv) *"Yes Yes Yes . . .,"*

(v) *"No No No . . ."*

The five different utterances were repeated three times by each Speaker to generate fifteen utterances per Speaker for the training dataset. One of these five utterances was read in languages other than English such as Zulu (South Africa), Yoruba (Nigeria), Halychyna (Carpathian region of Europe), and Hindi (India). This is to introduce diversity into the dataset and emulate the reality in a typical multiethnic population that may want to implement the voice biometric authentication in the proposed SMIV architecture. Furthermore, any two of the five phrases were uttered and recorded for each Speaker to generate two text-independent test samples per Speaker. The distribution of the generated datasets is shown in Table 2. A Graphical User Interface (GUI) shown in Figure 5 was designed and implemented in MATLAB R2012a for recording the speech utterances for our experiments. The waveforms of the "*Hello Hello Hello . . .*" utterances for eight of the enrolled twenty Speakers are shown in Figures 6 and 7 while the spectrograms of the utterances are shown in Figures 8 and 9. Both waveforms and spectrograms (Figures 6, 7, 8, and 9) clearly illustrate the variations in the patterns of the speech signal from one Speaker to the other. Only the plots for the first eight out of the twenty Speakers enrolled are reported in this paper because of space constraint. We created our own in-house speech database for this study rather than using existing databases such as TIMIT, NTIMIT, IISC, and YOHO [66, 67] because of the need to introduce diverse languages into the dataset. This is to add a unique flavor of investigating the acclaimed text and language independence of short-term spectral features [58]. We also made a choice of a compact dataset for our investigation in this study based on the established fact in the literature that only a small amount of data is necessary for short-term spectral features [58].

Table 2: Distribution of the dataset used in the experiments.

Number of Speakers	Numbers of speeches recorded per Speaker		Total
20	Training	15	300
	Testing	2	40
	Total	**17**	**340**

Four different experimental models were designed for this study so as to determine the appropriate combination of algorithms to realize an optimal Speaker recognition aspect of the SMIV architecture. All the experiments were

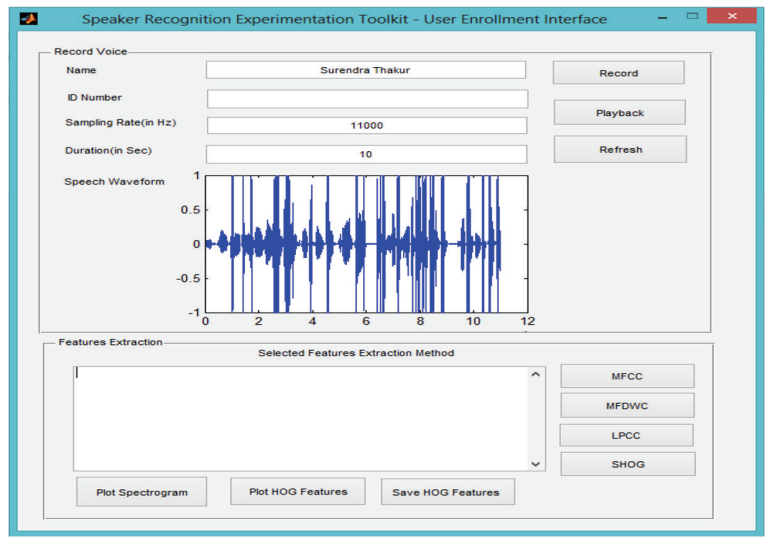

Figure 5: Speaker recognition experimentation toolkit (SRET) user enrollment interface.

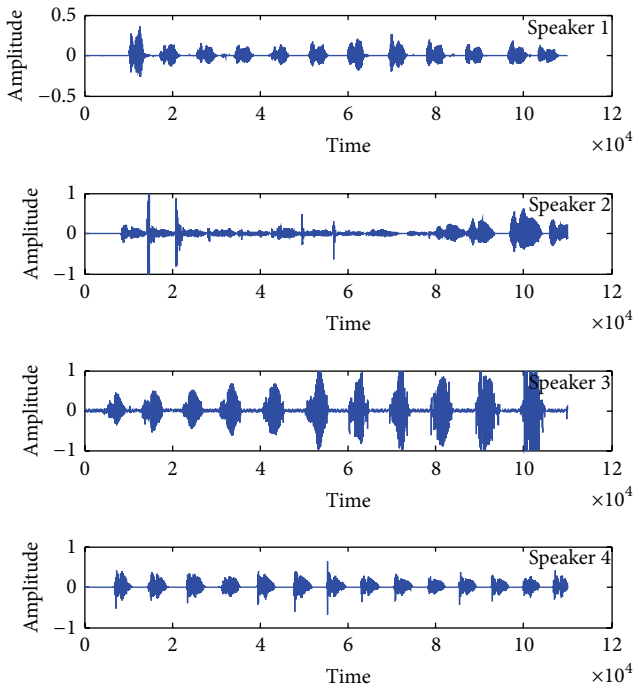

Figure 6: Waveforms corresponding to the utterance "*Hello Hello Hello* ..." by Speakers 1–4.

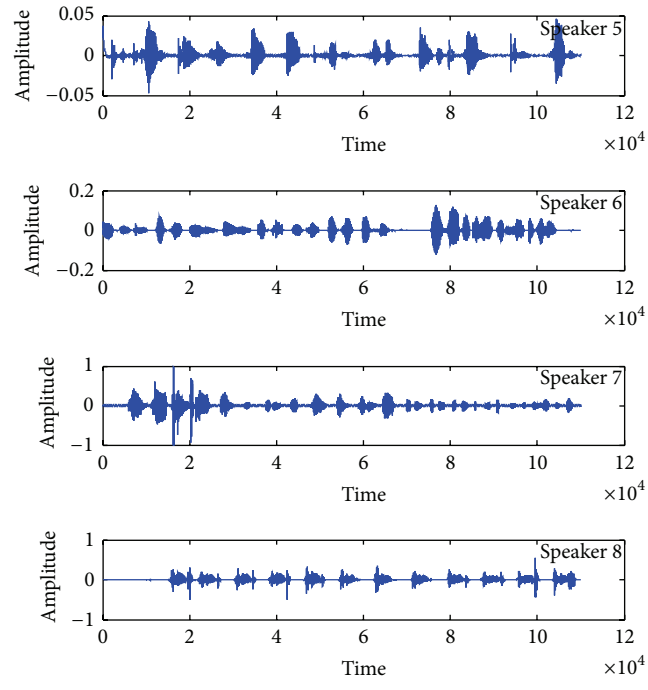

Figure 7: Waveforms corresponding to the utterance "*Hello Hello Hello* ..." by Speakers 5–8.

performed on a computer system with Intel Core i5-3210M CPU operating at 2.50 GHz speed. The computer system also has 6.00 GB RAM, 500 GB Hard disk and it runs 64-bit Windows 8 operating system. The experiments and the results obtained are reported in the subsequent subsections.

5.1.1. Experiment 1. The architecture of the model for experiment 1 is as shown in Figure 10. As illustrated in the figure, the first block involves the capturing and digitization of the analogue speech signal using a microphone and the Personal Computer (PC) in the MATLAB R2012a environment. Sample waveforms and spectrograms generated from this first block are as shown in Figures 6 to 9.

The preprocessing and features extraction block shown in Figure 10 were implemented with the Mel-Frequency Cepstral Coefficients (MFCCs) algorithm. The MFCC computational components shown in Figure 4 were implemented in this study using MATLAB R2012a. The digitized speech signal for each of the seventeen utterances from the different

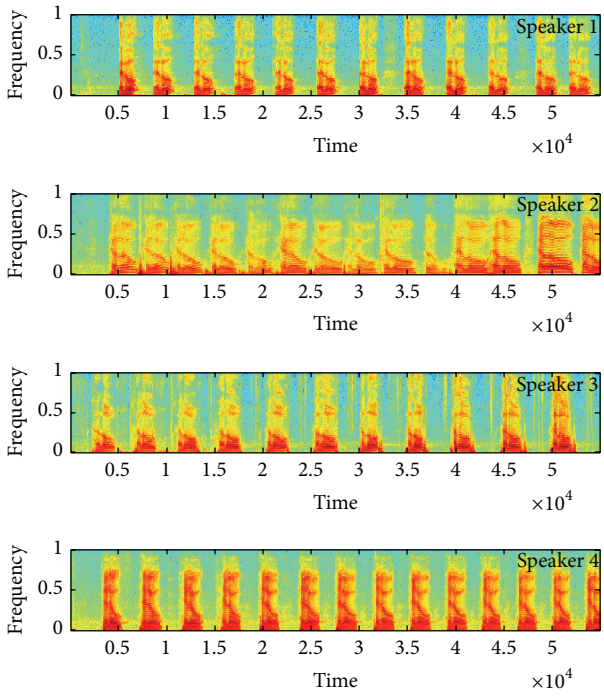

Figure 8: Spectrogram corresponding to the utterance "*Hello Hello Hello* ..." by Speakers 1–4.

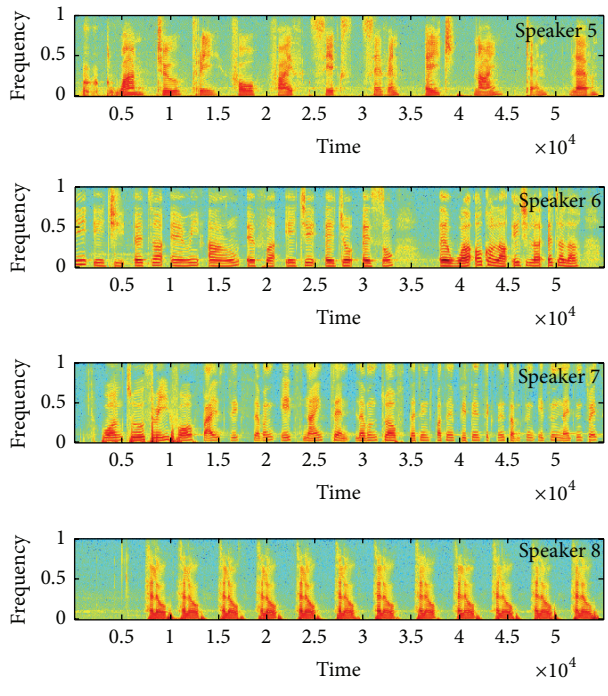

Figure 9: Spectrogram corresponding to the utterance "*Hello Hello Hello* ..." by Speakers 5–8.

Speakers served as inputs into the MFCC code and each utterance generated a 12 × 1374 MFCC matrix as outputs. The image plots of the MFCC matrices for the utterance "*Hello Hello Hello* ..." for the first eight Speakers are shown in Figures 11 and 12. As shown graphically in the figures, the patterns of the features extracted for each of the utterances by different Speakers are uniquely different.

In this study, the HOG block in Figure 10 was implemented in MATLAB R2012a to reduce the 12 × 1374 MFCC matrix for each utterance to a feature vector of 81 elements. This is an important procedure for reducing the complexity and computational time of the subsequent ensemble learning network in the model. The time and frequency domain plots of the HOG features for the utterance "*Hello Hello Hello* ..." for the first eight Speakers are shown in Figures 13 and 14. Both figures illustrate that the dimensionally reduced HOG features for the Speakers have similar patterns because they represent the utterance of the same set of words; however, despite the similarity in the patterns, the pattern for each of the Speakers is unique in both time and the frequency domains. This is an illustration of the capability of the HOG algorithm to both reduce the dimensions of the extracted MFCC features and still retain the discriminatory features for each of the Speakers in the dataset.

The next computational block in the model for the current experiment is the design and training of the pattern matching platform, which automates the Speaker recognition task. The selected pattern matching method is the ANN ensemble. As earlier stated in Section 4, ANN ensemble enhances the output of single ANNs by giving better accuracy than the individual base ANNs in the ensemble. In order to create an ensemble of ANNs, the base ANN has to be properly configured so as to achieve high performance with low network errors. For this study, the configuration of the base ANN is 500 training epochs with the dataset partitioned to 70% training, 15% testing, and 15% validation. It was shown by Cybenko [68] that a network with one hidden layer is able to approximate any continuous function. However, the authors in [69] posit that more hidden layers with a high number of neurons generally lead to small local minima. On these bases and from the outcome of our experimentations, we selected 2 hidden layers and 80 neurons in each hidden layer for the base ANN in this study. The activation functions selected for each layer of a network are also important in configuring the base ANN for ensemble learning. For the input layer, the linear activation function was selected because this layer is only required to convey the input data to the succeeding layer without any alteration. The power of MLP-ANN, which is the topology adopted for the base ANN in this study, comes from nonlinear activation in the hidden layer. The most commonly used functions are the logistic and hyperbolic tangent functions because of their nonlinearity and differentiability [69]. The hyperbolic tangent was however selected for the hidden layer neuron of the base ANN. This is because the function is symmetrical to the origin and decreases the speed of convergence during training. Hyperbolic tangent is also selected for the output layer neuron because the function is adjudged appropriately for binary output patterns [68, 69]. There are 81 neurons in the input layer for the base ANN in this study in conformity with the number of elements in the HOG features vector. We also have 5 neurons in the output layer since there are 20 unique

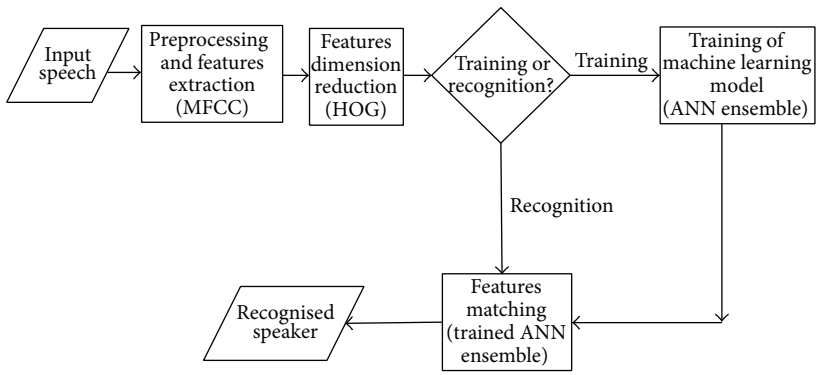

Figure 10: Architecture of the model for experiment 1.

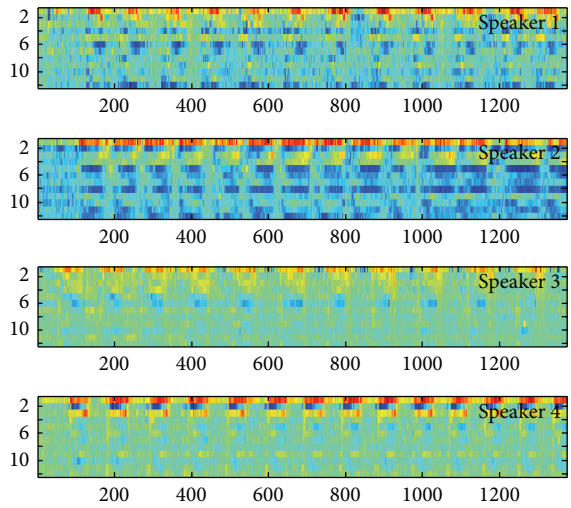

Figure 11: The MFCC images for the utterance "*Hello Hello Hello* ..." for Speakers 1–4.

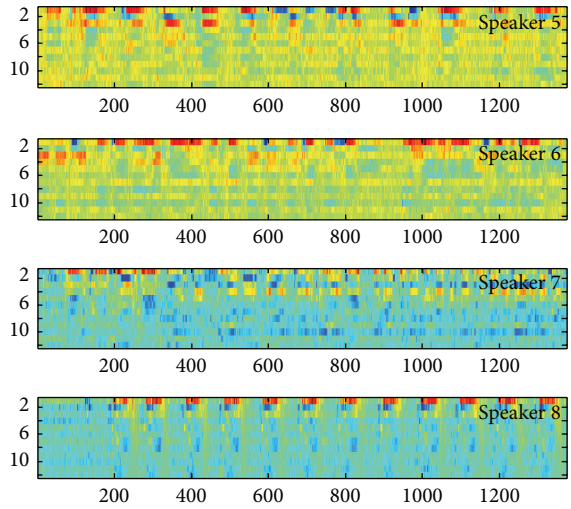

Figure 12: The MFCC images for the utterance "*Hello Hello Hello* ..." for Speakers 5–8.

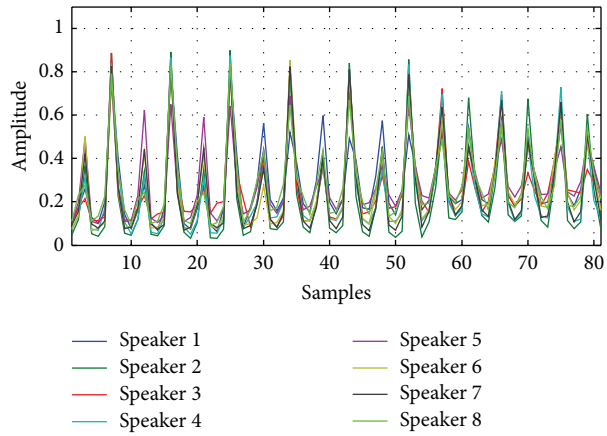

Figure 13: Time domain plot of the MFCC HOG features for the utterance "*Hello Hello Hello* ..." of the first 8 Speakers.

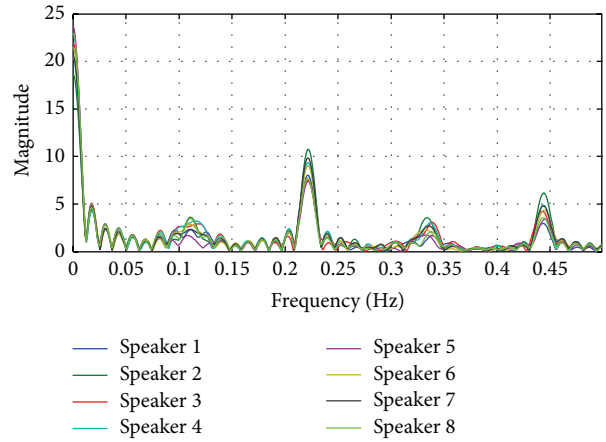

Figure 14: Frequency domain plot of the MFCC HOG features for the utterance "*Hello Hello Hello* ..." of the first 8 Speakers.

Speakers in the dataset and permutations of 5 binary patterns are sufficient for the unique identification of the 20 Speakers. The architecture of the fully configured base ANN used in this study is shown in Figure 15 while the target output binary patterns of the output neurons are shown in Table 3.

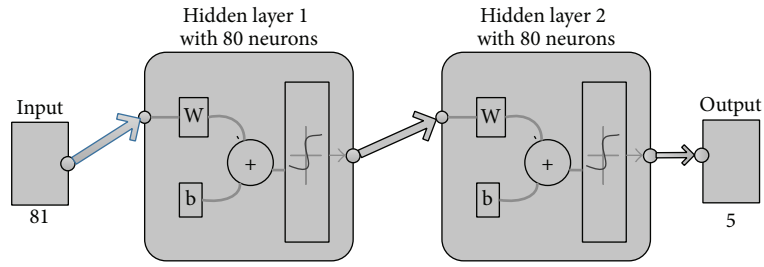

Figure 15: Architecture of the configured base ANN [56].

Table 3: Target outputs of the base ANN for each Speaker.

Speaker	Target output
Speaker 1	00001
Speaker 2	00010
Speaker 3	00011
Speaker 4	00100
Speaker 5	00101
Speaker 6	00110
Speaker 7	00111
Speaker 8	01000
Speaker 9	01001
Speaker 10	01010
Speaker 11	01011
Speaker 12	01100
Speaker 13	01101
Speaker 14	01110
Speaker 15	01111
Speaker 16	10000
Speaker 17	10001
Speaker 18	10010
Speaker 19	10011
Speaker 20	10100

The configuration of the ANN ensemble is another critical aspect in the design of ensemble learning systems. A study in [70] trained an ensemble of 32 neural networks to identify volcanoes on Venus. In the study at hand, 50, 100, and 200 base models were tested using bagging ensemble and plurality voting for combining their predictions. Our result gave better prediction accuracy with moderate complexity using 100 base models in the ensemble. It is however already established in the literature that having a high number of models is advantageous in problem domains with a small dataset, which is the case in this study [63]. The bagging ensemble algorithm adopted for this study was implemented in MATLAB R2012a using appropriate functions in the Statistical and Neural Network Toolboxes. The performances of the base ANNs were evaluated based on statistical measurements such as linear regression R value and mean square error (MSE) [71]. The R values and MSEs of the 100 base models in the ANN ensemble for the current experimental model are plotted as shown in Figure 16.

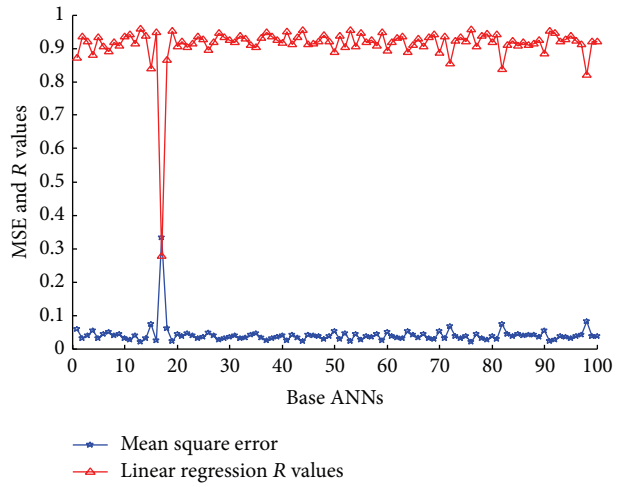

Figure 16: MSE and R values of the 100 base ANNs for experiment 1.

The average MSE for the ANN ensemble in this first experiment is 0.0411 and the average R value is 0.9123. This ANN ensemble was used to predict two test samples from each of the Speakers. Twenty-nine samples were correctly predicted out of the forty test samples. We performed more experiments as reported in subsequent subsections before a conclusion is made about the suitability of this result.

5.1.2. Experiment 2. In the current experiment, the experimental model was obtained by replacing MFCC in the first experimental model (Figure 10) with Mel-Frequency Discrete Wavelet Coefficients (MFDWC) as the short-term features. The MFDWC algorithm was implemented in MATLAB R2012a and the image plots of the output matrices for the utterance "*Hello, Hello, Hello* ..." by the first eight Speakers are shown in Figures 17 and 18. It is shown in these figures that the MFDWC pattern for each of the Speakers is unique.

Similar to the procedure in experiment 1, the HOG algorithm was further utilized to reduce the dimensions of the 12×1374 MFDWC feature matrices in order to obtain 81-element feature vector for each of the Speakers in our dataset. The time and frequency domain plots of the MFDWC-HOG features for the utterance "*Hello Hello Hello* ..." for the first 8 Speakers are shown in Figures 19 and 20.

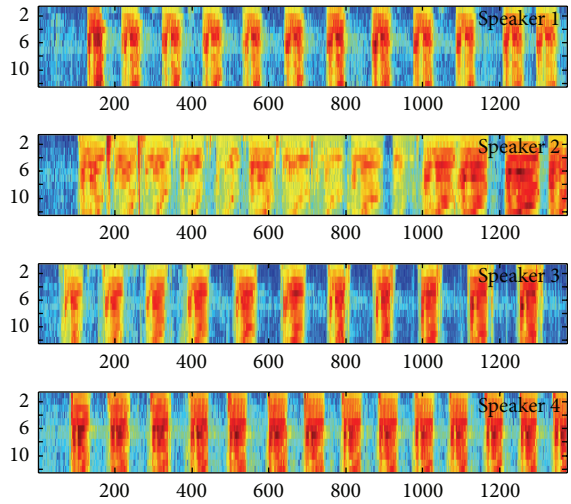

Figure 17: The MFDWC images for the utterance "*Hello Hello Hello* ..." for Speakers 1–4.

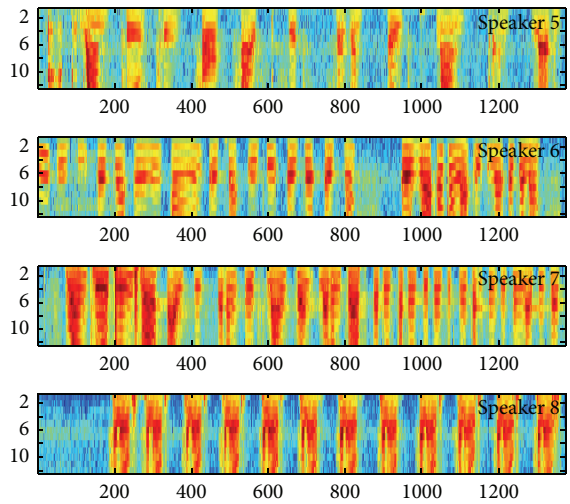

Figure 18: The MFDWC images for the utterance "*Hello Hello Hello* ..." for Speakers 5–8.

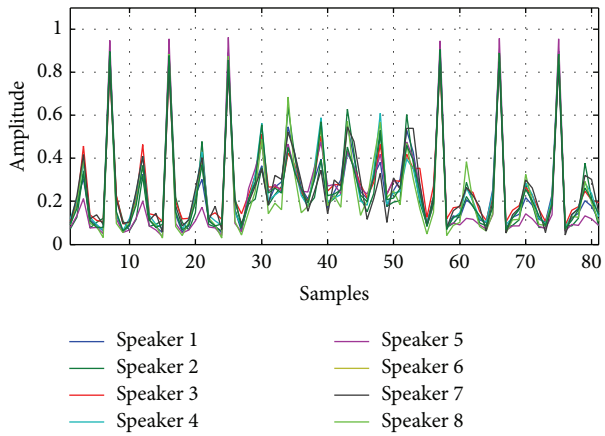

Figure 19: Time domain plot of the MFDWC HOG features for the utterance "*Hello Hello Hello* ..." of the first 8 Speakers.

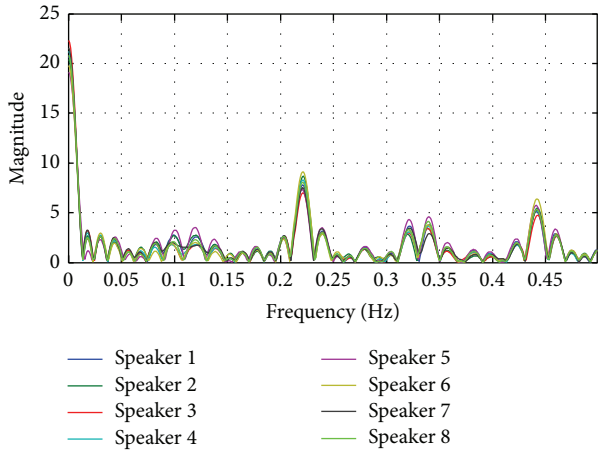

Figure 20: Frequency domain plot of the MFDWC-HOG features for the utterance "*Hello Hello Hello* ..." of the first 8 Speakers.

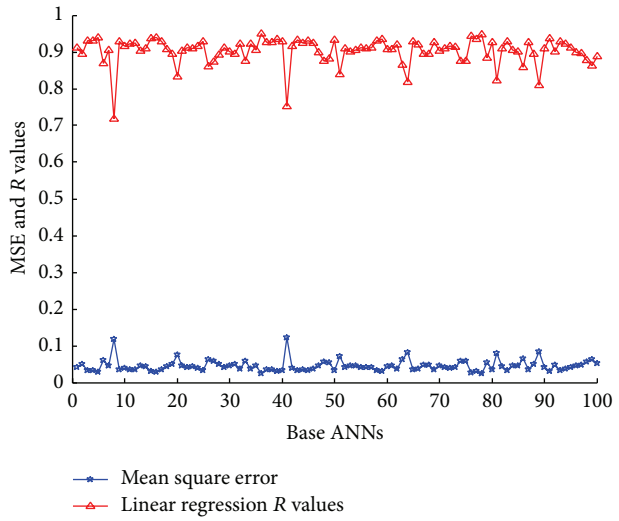

Figure 21: MSE and R values of the 100 Base ANNs for experiment 2.

As shown in Figures 19 and 20, although the shapes of the MFDWC-HOG features for the different Speakers are similar, the sizes of the shapes are unique and this provides a strong basis for using machine learning to uniquely identify each of the Speakers. Consequently, the next block in the model for experiment 2 is the training of the ANN ensemble with the MFDWC-HOG features. The configuration of the ANN ensemble in experiment 1 is also used in the current experiment and the results obtained for the 100 base ANNs in this experiment 2 are illustrated with the plot in Figure 21.

The average MSE for the ANN ensemble in the second experiment is 0.0455 and the average R value is 0.9028. Furthermore, the trained ANN ensemble was tested with two utterances from the test datasets for each of the Speakers. Out of the forty test samples, twenty-nine were correctly predicted. The ANN ensemble trained with MFDWC-HOG features in the current experiment gave a slightly lower statistical performance result than was obtained in experiment 1.

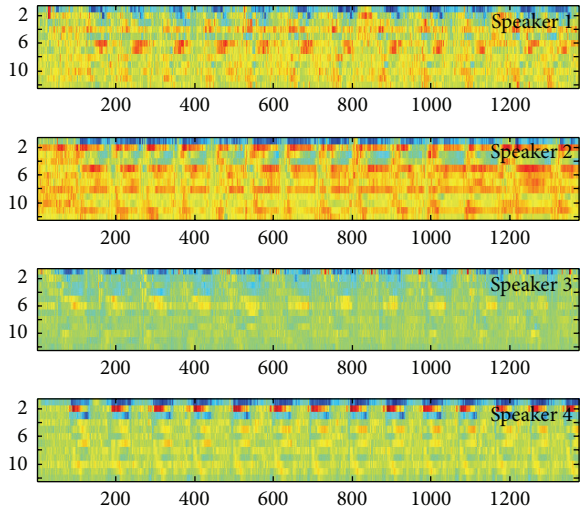

Figure 22: The LPCC images for the utterance "*Hello Hello Hello* ..." for Speakers 1–4.

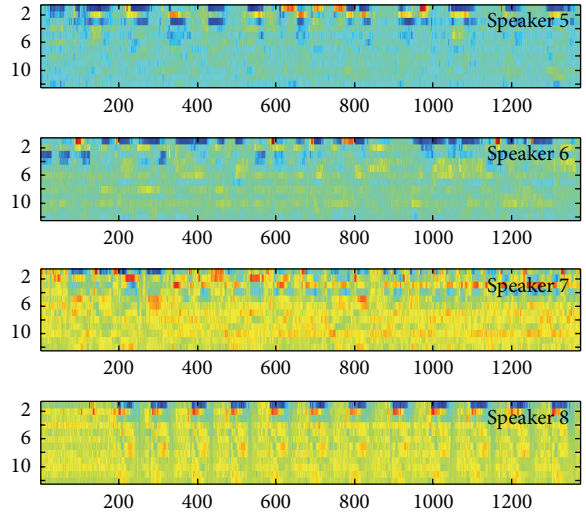

Figure 23: The LPCC images for the utterance "*Hello Hello Hello* ..." for Speakers 5–8.

However, the same number of test samples was correctly recognized in the two experiments. The similarity in the test results obtained from both experiments 1 and 2 is noteworthy, but these results are a little below the expected level of recognition for the authentication module of the SMIV architecture. Consequently, we set up another experimentation model, which is reported in the next subsection.

5.1.3. Experiment 3. The architecture of the model for experiment 3 was derived by using the Linear Predictive Cepstral Coefficients (LPCC) features extraction algorithm [48, 49] for the preprocessing and features extraction block in the architecture shown in Figure 10 and this distinguishes it from previous experiments 1 and 2. The LPCC features extraction algorithm was implemented in MATLAB R2012a and was similar to what was done in experiments 1 and 2; the image plots of the LPCC feature matrices for the utterance "*Hello, Hello, Hello* ..." by the first 8 Speakers are shown in Figures 22 and 23. The patterns of the outputs of LPCC feature matrices shown in these figures are different from the MFCC and the MFDWC patterns shown in Figures 11, 12, 17, and 18, respectively. This is a confirmation of the methodological differences among the different short-term spectral features. The patterns of the LPCC feature matrices for each Speaker are also unique and this is a reflection of the discriminatory power of the LPCC features.

Similar to experiments 1 and 2, the next procedure implemented was the dimension reduction of the 12×1374 LPCC feature matrices using HOG algorithm. The time and the frequency domain plots of the 81-element LPCC-HOG feature vector obtained for each of the Speakers in this experiment 3 are shown in Figures 24 and 25. These features which are unique for each Speaker as shown in both the time and the frequency domain plots are utilized to train the ANN ensemble of the same configuration as was used in experiments 1 and 2. The values obtained for the MSEs and R values for the 100 base ANNs in the current experiment are illustrated in Figure 26.

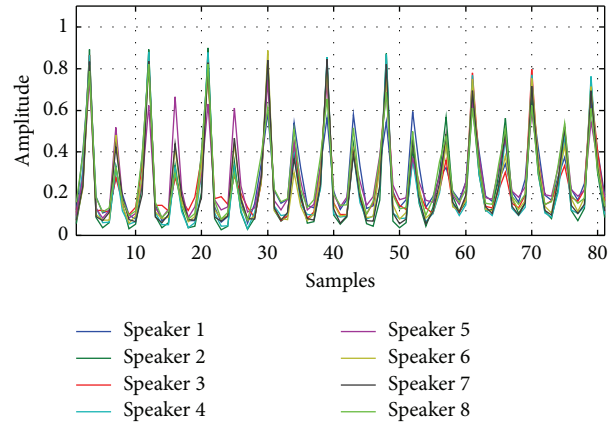

Figure 24: Time domain plot of the LPCC-HOG features for the utterance "*Hello Hello Hello* ..." of the first 8 Speakers.

An average MSE of 0.0407 and average R value of 0.9127 were obtained for the ANN ensemble trained with LPCC-HOG features in experiment 3. These statistical measures of performance obtained in the current experiment are better than those obtained in the two previous experiments. This is an illustration of a stronger discriminatory capability of LPCC features over both MFCC and MFDWC features. In order to further validate the current result, the ANN ensemble in this experiment was tested using two test samples from the test dataset for each of the Speakers. Out of the forty test samples, thirty samples were correctly predicted. This is a further validation of the stronger efficacy and discriminatory capability of the LPCC-HOG features over both MFCC-HOG and MFDWC-HOG features. The result obtained in this experiment 3 is apparently promising for developing the voters' authentication module of the SMIV architecture.

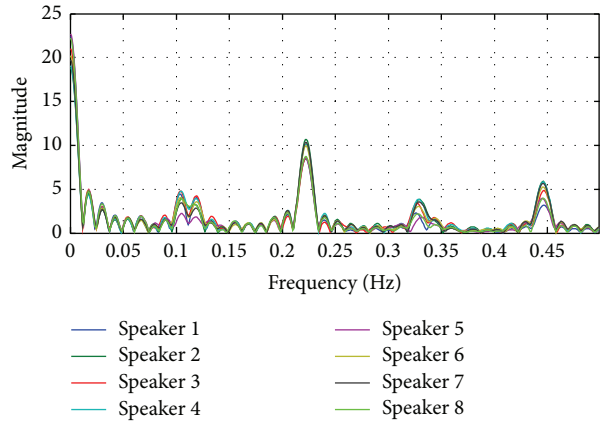

FIGURE 25: Frequency domain plot of the LPCC-HOG features for the utterance "*Hello Hello Hello* ..." of the first 8 Speakers.

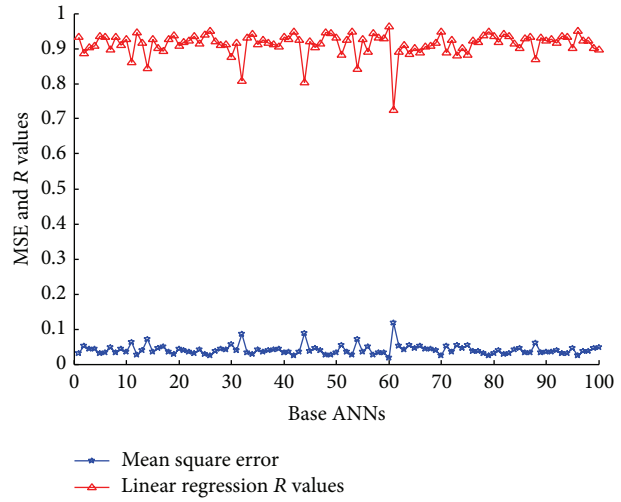

FIGURE 26: MSE and *R* values of the 100 Base ANNs for experiment 3.

5.1.4. Experiment 4. The fourth experimental model reported in this paper is based on the Spectral Histogram of Oriented Gradient (SHOG) features that were first reported in the speech processing research community by Selvan and Rajesh in [23] as efficient features for classification of Tamil language's male/female Speakers. Selvan and Rajesh [23] utilized the HOG algorithm to generate spectral features rather than for dimension reduction of the short-term spectral features (MFCC, MFDWC, and LPCC) as used in the three earlier experiments in this current study. However, the departure being pursued from the study by Selvan and Rajesh [23] in this experiment 4 is to examine the efficacy of SHOG features for Speaker recognition purpose rather than for speech based classification of persons into male or female gender. The architecture of the model for experiment 4 is derived from Figure 10 by using SHOG algorithm in the feature extraction and dimensionality reduction block. The computational components of the SHOG algorithm are also shown in Figure 27.

The spectrograms for the "*Hello Hello Hello* ..." utterance by the first 8 Speakers have been shown earlier in Figures 6 and 7. The computational components shown in the SHOG block diagram in Figure 27 were implemented in this study using appropriate functions in Image and Signal Processing Toolboxes of MATLAB R2012a. The time and the frequency domain plots of the 81-element SHOG features obtained as outputs from Figure 27 are shown in Figures 28 and 29. The SHOG features are unique for each Speaker as shown in time and frequency domain plots. These features are utilized to train the ANN ensemble with the same configuration as was used in experiments 1, 2, and 3. The results that were obtained (MSEs and *R* values) for each of the 100 base ANNs in this fourth experiment are illustrated with the plot in Figure 30.

Figure 30 shows the plot of the statistical results obtained from the training of the ANN ensemble with SHOG features in the current experiment. The average MSE is 0.0941 and the average *R* value is 0.7775. These results indicate that using SHOG features gave a far poorer performance than MFCC-HOG, MFDWC-HOG, and LPCC-HOG features in experiments 1, 2, and 3, respectively. In order to further test the performance of the SHOG features and ANN ensemble for Speaker recognition, the test samples utilized in the previous experiments are also used to test the model in the current experiment. Out of the forty test samples (i.e., two samples per Speaker), only ten samples were correctly recognized. Overall, the summary of the results obtained from all the experimental models in this study is shown in Table 4.

As shown in Table 4, the LPCC-HOG features with ANN ensemble machine learning model gave the best performance out of the four different models that were investigated in this study. On this basis, the LPCC-HOG features and ANN ensemble are nominated for the voters' authentication module of the SMIV system architecture. The result obtained in this study is in concordance with the position of the authors in [58] who recommended LPCC as one of the best short-term spectral features for practical applications. However, an important contribution of our work to the speech processing literature is the use of the HOG algorithm for dimension reduction of short-term spectral features. This contribution is significant because it serves as a consolidation of the earlier efforts by Selvan and Rajesh [23] in 2012, who developed the SHOG for classification of Speakers into different genders.

6. Conclusion

The SMIV architecture reported in this paper provides a paradigm shift for the implementation of e-voting systems by leveraging on the ubiquitous Internet, the pervasive mobile devices, GPS location services, NFC technology, and voice biometric authentication. This new architecture fulfills eight of the e-voting security requirements in creative ways and will potentially enhance conventional approaches to electoral processes. Another very important achievement of this current study is the discovery of LPCC-HOG as viable and compact short-term spectral features for implementing the authentication module of the SMIV architecture. These features are also very promising for other applications that

Table 4: Summary of the experimental results.

Experimental model	Extracted features	Average MSE	Average R	Number of correct predictions (total samples = 40)
1	MFCC-HOG	0.0411	0.9123	29
2	MFDWC-HOG	0.0455	0.9028	29
3	LPCC-HOG	0.0407	0.9127	30
4	SHOG	0.0941	0.7775	10

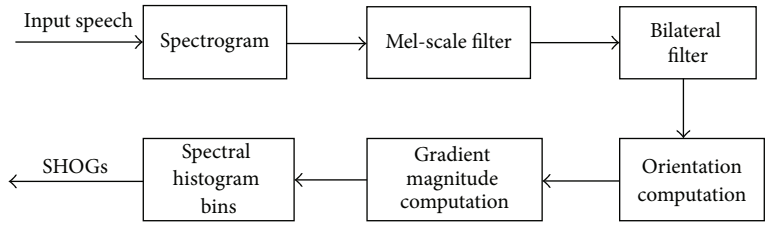

Figure 27: Computational components of SHOG [23].

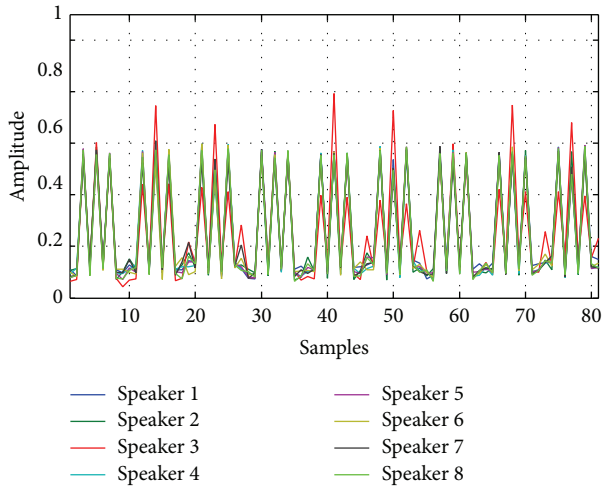

Figure 28: Time domain plot of the SHOG features for the utterance "*Hello Hello Hello* ..." of the first 8 Speakers.

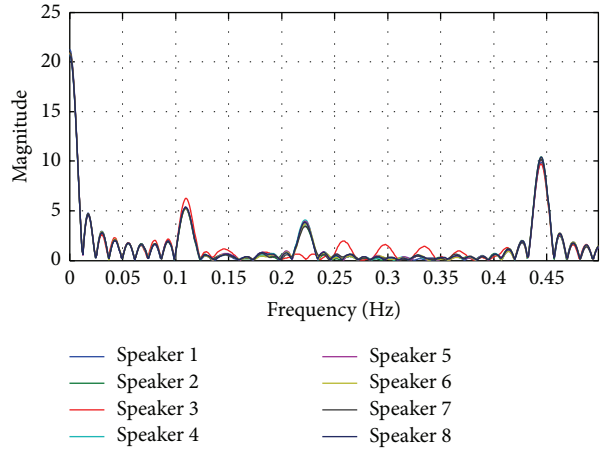

Figure 29: Time domain plot of the SHOG features for the utterance "*Hello Hello Hello* ..." of the first 8 Speakers.

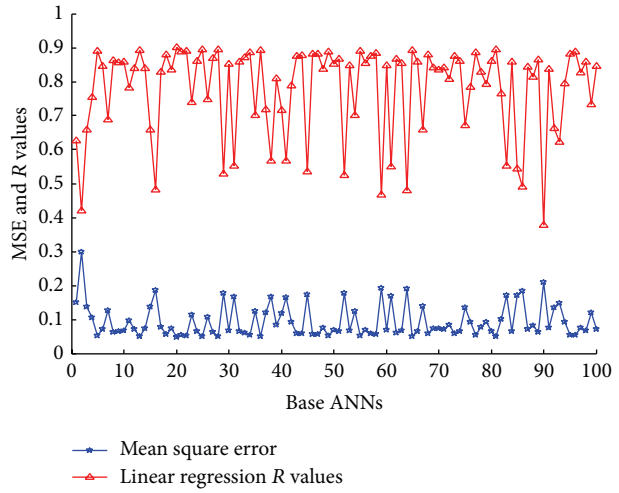

Figure 30: MSE and R values of the 100 Base ANNs for experiment 4.

require voice biometric based users' authentication module. However, we hope to further improve on the current result by adding other speech signals in more languages, record the speech signals over mobile phone lines, and experiment with other short-term features like the Line Spectral Frequencies (LSF) and Perceptual Linear Prediction (PLP). We also hope to experiment with other pattern matching models like Hidden Markov Model (HMM), Support Vector Machine (SVM), Radial Basis Function Neural Network (RBF-NN), and Deep Neural Network (DNN). This will hopefully help to further enhance the robustness of the authentication module of our proposed SMIV architecture.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

E. Adetiba is on Postdoctoral Fellowship at the ICT and Society (ICTAS) Research Group, Durban University of Technology, South Africa, funded by the Durban University of Technology Research and Development Directorate. He is on Postdoctoral Research Leave from the Department of Electrical and Information Engineering, College of Engineering, Covenant University, Ota, Ogun State, Nigeria.

References

[1] S. Yadav and A. K. Singh, "A biometric traits based authentication system for indian voting system," *International Journal of Computer Applications*, vol. 65, no. 15, pp. 28–32, 2013.

[2] R. K. Sinclair, *Democracy and Participation in Athens*, Cambridge University Press, 1991.

[3] P. J. Rhodes, *Athenian Democracy*, Oxford University Press, Oxford, UK, 2004.

[4] R. G. Saltman, *The History and Politics of Voting Technology: In Quest of Integrity and Public Confidence*, Palgrave Macmillan, 2006.

[5] D. W. Jones, *A Brief Illustrated History of Voting*, Department of Computer Science, University of Iowa, Iowa City, Iowa, USA, 2003, http://www.cs.uiowa.edu/~jones/voting/pictures/.

[6] Temple of Democracy, "Rural development and Panchayat raj department Policy note," 2012-2013, http://www.tnrd.gov.in/policynotes/ENGLISH%20POLICY%20NOTE%202014-15.pdf.

[7] S. D. Albright, *The American Ballot*, The American Council on Public Affairs, Washington, DC, USA, 1942.

[8] S. Thakur and R. Boateng, "Evoting for good governance and a green world," in *Proceedings of the Africa Digital Week*, pp. 55–62, African Institute of Development Informatics and Policy, Accra, Ghana, July 2011.

[9] M. I. Shamos and A. Yasinsac, "Realities of E-voting security," *IEEE Security & Privacy*, vol. 10, no. 5, pp. 16–17, 2012.

[10] R. Mercuri, "A better ballot box?" *IEEE Spectrum*, vol. 39, no. 10, pp. 46–50, 2002.

[11] K. Vollan, "Voting in uncontrolled environment and the secrecy of the vote," in *Proceedings of the 2nd International Workshop Conference: Electronic Voting*, pp. 155–169, 2006.

[12] A. M. Oostveen, "Outsourcing democracy: losing control of e-voting in the netherlands," *Policy & Internet*, vol. 2, no. 4, pp. 196–215, 2010.

[13] A. Khelifi, Y. Grisi, D. Soufi, D. Mohanad, and P. V. S. Shastry, "M-Vote: a reliable and highly secure mobile voting system," in *Proceedings of the Palestinian International Conference on Information and Communication Technology (PICICT '13)*, pp. 90–98, April 2013.

[14] R. Krimmer, S. Triessnig, and M. Volkamer, "The development of remote E-voting around the world: a review of roads and directions," in *E-Voting and Identity*, vol. 4896 of *Lecture Notes in Computer Science*, pp. 1–15, Springer, Berlin, Germany, 2007.

[15] R. M. Alvarez and T. E. Hall, *Electronic Elections: The Perils and Promises of Digital Democracy*, Princeton University Press, Princeton, NJ, USA, 2010.

[16] J. B. iEsteve, B. Goldsmith, and J. Turner, "International experience with e-voting," International Foundation for Electoral Systems, 2012.

[17] R. Allen, "Implementing electronic voting in the UK," UK Government Report, 2010, http://www.communities.gov.uk/corporate/.

[18] M. Reid, "E-Voting: a new electronic way to vote," *The New Brunswick Beacon*, 2010, http://www.newbrunswickbeacon.ca/2010/09/e-voting-a-new-electronic-way-to-vote/6687.

[19] L. M. Gallant, G. Boone, and C. S. LaRoche, "Mobile usability: state of the art and implications," in *Interdisciplinary Mobile Media and Communications: Social, Political, and Economic Implications*, pp. 345–346, IGI Global, 2014.

[20] L. F. Cranor and R. K. Cytron, "Sensus: a security-conscious electronic polling system for the internet," in *Proceedings of the 30th Hawaii International Conference on System Sciences*, vol. 3, pp. 561–570, IEEE, Wailea, Hawaii, USA, January 1997.

[21] F. Baiardi, A. Falleni, R. Granchi, F. Martinelli, M. Petrocchi, and A. Vaccarelli, "SEAS, a secure e-voting protocol: design and implementation," *Computers & Security*, vol. 24, no. 8, pp. 642–652, 2005.

[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, pp. 886–893, June 2005.

[23] A. M. Selvan and R. Rajesh, "Spectral histogram of oriented gradients (SHOGs) for Tamil language male/female speaker classification," *International Journal of Speech Technology*, vol. 15, no. 2, pp. 259–264, 2012.

[24] L. Kyrillidis, S. Cobourne, K. Mayes, S. Dong, and K. Markantonakis, "Distributed e-voting using the smart card web server," in *Proceedings of the 7th International Conference on Risks and Security of Internet and Systems (CRiSIS '12)*, pp. 1–8, October 2012.

[25] Q. Yuan-Yuan, Y. Jie, and L. Zhen-Ming, "Structural analysis of complex networks from the mobile internet," in *Proceedings of the National Doctoral Academic Forum on Information and Communications Technology*, pp. 1–7, August 2013.

[26] S. Juan and T. Shoulian, "Operator's mobile internet strategy in the process of converged network," in *Proceedings of the International Conference on Management and Service Science (MASS '10)*, pp. 1–4, August 2010.

[27] B. Sanou, "The World in 2014," ICT Facts and Figures, 2014, http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2014-e.pdf.

[28] K. McGrane, "The rise of the mobile-only user," *Harvard Business Review*, 2013, https://hbr.org/2013/05/the-rise-of-the-mobile-only-us/.

[29] Statistics South Africa, *General House Survey: 2013, Telecommunications. PO381*, Statistics South Africa, 2013.

[30] M. Souppaya and K. Scarfone, "Guidelines for managing the security of mobile devices in the enterprise," NIST Special Publication 800, 2013.

[31] T. Kinnunen, E. Karpov, and P. Fränti, "Real-time speaker identification and verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 277–288, 2006.

[32] F. Han, J. Hu, and R. Kotagiri, "Biometric authentication for mobile computing applications," in *Advanced Topics in Biometrics*, H. Li, Ed., pp. 461–481, World Scientific Publishing, River Edge, NJ, USA, 2012.

[33] S. Thakur, O. O. Olugbara, R. Millham, H. W. Wesso, M. Sharif, and P. Singh, "Transforming voting paradigm—the shift from inline through online to mobile voting," in *Proceedings of the 6th IEEE International Conference on Adaptive Science & Technology*

(ICAST '14), pp. 1–7, Covenant University, Ota, Nigeria, October 2014.

[34] S. Mauw, J. Verschuren, and E. P. de Vink, "Data anonymity in the FOO voting scheme," *Electronic Notes in Theoretical Computer Science*, vol. 168, pp. 5–28, 2007.

[35] R. Anane, R. Freeland, and G. Theodoropoulos, "E-voting requirements and implementation," in *Proceedings of the 9th IEEE International Conference on E-Commerce Technology and the 4th IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services (CEC/EEE '07)*, pp. 382–392, July 2007.

[36] G. Z. Qadah and R. Taha, "Electronic voting systems: requirements, design, and implementation," *Computer Standards & Interfaces*, vol. 29, no. 3, pp. 376–386, 2007.

[37] M. F. Mursi, G. M. Assassa, A. Abdelhafez, and K. M. Abo, "On the development of electronic voting: a survey," *International Journal of Computer Applications*, vol. 61, no. 16, pp. 1–11, 2013.

[38] A. Fujioka, T. Okamoto, and K. Ohta, "A practical secret voting scheme for large scale elections," in *Advances in Cryptology—AUSCRYPT '92: Workshop on the Theory and Application of Cryptographic Techniques Gold Coast, Queensland, Australia, December 13-16, 1992 Proceedings*, vol. 718 of *Lecture Notes in Computer Science*, pp. 244–251, Springer, Berlin, Germany, 1993.

[39] K. Visvalingam and R. M. Chandrasekaran, "Secured electronic voting protocol using biometric authentication," *Advances in Internet of Things*, vol. 1, no. 2, pp. 38–50, 2011.

[40] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4–20, 2004.

[41] J. L. Wayman, "Fundamentals of biometric authentication technologies," *International Journal of Image and Graphics*, vol. 1, no. 1, pp. 93–113, 2001.

[42] R. Krimmer and R. Schuster, "The e-voting readiness index," in *Proceedings of the 3rd International Conference on Electronic Voting*, pp. 127–136, Bregenz, Austria, August 2008.

[43] P. J. Phillips, A. Martin, C. L. Wilson, and M. Przybocki, "An introduction evaluating biometric systems," *Computer*, vol. 33, no. 2, pp. 56–63, 2000.

[44] D. Maio, D. Maltoni, J. Wayman, and A. Jain, "FVC2000: fingerprint verification competition 2000," in *Proceedings of the 15th International Conference on Pattern Recognition*, Barcelona, Spain, September 2000, http://www.csr.unibo.it/research/biolab/.

[45] T. Mansfield, G. Kelly, D. Chandler, and J. Kane, *Biometric Product Testing Final Report*, National Physical Laboratory, London, UK, 2001, http://us.allegion.com/IRSTDocs/DataSheet/110540.pdf.

[46] D. Blackburn, M. Bone, P. Grother, and J. Phillips, *Facial Recognition Vendor Test 2000: Evaluation Report*, 2001, http://www.dodcounterdrug.com/facialrecognition/FRVT2000/documents.htm.

[47] T. S. Gaafar, H. M. Abo Bakr, and M. I. Abdalla, "An improved method for speech/speaker recognition," in *Proceedings of the International Conference on Informatics, Electronics & Vision (ICIEV '14)*, pp. 1–5, May 2014.

[48] P. Rose, *Forensic Speaker Identification*, Taylor & Francis Forensic Science Series, Taylor & Francis, New York, NY, USA, 2002.

[49] J. J. Wolf, "Efficient acoustic parameters for speaker recognitio," *The Journal of the Acoustical Society of America*, vol. 51, no. 6, part 2, pp. 2044–2056, 1972.

[50] A. A. Malode and S. L. Sahare, "An improved speaker recognition by using VQ & HMM," in *Proceedings of the 3rd IET Chennai International on Sustainable Energy and Intelligent Systems (SEISCON '12)*, pp. 1–7, VCTW, Tiruchengode, India, December 2012.

[51] J. Deller, J. Hansen, and J. Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, New York, NY, USA, 2nd edition, 2000.

[52] X. Huang, A. Acero, and H. W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice-Hall, Upper Saddle River, NJ, USA, 2001.

[53] J. Bai, P. Xue, X. Zhang, and L. Yang, "Anti-noise speech recognition system based on improved MFCC features and wavelet kernel SVM," *Advances in Information Sciences and Service Sciences*, vol. 4, no. 23, pp. 599–607, 2012.

[54] J. Harrington and S. Cassidy, *Techniques in Speech Acoustics*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999.

[55] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1993.

[56] E. Adetiba and O. O. Olugbara, "Lung cancer prediction using neural network ensemble with histogram of oriented gradient genomic features," *The Scientific World Journal*, vol. 2015, Article ID 786013, 17 pages, 2015.

[57] M. Kumar, "Digital image processing," in *Satellite Remote Sensing and GIS Applications in Agricultural Meteorology, Proceedings of the Training Workshop, 7-11 July, 2003, Dehra Dun, India*, pp. 81–102, 2003.

[58] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[59] R. Bellman, *Adaptive Control Processes: A Guided Tour*, Princeton University Press, Princeton, NJ, USA, 1961.

[60] T. Kobayashi, A. Hidaka, and T. Kurita, "Selection of histograms of oriented gradients features for pedestrian detection," in *Proceedings of the 14th International Conference on Neural Information Processing (ICONIP '07), Kitakyushu, Japan, November 2007, Part II*, vol. 4985 of *Lecture Notes in Computer Science*, pp. 598–607, Springer, 2008.

[61] D. Das, "Activity recognition using histogram of oriented gradient pattern history," *International Journal of Computer Science, Engineering & Information Technology*, vol. 4, no. 4, pp. 23–31, 2014.

[62] E. Adetiba and F. A. Ibikunle, "Ensembling of EGFR mutations' based artificial neural networks for improved diagnosis of non-small cell lung cancer," *International Journal of Computer Applications*, vol. 20, no. 7, pp. 39–47, 2011.

[63] B. Parmanto, P. W. Munro, and H. R. Doyle, "Reducing variance of committee prediction with resampling techniques," *Connection Science*, vol. 8, no. 3-4, pp. 405–425, 1996.

[64] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 42, no. 4, pp. 463–484, 2012.

[65] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[66] B. Wildermoth and K. K. Paliwal, "Use of voicing and pitch information for speaker recognition," in *Proceedings of 8th Australian International Conference Speech Science and Technology*, pp. 324–328, Canberra, Australia, 2000.

[67] Z. Lei, Y. Yang, and Z. Wu, "Ensemble of support vector machine for text-independent speaker recognition," *International Journal of Computer Science, Network and Security*, vol. 6, no. 5, pp. 163–167, 2006.

[68] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals, and Systems*, vol. 2, no. 4, pp. 303–314, 1989.

[69] P. Marius-Constantin, V. E. Balas, L. Perescu-Popescu, and N. Mastorakis, "Multilayer perceptron and neural networks," *WSEAS Transactions on Circuits and Systems*, vol. 8, no. 7, pp. 579–588, 2009.

[70] K. J. Cherkauer, "Human expert-level performance on a scientific image analysis tasks by a system using combined artificial neural networks," in *Working Notes of the AAAI Workshop on Integrating Multiple Learned Models*, P. Chan, Ed., pp. 15–21, AAAI Press, Menlo Park, Calif, USA, 1996.

[71] M. Faúndez-Zanuy and D. Rodríguez-Porcheron, "Speaker recognition using residual signal of linear and nonlinear prediction models," in *Proceeding of the International Conference on Spoken Language Processing, (ICSLP '98)*, vol. 2, pp. 121–124, 1998.