Development of a Text-Dependent Speaker Recognition System

Aliyu E. O. Adewale O. S. Adetunmbi A.O.

Computer Science Department
Federal University of Technology, Akure, Nigeria

ABSTRACT
Speaker recognition is the ability of recognizing a person based on his voice. Many modalities and techniques have been applied to achieve the task of authentication, ranging from retina scans to finger prints. This paper presents a Text-Dependent model for automatic Speaker Recognition. Features extractor are based on Mel-Frequency Cepstrum Coefficients (MFCCs) and Linear Predictive Coding (LPC) analysis technique which aids the translation of incoming speech into a feature value. Also, a recognizer block, employs the two techniques (MFCC and LPC) to get an hybrid features for speaker identification/verification system. An experiment was carried out on the threshold from 75% to 95% at 5% differential to know the performance of the MFCCs and LPC identification/verification system. From the experiment carried out, the result shows that LPC outperforms MFCCs and a combination. The text-dependent speaker recognition system was implemented using Java Programming Language (Java Speech Application Programming Interface (JSAPI)). The system is developed for access control into computer systems and could be used for access control where security is considered to be of utmost important.

KEYWORDS: Text-Dependent,Speech Recognition, Mel-Frequence Cepstrum Coefficients (MFCCs), Linear Predictive Coding (LPC)

1. INTRODUCTION
The increasing emphasis being placed on security has led to a focus on the development of biometrics authentication technologies. Biometrics are the physical and behavioural traits that belong to an individual.
With the rapid development of automated applications for finances and E-commerce and in the context of the evolving internet and wireless communication technology, the importance of reliable, high-security and non-intrusive methods for personal authentication has been growing significantly. Today, the quality of these methods plays an essential role for the acceptability and ease of use of the target applications. Many modalities and techniques have been applied to achieve the task of authentication, ranging from retina scans to finger prints. In this paper, a particular modality is of interest - the voice. This modality has a unique advantage over other biometrics by relying on speech, the primary vector of communication and is especially important in applications such as telephony dialog systems where it is a natural and, besides the touchtone keypad, also the only communication means. By extracting appropriate features from a person's voice the uniqueness of the physiology of the vocal tract and the articulatory properties can be captured to a high degree and can serve the purpose of authentication (Stephane *et al*, 2001).
(Qin,2007) explained that Speaker recognition technology analyzing and modeling the voiceprints has been a major research effort for the past decades, but there are still many problems to be solved. One problem is to understand what characteristics in the speech signal convey the representation of speakers. This relates to understanding how humans listen to the speech signal and recognize the speaker. The other problem is to make automatic speaker recognition systems robust under different conditions.
In recent years, there has been a booming interest in the use of biometric characteristics as a means of recognizing and identifying a person. Human voice is one of the most important biometric identifiers of a person. Speaker recognition allows a device to determine who a speaker is, as distinct from speech recognition which determine what a speaker is saying [Ali et al., 2006] There are two types of speaker recognition systems : **Text-dependent** meaning the text must be the same for enrolment and recognition which improve performance especially with cooperative users. **Text independent** system has no advance knowledge of the presenter's phrasing and is much more flexible in situations where speaker is not cooperative. There are two main applications of speaker recognition : *Identification* is the process of determining which registered speaker provides a given utterance while V*erification* is the process of accepting or rejecting the identity claim of a speaker [Sunil et al., 2010].
In this paper, a concept of combining two techniques, the MFCCs and LPC model is presented to get an hybrid features for speaker identification/verification system, implement the model developed and experimental result are presented. We show that the text-dependent speaker recognition system is speech dependent which enables communication between cooperative users to enhance security.

2. RELATED WORKS
In Ali *et al*., (2006) "Speaker Recognition", was motivated to derive a more reliable method of distinguishing between two different types of speeches or voices and to develop an algorithm to recognize and identify different input voices. The type of speaker verification used is text-independent in which the voice sample of the speakers were collected with standard computer microphone and carried out the following analyses: frequency analysis, power spectral density analysis and energy analysis but the method of feature extraction was not explained. The program was coded using the capabilities of Matlab that record a persons voiceprint, certain threshold was set that determine whether the aspect of the voiceprint is male or female. The accuracy of the system is 79.85% with the shortcomings includes too much processing power if one

records for too long, the program gives a divide by zero error if the amplitude of the sound is overly high and it is highly CPU intensive.

Ahsanul et al., [2007] presented a "Vector Quantization In Text – Dependent Automatic Speaker Recognition Using Mel-Frequency Cepstrum Coefficient (MFCC)". MFCCs was used for feature extraction and Vector Quantization for feature matching using VQ-LBG algorithm for implementing the vector quantization. However, it is very important to develop a method to cope with the problem of distortion due to telephone sets and background noise, the speaking rate and robust against variations in voice quality due to causes such as voice disguise or colds.

Furthermore, Sunil et al., [2010] presented "Prosodic Feature Based Text-Dependent Speaker Recognition Using Machine Learning Algorithms". This paper explained early methods of speaker recognition consisted of classifiers like Vector Quantization, Dynamic Time Warping, Hidden Markov Models, Gaussian Mixture Models and identified the drawback of each methods. Based on the lapses found in the previous methods, the paper proposed the use of machine learning classifier to overcome the drawbacks of traditional classifiers by investigating and comparing the performance of four machine learning (ML) algorithms namely MLP, RBFN, C4.5 and BayesNet. It found that RBFN algorithms outperform all other ML algorithms with an improvement in classification accuracy as the number of speaker increases. Moreover, in order to further decrease the misclassification rate and improve the performance rate, emphasis can be laid on including more number of features using hybrid features set.

Rashidul et al., [2004] presented "Security System Based on Speaker Identification". Mel-frequency cepstral coefficients (MFCCs) were used for feature extraction and vector quantization technique was used to minimize the amount of data to be handled. The system was implemented in Matlab 6.1 on window XP platform. The speech database consists of 21speakers, which includes 13 male and 8 female speakers in which the identification rate is defined as the ratio of the speakers identified to the total number of speakers tested. In their conclusion, they suggested that HMM may be used to improve the efficiency and precision of the segmentation to deal with crosstalk, laughter and uncharacteristic speech sounds. Also, a combination of features (MFCC, LPC, LPCC etc) may be used to implement a robust parametric representation for speaker identification.

3. The Proposed Model For Speaker Recognizer System

A schematic diagram of automatic speaker verification/identification system is presented in Figure 3.1 This research work adopted Linear Prediction Coding (LPC) and Mel Frequency Cepstrum Coefficients (MFCCs) techniques. The speaker recognizer system comprises of two distinct blocks, a feature extractor to create a template for the user's and a recognizer uses both techniques for speaker identification/verification.

3.1 Speech Signal Acquisition

The first step for achieving voice recognition is to capture the sound signal of the voice. A standard microphone coupled with Java Speech Application Programming Interface (JSAPI 2.0) was used for capturing the voice signal in a room. An analog-to-digital converter (ADC) available in JSAPI digitizes the amplitude (volume) of each sampled element onto a digital scale. From that sample, a digital representation of the voice, called a voiceprint is created. The encoding format for the voice sample used for feature extraction and identification/verification are :

a) A 16bits per sample rate: The rate at which voice sample values obtained are measured.
b) 44.1kHz sampling rate: The frequency at which the voice sample must be output in order to produce the original speed and pitch.
c) Stereo : To make the recorded sound seem more natural when reproduced.
d) Big Endian : The way the bytes are stored in the memory, big endian means that the high-order byte of a number is stored in memory at the lowest address, and the low-order byte at the highest address.
e) Linear PCM (Pulse Code Modulation): Is a method used to digitally represent sampled analog signals. It is a standard form for digital audio in computer.

3.2 Mel–Frequency Cepstrum Coefficients Algorithm (MFCCs)

MFCCs are based on the known variation of the human ear's critical bandwidths with frequency [Rashidul, 2004]. Speech is a signal produced as a result of several transformations occurring at several different levels: semantic, linguistic, articulatory, and acoustic. Differences in these transformations appear as differences in the acoustic properties of the speech signal. In Java, sampled sounds are stored as a series of bytes. This series is also stored as an array, where an array represents a series of values at specific indexes denoted by brackets. Java converts an analogue speech signal into an array of byte which can be stored as a wave in aiff or au file format. aiff or au file format are audio file format like wave, mp3, midi and so on. Here, the voiceprint is saved in wave file format because it is easier to work with this file. MFFC analyzes the speech signal by determine the amplitude of different frequencies retrieve from the voices of different persons.
Applying Discrete Fourier Transform (DFT) to convert digital signals from the time domain into the frequency domain. This paper employed the DFT in Aritra [2011] defined as:

$$f_j = \sum_{k=0}^{n-1} x_k e^{-\frac{2\pi i}{n}jk} = \sum_{k=0}^{n-1} x_k \left(\cos\left(\frac{2\pi}{n}jk\right) - i\sin\left(\frac{2\pi}{n}jk\right)\right) \quad (3.1)$$

The Filterbank is defined as

$$X(m) = \ln \sum_{k=0}^{N-1} |f_j|\, H(k,m) \quad (3.2)$$

Obtaining the centre frequency in Hertz is given by

$$F_c(m) = 700(\log_{10}\emptyset_c(m)^{2595} - 1) \quad (3.3)$$

To save the MFCCs value using Discrete Cosine Transform (DCT) by computing C(l) as given by

$$C(l) = \sum_{m=1}^{M} X(m).\cos\left(l\frac{\pi}{M}(m-\frac{1}{2})\right) \quad (3.4)$$

3.3 Linear Predictive Coding (LPC)

The resonance of some of the signal frequencies produced by the vocal tract results in sharp frequency peak is the voice signal. These sharp frequency peaks are pitchmarks. The pitchmarks are formed when there is a resonance with the frequency of the vocal tract with the signal produced by the source. The positions of pitchmarks are determined by using the following methods adopted from Sachin and Pallavi [2005]: The positions of the pitchmarks are calculated for all target values (p(n)), using $p(n) = p(n-1) + \frac{1}{lf0+(m*p(n-1))}$ (3.5)

where

$$m = \frac{(f0-lf0)}{position} \quad (3.5a)$$

In the above equation, 'position' is the time-position of target value 'f0' and 'lf0' is the pitch value for the low pass filter. Low-pass filters are used to implement anti-alias that limit the bandwidth of the signal to approximate the Nyquist rate before sampling. In this paper, lf0 represent lower frequency in a voiceprint in pitchmark generation which eliminate short term flunctuation frequency and leaving the long term to obtain a smoother form of signal frequencies.

The LPC coefficients are calculated by the values of the formant at the pitchmarks, at every channel by using the relation below:

$$LPC_{coefficient} = (\text{formant value} * LPC_{range}) + LPC_{minimum} \quad (3.6)$$

LPC_{range} and $LPC_{minimum}$ are determined while encoding the file (voiceprint), voiceprint is the digital representation of the voice. The formant value are the pitchmark calculated for the sharp frequency peaks as follows :

$$s'(n) = \sum_{k=1}^{p} a_k * s(n-k) \quad (3,7)$$

3.4 Speaker Recognition Based On LPC and MFCCs Techniques

In the recognition process, the speakers are required to provide sample speech for the purpose of identification/verification. Extraction of features using MFCC and LPC are carried out on the sample voice to determine the speaker. The threshold for recognition is varied and set at 75%, 80%, 85%, 90% and 95%. The features extracted using MFCC and LPC are compared with the presented feature sample to the stored data for identification/verification as shown in equation 3.8 and 3.9. Then, the output is later combined to give an hybrid value for speaker recognition.

$$MFCC_p = \frac{MFCC_n}{\sum_{i=1}^{N} x_i} * 100 \quad (3.8)$$

$$MFCC_n = \begin{cases} n = n+1 \ni MFCC_i = MFCC_j ; = i = j = 1, 2,\ldots, N \\ n = n + 0 \ni MFCC_i \neq MFCC_j ; i = j = 1, 2,\ldots, N \end{cases}$$

where $MFCC_i$ is the MFCC value i of frequency bank in enrollment, $MFCC_j$ is the MFCC value j of frequency bank in verification, $MFCC_n$ represent number of distinct value of $MFCC_i$ frequency bank in enrollment and when compared with $MFCC_j$ frequency bank in verification, N represent number of observation of frequency bank, and $MFCC_p$ represent the percentage value obtained for verification/identification which determines if access will be granted or denied.

The value of LPC_p for a particular speaker is given in equation 3.9

$$LPC_p = \frac{LPC_n}{\sum_{i=1}^{N} x_i} * 100\% \quad (3.9)$$

$$LPC_n = \begin{cases} L = L + 1 \ni Fi = Fj \quad i = j = 1, 2 \\ L = L + 0 \ni Fi = Fj \quad i = j = 1, 2,\ldots, N \end{cases}$$

where F_i is the LPC stress value found in the enrollment, F_j is the LPC stress value found in verification / identification, N is the number of observation stress position found in a voiceprint, L is the total match found, which becomes the value of LPC_n and LPC_p is the percentage value obtained for verification or identification which determine if access will be granted or denied.

3.5 Hybrid Feature Value

In order to reduce the misclassification rate, an hybrid value is obtained from the result of MFCC and LPC feature value. Both techniques are capable of identifying a speaker based on different voice features, therefore equal contribution (weight) from both technique will compensate for each other deficit. The hybrid value is determined by taking the average percent of the outcome of both techniques as expressed in equation 3.10. As a result of this, identification/verification system is being enhanced. Over reliance on one technique will make the system dependent on that technique and hence reduce accuracy. Where one technique seems to be inadequate due to the factors like noise or ill-health, the other factor would compensate for it to provide recognition accuracy.

$$H_n = (LPC_n + MFCC_n) * 0.5 \quad (3.10)$$

where H_n is the hybrid value for speaker and n represents distinct values for a speaker.

4. Experimental Setup and Results

The experiment was carried out with 20 persons, 9 males and 11 females. Their age range was between 15 to 28 years. The experiment with speaker identification and verification is divided into two major phases. The first phase is the **enrollment** which involves data collection (voice samples) from the speaker and processing the voice sample to obtain certain spectral and prosodic features using LPC and MFCC respectively. This process was quite slow; in order to quicken the process, the system was divided into two. Enrollment was done simultaneously for two persons. At the end of enrollment

for each person, the following information was recorded: surname, other name, gender, age, and an identification number (id) is assigned by the system for each of the persons used in the experiment. This information is shown below in table 4.1.

The second phase is the **identification and verification** phase. The verification of a person requires that the id assigned to the individual and the voice sample are submitted as input into the system and the identification determines which registered speaker provides a given utterance. The system performs feature extraction and compares the outcome with the values available in the system for that id and for extracted features for identification. The result of LPC, MFCC, and a combination are expressed in percentage.

4.1 Results of The Experiment

An experiment was carried out on the threshold from 75% to 95% at 5% differential to know the performance of the MFCC and LPC identification/verification system. Table 4.2 and 4.3 shows the percentage score obtained for each threshold mark for identification/verification system with their corresponding graph.

Identifying the speaker is based on which registered speaker provides a given utterance. Table 4.2 show that the higher the threshold for MFCC, LPC and hybrid, the less an access is granted. Verification system requires the speaker to supply his/her ID and voiceprint to the system for verification. Table 4.3 show the result obtained for MFCC varied for each mark limit which could be as a result of background noise of the computer system while the result of LPC was stable even more than the hybridization.

Identifying the speaker is based on which registered speaker provides a given utterance. Table 4.2 show that the higher the threshold for MFCC, LPC and hybrid, the less an access is granted. Verification system requires the speaker to supply his/her ID and voiceprint to the system for verification. Table 4.3 show the result obtained for MFCC varied for each mark limit which could be as a result of background noise of the computer system while the result of LPC was stable even more than the hybridization.

5. Conclusion and Recommendations

The study reveals that Linear Predictive Coding (LPC) outperforms Mel Frequency Cepstral Coefficients (MFCCs) where MFCCs denied a speaker due to the factor like noise, LPC compensates for it by identifying the speaker to provide powerful authentication solution. Also, the results show that LPC is more stable than MFCC and a combination. The results of the developed system are satisfactory though, it can be improved upon to ensure security of data and computerized system.

Moreover, future applications of voice biometrics will be text independent incorporating much higher level characteristics such as phonotactic, idiolectal, dialogic and semantic. Also, speaker recognition could be improved upon as communication and computing technology advance. The system is speech dependent which requires that a constant word or sentence is spoken by every person otherwise the voice input is rejected.

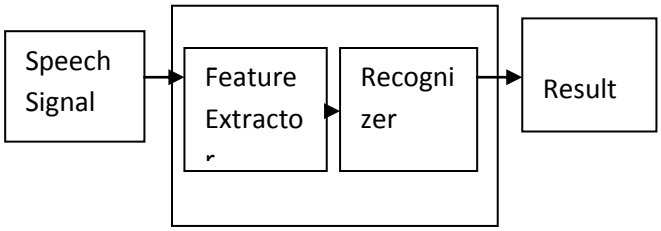

Figure 3.1 Block-Diagram of Proposed Speaker Recognizer

The line colours below are used to illustrate the graphs

MFCC graph

LPC graph

MFCC and LPC combine graph`

Table 4.1: Persons used for the experiment

ID	Surname	Other name	Gender	Age
51	Ajokwu	Chinyere	Female	15
52	Ejike	Arinze	Male	20
53	Sanni	Sanni	Male	25
54	Anichebe	Chioma	Female	18
55	Sunday	Martha	Female	17
56	Mayomi	Kehinde	Male	21
57	Samuel	Ndidiamaka	Female	17
58	Ganiyu	Mosurat	Female	23
59	Olisa	Chinanza	Female	17
60	Ifeanyi	Osinachi	Male	26
61	Ojo	Olubunmi	Male	28
62	Akaeze	Emeka	Male	18
63	Nweke	Ifeoma	Female	21
64	Ndubuze	Amarchi	Female	23
65	Ndubuze	Peace	Female	23
66	Ajokwu	Ugochukwu	Male	25
67	Rafiu	Saidi	Male	25
68	Alex	David	Male	15
69	Ibrahim	Fatimat	Female	23
70	Obi	Amarachi	Female	24

Table 4.2 : Access granted at different class limit for Identification

Mark Limit	Access Granted(MFCC)	Access Granted(LPC)	Access Granted(hybrid)
75%	18	20	20
80%	15	20	20
85%	8	20	16
90%	3	14	4
95%	1	6	0

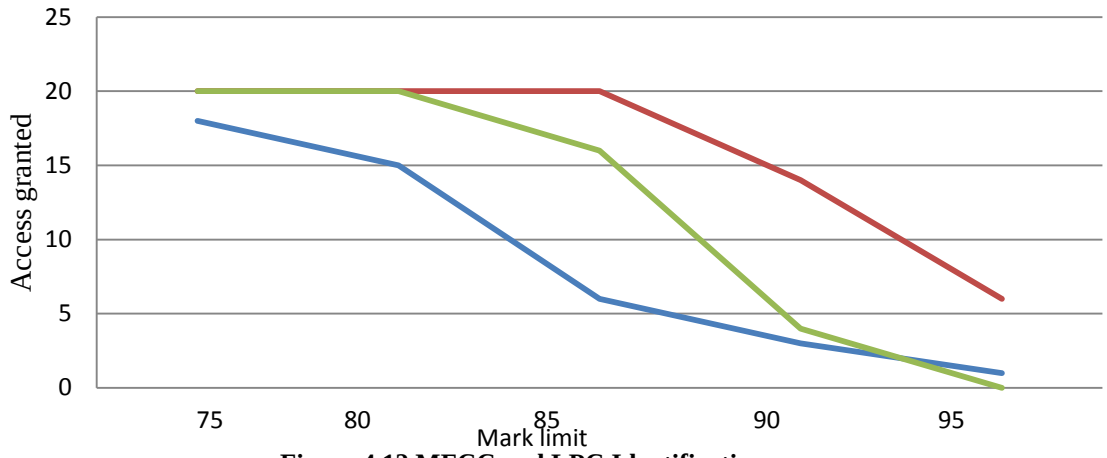

Figure 4.12 MFCC and LPC Identification

Table 4.3 : Access granted at different class limit for Verification

Mark Limit	Access Granted(MFCC)	Access Granted(LPC)	Access Granted(hybrid)
75%	18	20	20
80%	15	20	20
85%	9	20	16
90%	4	14	5
95%	2	6	0

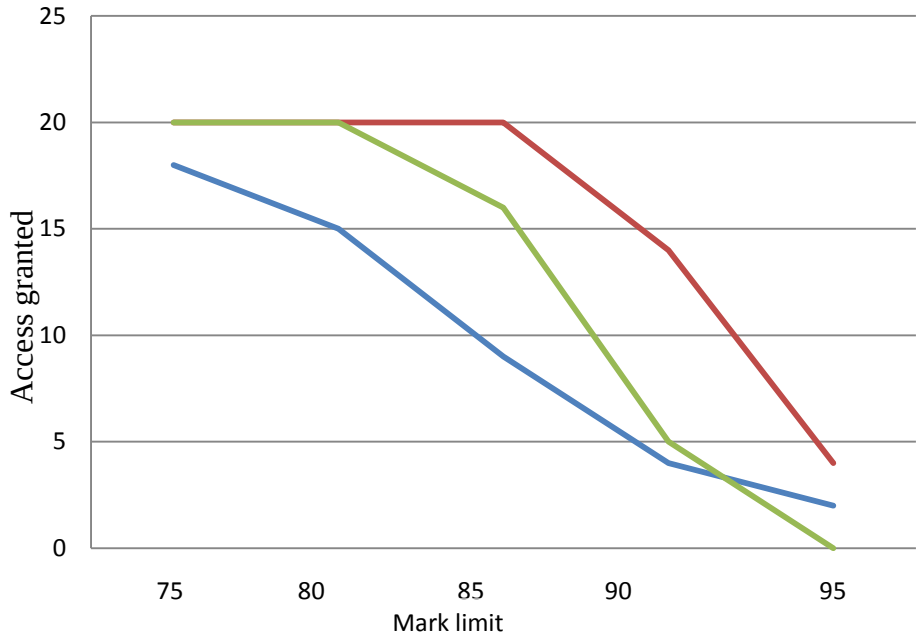

Figure 4.2 MFCC and LPC Verification

6. REFERENCES

[1] Ahsanul, K., Sheikh, M., and Masudul, A., (2007). " Vector Quantization in Text-Dependent Automatic Speaker Recognition Using Mel-Frequency Cepstrum Coefficient". Retrieved from

[2] Ali, H., Nawaz, I., Shivam, L., and Zheng, H., (2006). "Automatic Speaker Recognition".

[3] Aritra, A., (2011). "A Novel Approach to Design Automatic Speaker Verification Based Security System Using Neural Network". Retrieved from

[4] Asterios, T., and Margaritis, K. G., (2002). "Development of a Text-Dependent Speaker Identification System". Department of Applied Informatics University of Macedonia, pp 525 – 530

[5] Qin, J., (2007). "Robust Speaker Recognition".

[6] Rashidul, H., Mustafa, J., Golam, R., and Saifur, R., (2004). "Speaker Identification Using Mel-Frequency Cepstral Coefficients". 3rd International Conference on Electrical $ Computer Engineering ICECE 2004, pp : 28 – 30 December 2004, Dhaka, Bangladesh.

[7] Sachin, A., and Pallavi, A., (2005). "A Pragmatic Solution to An Indian Accented English Speech Synthesizer Using Residual Excited Linear Predictive Coded Voice". Indian Institute of Information Technology-Allahabad Jhalwa, Allahabad-Indian. Retrieved from

[8] Stephane, H., Jiri, N., and Upendra, V. (2001). "Conversational Speech Biometrics". IBM T. J. Watson Research Center Rt. 134, Yorktown Heights, NY, USA. Retrieved from

[9] Sunil, A., Shruti, A. K., and Rama, C. K., (2010). "Prosodic Feature Based Text Dependent Speaker Recognition Using Machine Learning Algorithms". International Journal of Engineering Science and Technology Vol. 2(10), 2010, pp : 5150 – 5157.