Voice Signal Processing for Personal Identification & Authentication

W.O. Adesanya[1], M. K. Adu[2] & G. O. Odulaja[3]
[1]Department of Computer Science, Federal College of Agriculture, Akure, Nigeria
[2]Department of Computer Science, Federal Polytechnic, Ado-Ekiti, Nigeria.
[3]Department of Computer Science, Tai Solarin University of Education, Ijagun, Ijebu-Ode. Nigeria.
E-mails: sanyalanre2003@yahoo.com, memokadu@yahoo.co.uk, goddyseyi@gmail.com

ABSTRACT

Voice-signal processing is a component of Speaker Recognition System that involves identification and verification or authentication of the speaker. At each stage, the voiceprint is compared with model voices of all speakers in the database. The comparison is a measure of the similarity (score) from which rejection or acceptance of the verified speaker is chosen. The Dynamic Time Warping (DTW) and Vector Quantization (VQ) models were employed to investigate processing time and memory requirements. The Linear Predictive Coding (LPC) and Cepstral analysis for feature extraction techniques were used for damping. The system was trained and tested using a population of ten users, with additional ten impostors. The DTW was found to be more suitable for real-time application with the real-time average speaker recognition time of 7.80 seconds. The system was able to make access decision in an average of 2.80 seconds after the voice sampling was completed. In general, our model compares favourably with literature with better recognition and access decision times.

Key words: Voice-signal, Real-time, Dynamic Time Warping, Vector Quantization, Linear Predictive Coding, Personal Identification and authentication

1. INTRODUCTION

Speaker Recognition System is one of the Biometrics that uses voice for individual recognition in which Voice-signal processing is a component. Accessing protected resources is always carried out through the use of personal tokens like a key or badge, knowledge of certain information like a password or combination of numbers [2]. A password is a string of characters used to login to a computer and other systems for files access, program access, and other resources. They are used to ensure that people do not access any system unless they are authourized to do so [1]. It is however observed that these passwords (or keys or badges) can be lost, stolen or counterfeited, thereby posing a threat to information or data security.

Thus, in order to reduce this security threat, this paper focuses on real-time voice-driven access to the restricted resources, since voice is unique to each person and cannot be lost or stolen. Voice-driven based solutions are able to provide for confidential financial transactions and personal data privacy. The remaining section is organized as follows: section 2 reviews a number of relevant literatures on speaker recognition system; section 3 describes the methodology for the proposed system while 4 and 5 describe the results and concludes the work respectively.

2. RELATED WORK

There have been numerous researches in the application of techniques and models used in extracting voice feature or matching feature in order to identify and verify speaker in speaker recognition system. A number of such relevant researches were reviewed in this paper. [7] identified that Verification system authenticates a person's identity by comparing the captured biometric characteristic with its own biometric template(s) pre-stored in the system which conducts one-to-one comparison to determine whether the identity claimed by the individual was true. A verification system either rejects or accepts the submitted claim of identity and that the identification system recognizes an individual by searching the entire template database for a match which conducts one-to-many comparisons to establish the identity of the individual. The delimitations of [6] were that the rate of fingerprint capture and feature extraction were not considered, although in a real-time world scenario, this is an important factor.

In [10], a stochastic model was developed to solve the problem of speech processing in speaker recognition. The research was able to develop a high-quality, multivariate and Hidden Markov Model (HMM) by means of Hidden Markov Toolkit (HTK) tool software to determine the speaker but provision for grammar testing enlargement as the new models are needed for the new words training. However, the limitations of the research were the direct counting of the probability was very complicated; and that the current state depends on the previous state. A new feature selection method for speaker recognition was proposed by [9] to keep the high quality speech frames for speaker modelling and to remove noisy and corrupted speech frames. The research adopted spectral subtraction algorithm to estimate the frame power.

An energy based frame selection algorithm was then applied to indicate the speech activity at the frame level. The research was able to use the eigenchannel based GMM-UBM speaker recognition system to evaluate the proposed method. However, the research required long-term spectral analysis and computation found to be complex. [18] concentrated on optimized speech processing in the DSP56001 hardware platform, especially in the application of noise reduction and speech enhancement. [12] worked on a hardware based speech recognition system.

Both work by [12, 18] were hardware based but were not concentrated in the area of speaker recognition, which is the focus of this paper, based on the observation that the size of the speaker database grows when the number of speakers in a system is increased. This poses two problems in terms of memory requirement for voice database storage, and processing time required by the system and these problems are being analyzed in this paper using a comparative analysis on Dynamic Time Warping and Vector Quantization based models to determine a suitable model with better response time in real-time application for voice-driven recognition system.

3. METHODOLOGY

A voice-driven system involves two phases. In the first phase, a user enrols by providing voice samples to the system. The system extracts speaker-specific information from the voice samples to build a voice model of the enrolling speaker, Figure 1.

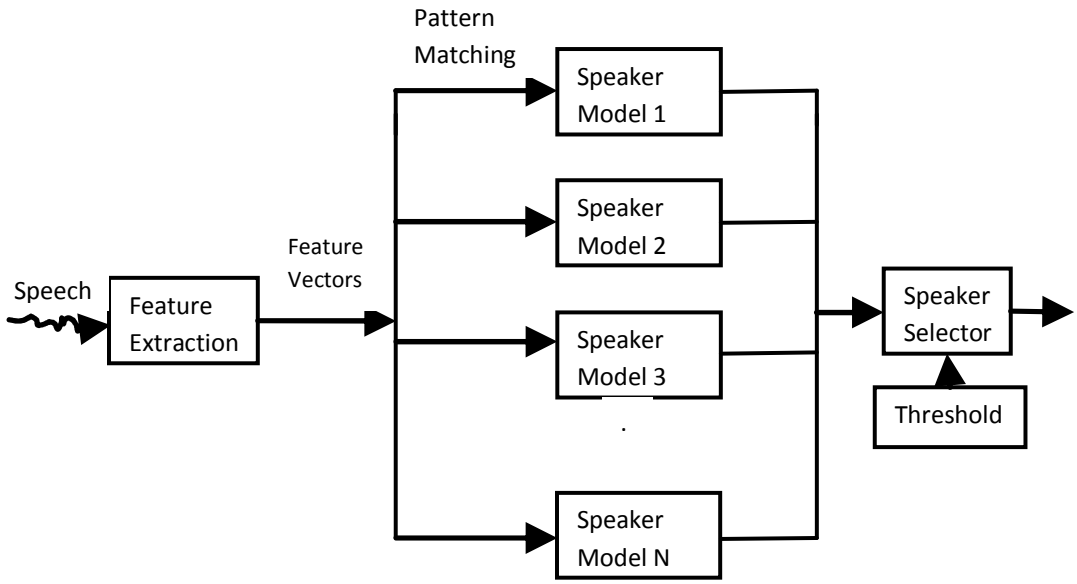

Figure 1: Model of a Voice Identification System

In the second phase, a user provides a voice sample (also referred to as test sample) that is used by the system to measure the similarity of the user's voice to the model(s) of the previously enrolled user(s) and, subsequently, to make a decision. In a speaker identification task, the system measures the similarity of the test sample to all stored voice models. In speaker verification task, Figure 2, the similarity is measured only to the model of the claimed identity.

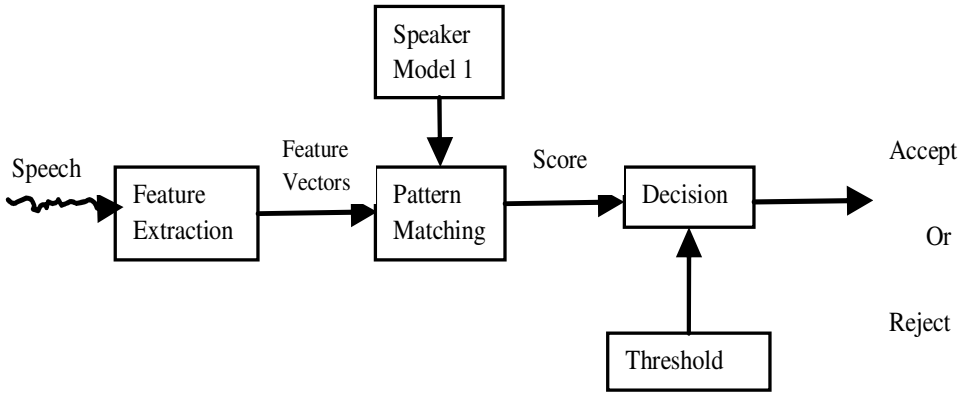

Figure 2: Model of a Voice Verification System

Several conversational telephone calls in English and Yoruba languages were conducted and recorded. The collected voices were processed through the use of notebook computer with an external microphone attached, where all the voices were recorded digitally into the computer via the microphone. Voice sampling was required to convert an analogue signal into a discrete signal, to be digitally processed by a digital computer. Further pre-processing such as speech framing, edge detection and windowing were performed to improve the raw digitized signal to be used in the feature extraction process, further steps taken as shown in Figure 3. A digital signal processor running at 50MHz was used to execute the voice recognition algorithm. The Linear Predictive Coding (LPC) Cepstral technique was used for feature extraction of speech signal as the speech sample $s(t)$ at time t was approximated as a linear combination of the past p samples

$$s(t) \approx a_1 s(t-1) + a_2 s(t-2) + \ldots + a_p s(t-p) \quad (1)$$

where the coefficients $a_1, a_2, \ldots, a_p$ were assumed constant over a single speech frame. The autocorrelation method with function

$$r_i(m) = \sum_{i=1}^{t} x_i(t) x(t+m), \quad m = 0, 1, 2, \ldots, p \quad (2)$$

was used for estimating the coefficients which provided the energy of the speech frame and was used for discarding silent frames. The LPC coefficients $a(t)$, $0 \leq t \geq$ T-1 were computed from the autocorrelation vector using a recursion method known as Durbin's method where the equations were solved recursively for i = 1,2,...,p. On completion of the algorithm, the final solution was given as:

$$a_m = LPC\ coefficients = a_m^{(p)}, \quad 1 \leq m \leq p \quad (3)$$

Vector quantization (VQ) codebook was used for feature matching, to efficiently represent speaker specific characteristics. One codebook was created for each i speaker during the training stage. During recognition, the total distance for the i-th speaker was computed by:

$$D^i = \sum_{i=1}^{L} \min_{1 \leq j \leq N} d(y_i, C_j^i) \quad (4)$$

where C_j^i is the j-th code vector of the i-th speaker's codebook, N_c is the codebook size, $y_1, y_2, \ldots, y_L$ represent the feature vector of the test utterance, D^i is the matching score and $d(y_i, C_j^i)$ the distance between the feature vector and the codebook vector, where the speaker identification decision was based on the matching score. The speaker model with the smallest matching score, D was accepted as the producer of the voice sample, otherwise, rejected. Speaker identification using Dynamic Time Warping (DTW) was implemented using a training or reference template for each speaker. During identification stage, a DTW score of the test utterance was made against each training template. Speaker identification was carried in favour of the speaker whose training template produced the lowest score, provided the score is within the threshold value. For speaker verification application, the test utterance was compared against the training template of the speaker who was being verified. The obtained DTW score was compared against a threshold value and the user was only verified if the score was lower than the threshold value set for the speaker. The verification threshold T was computed using this equation:

$$T = \frac{\mu_{spk}\,\mu_{imp} + \mu_{imp}\,\sigma_{spk}}{\sigma_{spk} + \sigma_{imp}} \quad (5)$$

Where (i) The mean, μ_{spk}, and standard deviation, σ_{spk}, is computed from the DTW score from each digit; and (ii) The mean, μ_{imp}, and standard deviation, σ_{imp}, is computed from the DTW score against this users template and speech samples of impostors.

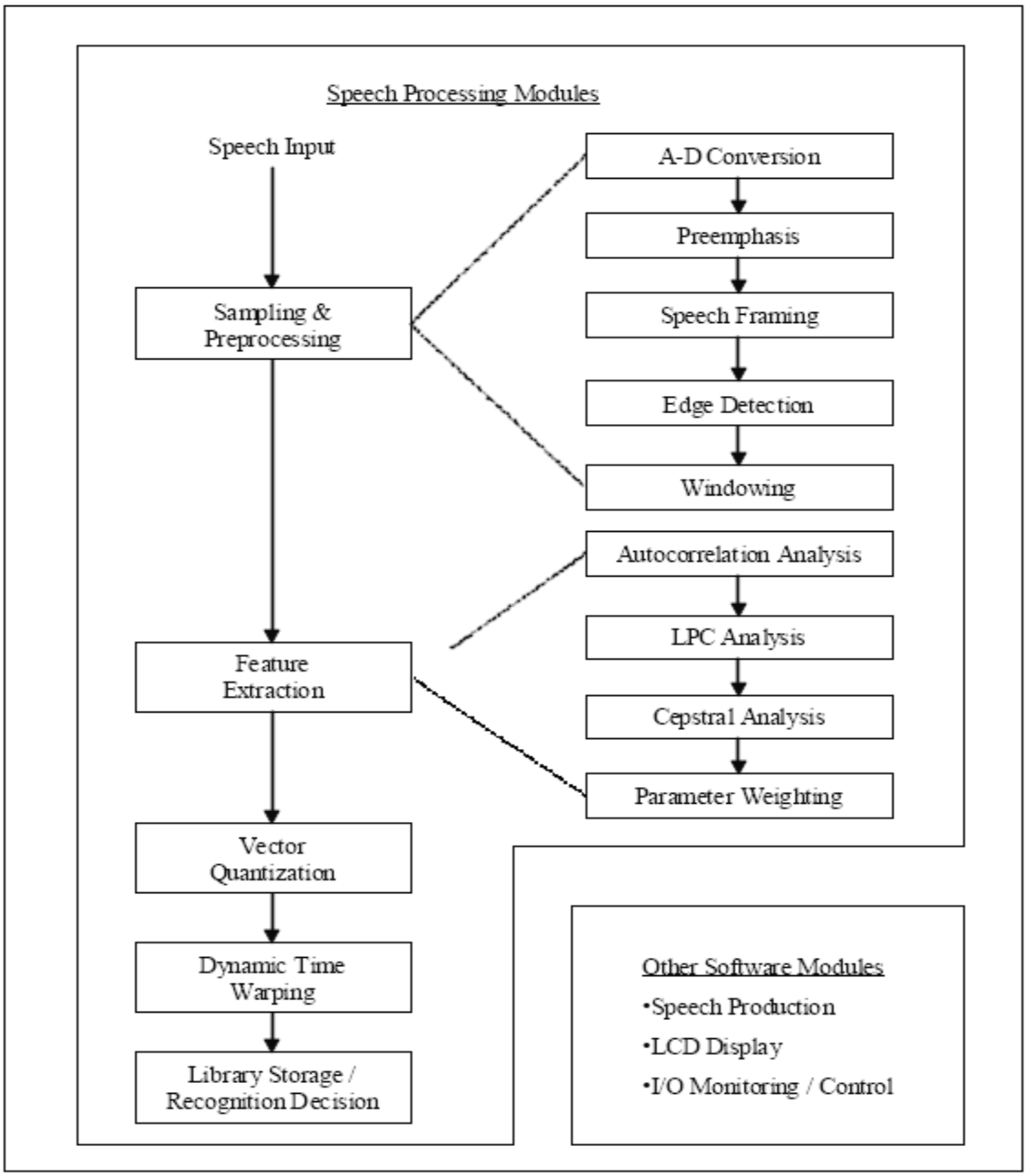

Figure 3: Block diagram of software modules developed for the voice recognition system

For the real-time experiment, the memory requirements for VQ and DTW were computed when implemented on a voice recognition system. The classifiers compared were the VQ and DTW. The memory required for the types of classifier implementations were noted along with the execution time. The execution time was only given for the classifier training and recognition routine. The memory and processing time results were recorded. Voice access was only granted if both identification and verification were successful. An application software was developed using C programming language with Code Composer Studio (CCS) to generate the source codes for autocorrelation analysis, LPC Cepstrum and DTW. The DSP Starter Kit (DSK) Debugger was used to download source code to the speaker recognition system, which executes decoding and monitoring. The average identification success rate and average verification success rate for both original speakers and impostors were given in percentage. The system performance was evaluated using Equal Error Rate.

4. DISCUSSION OF RESULTS

Personal identification and verification result of a true speaker is given in Table 1. The average identification success rate was 96%, and average verification success rate was 97%. The overall call success result of a true speaker is given in Table 2. The average call success rate for a true speaker was 96.5%.

Table 1: The Voice identification and verification success count

Spk	Successful Identification (True Acceptance)	Unsuccessful Identification (False Rejection)	Successful Verification (True Acceptance)	Unsuccessful Verification (False Rejection)
S 1	47	3	47	3
S 2	50	0	50	0
S 3	45	5	50	0
S 4	50	0	49	1
S 5	47	3	48	2
S 6	47	3	50	0
S 7	49	1	44	6
S 8	48	2	50	0
S 9	49	1	47	3
S10	48	2	50	0

Table 2: True Voice call attempts success count

Speaker	Successful Entry (True Acceptance)	Unsuccessful Entry (False Rejection)
S 1	45	5
S 2	50	0
S 3	45	5
S 4	49	1
S 5	45	5
S 6	47	3
S 7	44	6
S 8	48	2
S 9	46	4
S10	48	2

The total Storage/memory and processing time is summarized in Table 3.

Table 3: Storage and processing time for different classifiers

	Storage Location	Training Time	Speaker Identification Time	Speaker Verification Time
VQ	1.0Mb	8.40s	15.75s	0.16s
DTW	4.0Mb	0.00s	0.80s	0.02s
HMM	5.2Mb	250.0s	1.23s	0.02s
ANN	0.3Mb	1400.s	13.40s	0.26s

The training time listed is for each enrolment session. The speaker identification time was calculated on assumption that there were 100 enrolled users. From Table 3, the storage requirement needed for the VQ implementation was the least, with the DTW implementation required larger storage area. The VQ implementation requires a comparatively moderate amount of memory. The VQ consumes less memory than the DTW, which was expected due to the lousy compression nature of the VQ implementation. All the classifiers evaluated required memory location which was easily made possible in current design.

The time needed to enrol a user varies drastically between the classifiers. The DTW implementation required 0.50 second for training and found to be acceptable, and may be used for online training. A person can be made to wait during an enrolment session, and thereafter the trained database may be verified. If the verification is unsuccessful, speech samples may be prompted again from the user to retrain the user database. The training time of VQ was well beyond the waiting time for a user who was enrolling. The training may be carried out offline, during the idle processing time of the voice recognition system.
The speaker identification time for DTW classifier was within acceptable limit. The identification time of the VQ was quite long and may not be suitable in certain applications like telephone banking and telephone credit cards. The training time can be reduced by using a more powerful DSP.

The time needed for all 10 speakers who enrolled in the speaker recognition system were recorded. Prior to training, all speakers were briefed of the training procedure. Average training time was noted at 50.0 seconds. This included the voice sampling time of a minimum of 16.72 seconds. Sampling time increased due to verification of digit and login name sample. Speakers were requested by the system three times, if verification failed. The average speaker recognition time was noted at 7.80 seconds. This timing included the prompt and sample time of 5 seconds. The system was able to make access decision in an average of 2.80 seconds after the voice sampling was completed.

5. CONCLUSION

As the level of security breaches and transaction fraud increases, the need for highly secure identification and personal verification technologies is becoming apparent. Therefore, in order to aid forensics in criminal identification, authentication in civilian applications and for preventing un-authorized access, there is a need to develop a voice recognition system that would be able to provide solutions for confidential financial transactions and personal data privacy that reduces the high-tech computer theft or fraud in terms of access control, telephone banking and telephone credit cards.

This paper presents a model for maintaining data security and authenticity in voice-driven system whereby a system designed consists of memories and data acquisition modules that were well suited for a voice recognition system. Voice as a special characteristic of an individual, a form of biometric feature, could be used as a form of personal system identification and verification, and is recommended to be part of feature to be captured in the on-going government's activities like the acquisition of National Identification Number, Drivers Licence, International Passport, Integrated Payroll and Personnel Information System (IPPIS), etc.

REFERENCES

[1] Aborisade D.O., Alowosile O.Y., Odunlami K.O., and Odumosu A. (2013). Nigeria Computer Society 2013 Conference Proceedings. June 2013, pp 5-12.

[2] Aladesanmi O. A. T., Afolabi B .S and Oyebisi T. O. (2012). "Assessing Network Services and Security in Nigeria Universities". An International Journal of the Nigeria Computer Society (NSC). Vol.19 No 1, June 2012, pp 60-65.

[3] Alexandre Preti, Bertrand Ravera, François Capman,and Jean-François Bonastre (2008)."An Application Constrained Front End for Speaker Verification" 16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

[4] Campbell, J. and Reynolds D. A. (2004), "The MMSR Bilingual and Crosschannel Corpora for Speaker Recognition Research and Evaluation", *Proc. Odyssey '04*

[5] Champod C. and Meuwly, D. (2000). "The inference of identity in forensic speaker recognition". *Speech Communication.* 31(2-3): 193 - 203.

[6] David D. Zhang (2000). "Automated Biometrics. Boston: Kluwer: Academic Publishers McDowall, R.D., "Biometrics: The Password You'll Never Forget" retrieved from http://www.21cfr11.com/files/library/compliance/lcgc10_00.pdf

[7] David P. Beach (2004), " Fingerprint Recognition and Analysis System". Ph.D thesis submitted to the Department of Electronics and Computer Technology, Indiana State University.Terre Hante, Indiana.

[8] Driss Matrouf, Jean-Franc Bonastre, and Jean-Pierre Costa (2007) "Effect of impostor speech transformation on automatic speaker recognition" retrieved from http://www.nist.gov/speech/test/spk/Index.htm

[9] Hanwu Sun, Bin Ma and Haizhou Li (2008) "An efficient feature selection method for speaker recognition" retrieved from http://www.isca-speech.org

[10] Julius Zimmermann and Julius Zimmermann, Jr (2005). "Stochastic Speaker Recognition Model". Retrieved from zimmer@unipo.sk <htk.eng.cam.ac.uk>

[11] Kinnunen, T. and Kärkkäinen, I. (2002). "Class-Discriminative Weighted Distortion Measure for VQ-Based Speaker Identification". *Proc. Joint IAPR International Workshop on Statistical Pattern Recognition (S+SPR 2002.* August 6-9, Windsor, Canada, 681-688.

[12] Kwek, Ser Wee. (2000). "*Real Time Implementation of Speech Recognition System Using TI Floating-Point Processor TMS320C31".* Universiti Teknologi Malaysia: Final Year Project Thesis.

[13] Martin, A. and Przybocki, M. (2004), "The NIST Speaker Recognition Evaluation Series", *National Institute of Standards and Technology's web-site, http://www.nist.gov/speech/tests/spk*

[14] Martin, A. (2004), "Conversational Telephone Speech Corpus Collection for the NIST Speaker Recognition Evaluation 2004". *Proc LREC 2004*

[15] Polemi, D. (2005), "Biometric Techniques: Review And Evaluation Of Biometric Techniques For Identification And Authentication", retrieved from http://www.cordis.lu/infosec/src/stud5fr.htm

[16] Prakash Sowndappan (2006). Thesis titled: "Development of a real-Time Speaker Recognotion System Using TMS320C31". Faculty of Electrical Engineering, Universiti Tecknologi, Malaysia.

[17] Reynolds D.A. (2002). "Automatic Speaker Recognition: Acoustics and Beyond". In superSID project at JHU Summer Workshop. http://www.clsp.jhu.edu/ws2002/groups/supersid/.

[18] Sara Grassi (1998*). "Optimized implementation of speech processing algorithms".* University of Neuchatel: Ph.D. Thesis.

[19] Satish Gunnam (2004). Thesis titled: Fingerprint Recognition andAnalysis System". Dept. Of Electronics & Computer Technology. Indiana State University. Terre Hante, Indiana.