International Science, Technology, Education, Arts, Management & Social Sciences Conference (iSTEAMS) Research Nexus Conference, UNILORIN 2015

Collocated with

The African Mobile Apps Development Conference & Exhibition &
The Annual Workshop On Contemporary Classroom Teaching Using Digital Skills
(D-SKILLS Workshop @UNILORIN)

BOOK OF PROCEEDINGS

SERIES 7

Theme: Together Towards Tomorrow – Leveraging on Multidisciplinary Research for Human Capital & Global Development

Venue: University Auditorium
University of Ilorin
Ilorin, Kwara State, Nigeria

Date: 11th – 13th March, 2015

Organized By

Research Nexus Africa Networks & The Creative Research Networks

in Partnership with

University of Ilorin
Ilorin, Kwara State, Nigeria

Table of Contents

Note: Affiliations are 1st Author Affiliations

Comparative Analysis of the Digital Classifier Models for Real-Time Voice Recognition System

W.O. Adesanya
General Studies Department
Federal College of Agriculture, Akure, Nigeria
Department of Computer Science
Tai SolarinUniversity of Education, Ijagun, Ijebu-Ode, Nigeria
sanyalanre2003@yahoo.com

J. O. Okesola
Department of Computer Science
Tai SolarinUniversity of Education
Ijagun, Ijebu-Ode, Nigeria
School of Computing
University of South Africa (UNISA)
48948535@mylife.unisa.ac.za

ABSTRACT

The study gives the analysis of different classifier models, in terms of their processing time and memory requirements to identify and verify person through his or her voice in order to access restricted information or resources. Voice was chosen as a means for identification and verification purpose, since password can be lost or stolen or even counterfeited. The classifiers compared were the VQ, DTW, HMM and ANN. For the purpose of simulation and comparison, the voice recognition system has both the components of speaker identification and verification. At each stage of identification and verification, 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 Linear Predictive Coding (LPC) and Cepstral analysis for feature extraction techniques were used for damping. 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.

Keywords - Voice Recognition, Real-time, Dynamic Time Warping, Vector Quantization, Linear Predictive Coding

1. INTRODUCTION

Each person has unique anatomy, physiology and learned habits that familiar persons use in everyday life to recognize the person such as signature, fingerprints, voice, facial, etc. Increased computing power and decreased microchip size has given impetus for implementing realistic biometric authentication methods. Spoken language is the most natural way used by humans to communicate information. The voice signal conveys several types of information. From the voice production point of view, the voice signal conveys linguistic information (e.g., message and language) and speaker information (e.g., emotional, regional, and physiological characteristics). From the voice perception point of view, it also conveys information about the environment in which the voice was produced and transmitted. Even though this wide range of information is encoded in a complex form into the voice signal, humans can easily decode most of the information. Such human ability has inspired many researchers to understand voice production and perception for developing systems that automatically extract and process the richness of information in voice Qin (2007).

The voice technology has found wide applications such as automatic dictation, voice command control, audio archive indexing and retrieval etc. The application defines which information in the voice signal is relevant. For example, the linguistic information will be relevant if the goal is to recognize the sequence of words that the speaker is producing. The presence of irrelevant information (like speaker or environment information) may actually degrade the system accuracy. Voice recognition is a biometric modality that uses an individual's voice for recognition purposes. The speaker recognition process relies on features influenced by both the physical structure of an individual's vocal tract and the behavioural characteristics of the individual. 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 Aladesanmi et al (2012).

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 authorized to do so Aborisade et al (2013). 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 and for real-time operation, there is the need for the comparative analysis of classifier models that can aid the security of real-time utilization of voice-driven access to the restricted resources, since voice is unique to each person and cannot be lost or stolen. 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 analysis 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. David (2004) found 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 David (2000) 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 Julius et al (2005), 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 Hauwu et al (2008) 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. Sara (1998) concentrated on optimized speech processing in the DSP56001 hardware platform, especially in the application of noise reduction and speech enhancement. Kwek (2000) worked on a hardware based speech recognition system.

Both work by Sara (1998) and Kwek (2000) 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 (DTW), Vector Quantization (DQ), Hidden Markov Models (HMM) and Artificial Neural Networks (ANN) based models to determine a suitable model with better response time in real-time application for voice-driven recognition system.

3. METHODOLOGY

The classifiers compared were the VQ, DTW, HMM and ANN. For the purpose of simulation and comparison, the voice recognition system has both the components of speaker identification and verification. A user has to first utter a phrase to identify his or her self, similar to giving a "login name". The system performs a speaker identification routine based on the login utterance. The user is later required to give a text prompted four combination digit, for example "3-6-7-4". This prompted password is edge detected, and sliced into four different utterances of digits ranging from 0 to 9. A speaker verification routine is executed based on the database of the person verified during the "login" phase.

All classifiers are trained or implemented on digits 0 to 9, and each digit being sampled in 1 second duration. 5 samples are taken for each digit during the training phase. For speaker identification, 5 samples of 2 second phrase were recorded for each user. All 50 digit samples and 5 identification phrases were used to train the VQ, HMM and ANN implementation. Only 1 sample of digits and recognition phrase was used for the DTW implementation. The memory requirements for VQ, DTW, HMM and ANN were computed when implemented on a voice recognition system. The memory required for the types of classifier implementations were noted along with the execution time.

The total memory required for VQ is calculated as follows:

memory required = (number of users) x (codebook size) x (coefficients per frame/codebook entry) x (bytes per coefficient). .... (1)

The amount of storage required for DTW is completed as follows:

memory required = (number of users) x (number of models per user) x (number of frames per model) x (coefficients per frame) x (bytes per coefficient). ....(2)

The storage required to store the neural network depends on the number of nodes contained in the network. The storage is needed to store the weights of interconnecting nodes, and the offset of each node in the hidden and outer layer.

The storage required for three layer network is computed as following:

total storage = $H(I+1) + O(H+1)$, where H is the number of hidden nodes, I is the number of input nodes, and O is the number of output nodes. .... (3)

The storage allocation to store the HHM parameters required for each word is calculated as:

$N(N+M+1)$ .... (4)

where N and M represent the number of state and the number of symbol respectively.

The real-time clock (RTC) was used to keep track of time on events and activities within the system. 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.

4. DISCUSSION OF RESULTS

The total memory and processing time is summarized in Table 1. The training time listed is for each enrolment session. The speaker identification time is calculated on assumption that there are 100 enrolled users. The storage rèquirement needed for the ANN implementation is the least, with the DTW and HMM implementation requiring larger storage area. The VQ implementation requires a comparatively moderate amount of memory. All the classifiers evaluated require memory location which is easily made possible in current designs.

Table 1: 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 time needed to enroll a user varies drastically between the other types of classifier. The DTW implementation requires no training at all. The VQ implementation requires training time which is 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 HMM and ANN is well beyond the waiting time for a user who is enrolling. The training may be done offline, during the idle processing time of the speaker recognition system. The HMM training time does not show a significant change, but the ANN training time varies drastically.

The speaker identification time for DTW and HMM classifiers are within acceptable limit. The identification time of the VQ and ANN are quite long and may not be suitable in certain applications like door-access. The training time can be reduced by using a more powerful DSP. The usage of a more powerful DSP has its drawback in actual application due to extra power consumed by the DSP. Power consumption is a major design trade-off in hand-held applications, especially systems that run on battery, or systems being battery backed-up in the absence of power. It is apparent that the speaker identification is still feasible using any of the classifiers when the population is up to 100 users. The identification time increases when the number of users increase, which causes speaker identification not feasible in application like banking. The speaker verification is computably feasible using any of the classifier evaluated here. The verification time is not affected by the number of enrolled users.

5. CONCLUSION

This paper presents a platform for comparing classifier models for maintaining data security and authenticity in voice-driven system in terms of memories and data acquisition modules that were well suited for a voice recognition system. Evaluations of various classifiers have shown that speaker recognition is feasible to be implemented in a real-time environment, and in a stand-alone hardware platform. The training time of certain classifier which is lengthy may be done offline or during the system's idle time. The speaker identification may not be feasible in application with a large population of users. A security system may alternately apply speaker verification along with conventional identification token or another biometric feature instead of applying speaker identification. A conventional token like an ATM card or key may be used to identify the user before speaker verification is performed.

REFERENCES

[1] Aborisade D.O., Alowosile O.Y., Odunlami K.O., and Odumosu A. (2013). Nigeria Computer Society 2013 Conference Proceedings. June 2013, Vol 24. pp 5-12.

[2] Adesanya W.O., Alaba O. B., and Adeyemo A. B. (2014). "A Model For Maintaining Data Security and Authenticity In Voice-Driven System" Nigeria Computer Society 2014 Conference Proceedings. June 2014, Vol 25. pp 8-12.

[3] 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.

[4] 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

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

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

[7] 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

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

[9] DrissMatrouf, 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

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

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

[12] 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.

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

[14] 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

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

[16] 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

[17] PrakashSowndappan (2006). Thesis titled: "Development of a real-Time Speaker Recognotion System Using TMS320C31". Faculty of Electrical Engineering, UniversitiTecknologi, Malaysia.

[18] Qin Jin (2007). "Thesis titled: Robust Speaker Recognition". Carnegie Mellon University. Pittsburgh.

[19] 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/.

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

[21] SatishGunnam (2004). Thesis titled: Fingerprint Recognition and Analysis System". Dept. Of Electronics & Computer Technology. Indiana State University. Terre Hante, Indiana.