International Journal of Research in Computer Science
eISSN 2249-8265 Volume 3 Issue 6 (2013) pp. 9-15
www.ijorcs.org, A Unit of White Globe Publications
doi: 10.7815/ijorcs.36.2013.073

TEXT INDEPENDENT BIOMETRIC SPEAKER RECOGNITION SYSTEM

Luqman Gbadamosi
Computer Science Department, Lagos State Polytechnic, Lagos, Nigeria
Email: luqmangbadamosi@yahoo.com

Abstract: *Designing a machine that mimics the human behavior, particularly with the capability of responding properly to spoken language, has intrigued engineers and scientists for centuries. The earlier research work on voice recognition system which is text-dependent requires that the user must say exactly the same text or passphrase for both enrollment and verification before gaining access. In this method the testing speech is polluted by additive noise at different noise decibel levels to achieve only 75% recognition rate and would require full cooperation by the speaker which could not be used for forensic investigation. This paper presents the historical background, and technological advances in voice recognition and most importantly the study and implementation of text-independent biometric voice recognition system which could be used for speaker identification with 100% recognition rate. The technique makes it possible to use the speaker's voice to verify their identity and control access to services such as voice dialing, telephone shopping, database access services, information services, voice mail, and remote access to computers. The implementation mainly incorporates Mel frequency Cepstral Coefficient (MFCCs) which was used for feature extraction and Vector quantization using the Linde-Buzo-Gray (VQLBG) algorithm used to minimize the amount of data to be handled. The matching result is given on the basis of minimum distortion distance. The project is coded in MATLAB.*

Keywords: *MFCC, Voice Print, VQLBG, Voice Recognition*

I. INTRODUCTION

Speaker recognition refers to recognizing every human from their voice. History has shown time and time over that no two individuals sound identical because their vocal tract shapes, larynx sizes, and other parts of their voice production organs are different. However, these physical differences, each speaker has his or her characteristic manner of speaking, including the use of a particular accent, rhythm, intonation style, pronunciation pattern, choice of vocabulary and so on. Most recent state of-the-art speaker recognition systems use a number of these features in parallel, attempting to cover these different aspects and employing them in a complementary way to achieve more accurate recognition rates.[1][2][3][4][5]

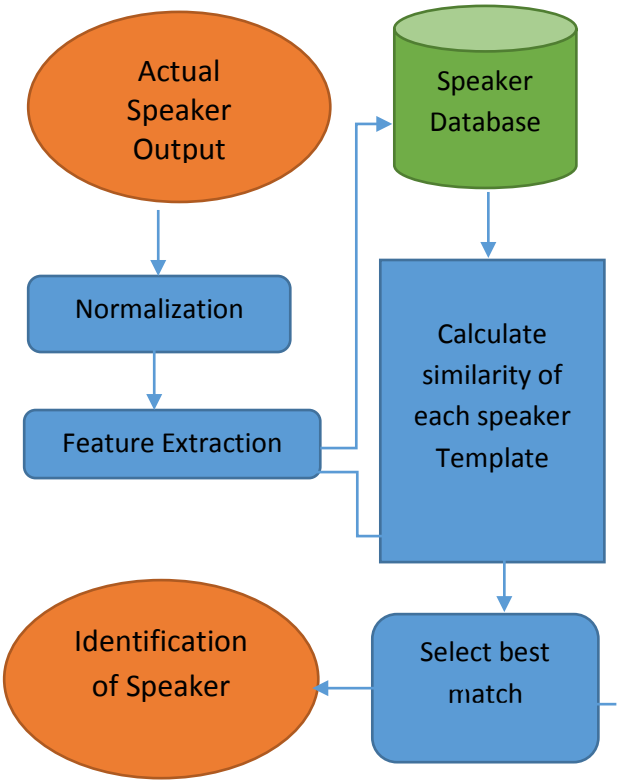

Figure 1: Speaker Identification and Recognition Diagram

Speech Recognition does not only cut across every facet of human endeavours, but has found its applications on various aspects of our daily lives from automatic phone answering service to dictating text and issuing commands to computers. The main essence of communication is understanding while language is human most important means of communication and speech is its primary medium. [6]

An important application of speaker recognition technology which this research work would be most useful to is forensics. Much of information is exchanged between two parties in telephone conversations, including between criminals, and in recent years several efforts has been made to accomplish this.

In forensics and speaker diarization, the speaker can be considered non-cooperative as they do not specifically wish to be recognized. On the other hand, in telephone-based services and access control, the users are considered cooperative. Speaker recognition systems, on the other hand, can be divided into text-dependent and text-independent ones. In text-dependent systems, suited for cooperative users, the recognition phrases are fixed, or known beforehand. For instance, the user can be prompted to read a randomly selected sequence of numbers. Meanwhile, in text-independent systems, there are no constraints on the words which the speakers are allowed to use. Thus, the reference (spoken in training) and the test (what are uttered in actual use) utterances may have completely different content, and the recognition system must take this phonetic mismatch into account. Text independent recognition is the much more challenging of the two tasks. [1][2][3]

A. Classification of Speaker Recognition Methods

The problem of speaker recognition can be divided into two major sub problems:

Speaker identification can be thought of, as the task of determining who is talking from a set of known voices of speakers. It is the process of determining who has provided a given utterance based on the information contained in speech waves. The unknown voice comes from a fixed set of known speakers, thus the task is referred to as closed set identification.

Speaker Verification on the other hand is the process of accepting or rejecting the speaker claiming to be the actual one. Since it is assumed that imposters (those who fake as valid users) are not known to the system, this is referred to as the open set task. Adding none of the above option to the closed set identification task would enable merging of the two tasks, and it is called open set identification.

Both the text dependent and independent methods share a problem. These systems can be deceived because someone who plays back a recorded voice of a registered speaker saying the key words or sentences can be accepted as the registered speaker. Even the use of pre-determined set of words or digits that are randomly chosen every time can be reproduced in the requested order by an advanced electronic recording equipment. Therefore a text prompted (machine driven text dependent) speaker recognition system could be considered.

With the merger of speaker and speech recognition systems and improvement in speech recognition accuracy, the distinction between text dependent and independent applications will eventually decrease. The text dependent speaker recognition is the most commercially viable and useful technology, although there has been much research conducted on both the tasks.

However, due to the possibilities offered, more attention is being paid to the text independent methods of speaker recognition irrespective of their complexity. [12].

II. HISTORY

In the last 50 years, research in speech and speaker recognition has been intensively carried out worldwide, spurred on by advances in signal processing, algorithms, architectures, and hardware. The technological progress in the last 50 years can be summarized by the following changes [11]:

- From template matching to corpus-base statistical modeling, e.g. HMM and n-grams,
- From filter bank/spectral resonance to cepstral features (cepstrum + Δcepstrum + ΔΔcepstrum),
- From heuristic time-normalization to DTW/DP matching,
- From “distance”-based to likelihood-based methods,
- From maximum likelihood to discriminative approach, e.g. MCE/GPD and MMI,
- From isolated word to continuous speech recognition,
- From small vocabulary to large vocabulary recognition,
- From context-independent units to context-dependent units for recognition,
- From clean speech to noisy/telephone speech recognition,
- from single speaker to speaker-independent/adaptive recognition,
- from monologue to dialogue/conversation recognition,
- From read speech to spontaneous speech recognition,
- From recognition to understanding,
- From single-modality (audio signal only) to multimodal
- (audio/visual) speech recognition,
- From hardware recognizer to software recognizer, and
- From no commercial application to many practical commercial applications.

Most of these advances have taken place in both the fields of speech recognition and speaker recognition. The majority of technological changes have been directed toward the purpose of increasing robustness of recognition, including many other additional important techniques not noted above. Recognition systems have been developed for a wide variety of applications,

ranging from small vocabulary keyword recognition over dialed-up telephone lines, to medium size vocabulary voice interactive command and control systems for business automation, to large vocabulary speech transcription, spontaneous speech understanding, and limited-domain speech translation.[11]

Although we have witnessed many new technological promises, we have also encountered a number of practical limitations that hinder a widespread deployment of applications and services. The first speech recognizer appeared in 1952 and consisted of a device for the recognition of single spoken digits. Another early device was the IBM Shoebox, exhibited at the 1964 New York World's Fair. Speech recognition technology has also been a topic of great interest to a broad general population since it became popularized in several blockbuster movies of the 1960‟s and 1970‟s, most notably Stanley Kubrick‟s acclaimed movie “2001: A Space Odyssey”[11]. In this movie, an intelligent computer named “HAL” spoke in a natural sounding voice and was able to recognize and understand fluently spoken speech, and respond accordingly. This anthropomorphism of HAL made the general public awareness of the potential of intelligent machines. In the famous Star Wars saga, George Lucas extended the abilities of intelligent machines by making them mobile as well as intelligent and the droids like R2D2 and C3PO were able to speak naturally, recognize and understand fluent speech, and move around and interact with their environment, with other droids, and with the human population at large. In 1988, in the technology community, Apple Computer created a vision of speech technology and computers for the year 2011, titled “Knowledge Navigator”, which defined the concepts of a Speech User Interface (SUI) and a Multimodal User Interface (MUI) along with the theme of intelligent voice-enabled agents. This video had a dramatic effect in the technical community and focused technology efforts, especially in the area of visual talking agents. [6]

III. IMPLEMENTATION

A. Design Description

The text independent Biometric Speaker Recognition system is accomplished by reading audio data from train folder and also from test folder for performing operations to compute MFCC of the audio data to be used in speech processing for both test and train folders and lastly compute voice quantization of the audio data used to be used in speech processing for both test and train voices.

Mel Frequency Cepstral Coefficient (MFCC): The first step in any automatic speech recognition system is to extract features i.e. identify the components of the audio signal that are good for identifying the linguistic content and discarding all the other stuff which carries information like background noise, emotion etc. The main point to understand about speech is that the sounds generated by a human are filtered by the shape of the vocal tract including tongue, teeth etc. This shape determines what sound comes out. If we can determine the shape accurately, this should give us an accurate representation of the phoneme being produced. The shape of the vocal tract manifests itself in the envelope of the short time power spectrum, and the job of MFCCs is to accurately represent this envelope. Mel Frequency Cepstral Coefficents (MFCCs) are a feature widely used in automatic speech and speaker recognition. They were introduced by Davis and Mermelstein in the 1980's, and have been state-of-the-art ever since. Prior to the introduction of MFCCs, Linear Prediction Coefficients (LPCs) and Linear Prediction Cepstral Coefficients (LPCCs) and were the main feature type for automatic speech recognition (ASR). This page will go over the main aspects of MFCCs, why they make a good feature for ASR, and how to implement them. [13]

Vector Quantization: Vector quantizer encoder it computes for a given input, the index of nearest code word based on Euclidean or weighted Euclidean distance measure.

The Vector Quantizer Encoder block compares each input column vector to the codeword vectors in the codebook matrix. Each column of this codebook matrix is a codeword. The block finds the codeword vector nearest to the input column vector and returns its zero-based index. This block supports real floating-point and fixed-point signals on all input ports. The block finds the nearest codeword by calculating the distortion. The block uses two methods for calculating distortion: Euclidean squared error (unweighted) and weighted Euclidean squared error. Consider the codebook, $CB=[CW_1\, CW_2 \ldots CW_N]$. This codebook has N code words; each codeword has k elements. The i-th codeword is defined as a column vector, $CW_i = [a_{1i}\, a_{2i} \ldots a_{ki}]$. The multichannel input has M columns and is defined as $U=[U_1\ U_2 \ldots U_M]$, where the pth input column vector is $U_p=[U_{1p}\, U_{2p} \ldots U_{kp}$. The squared error (un-weighted) is calculated using the equation

$$D = \sum_{j=1}^{k} \left(a_{ji} - u_{jp}\right)^2$$

The weighted squared error is calculated using the equation

$$D = \sum_{j=1}^{k} w_j \left(a_{ji} - u_{jp}\right)^2$$

where the weighting factor is defined as , $W = [w_1 w_2 \ldots w_k]$. The index of the codeword that is associated with the minimum distortion is assigned to the input column vector.[14]

B. Voice Recognition Algorithm

A voice analysis is done after taking an input through microphone from a user. The design of the system involves manipulation of the input audio signal. At different levels, different operations are performed on the input signal such as Pre-emphasis, Framing, Windowing, Mel-Cepstrum analysis and Recognition (Matching) of the spoken word. The voice algorithms consist of two distinguished phases. The first one is training sessions, whilst, the second one is referred to as operation session or testing phase as described in figure 2 [24][26].

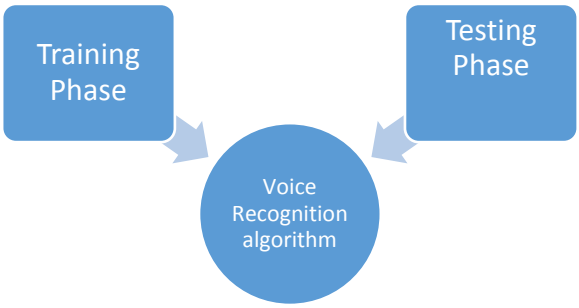

Figure 2: Voice Recognition Algorithm

Training Phase: Each speaker has to provide samples of their voice so that the reference template model can be built.

Testing Phase: To ensure that input that input test voice match with stored reference template model and recognition decision is made.

Feature Extraction: The extraction of the best parametric representation of acoustic signals is an important task to produce a better recognition performance. The efficiency of this phase is important for the next phase since it affects its behavior. MFCC is based on human hearing perceptions which cannot perceive frequencies over 1Khz. In other words, MFCC is based on known variation of the human ear's critical bandwidth with frequency [8-10]. MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz. A subjective pitch is present on Mel Frequency Scale to capture important characteristic of phonetic in speech. The overall process of the MFCC is shown in Figure 3, [24, 25]

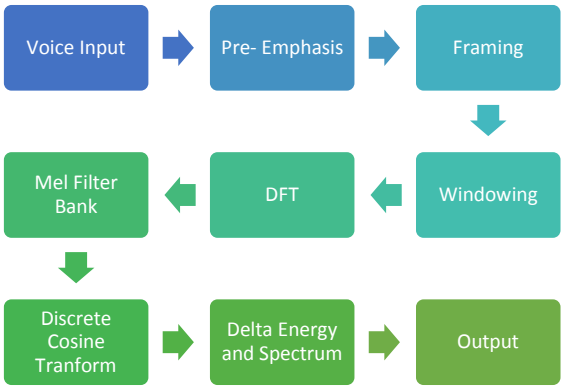

Figure 3: MFCC Block Diagram

As shown in Figure 4, MFCC consists of seven computational steps. Each step has its function and mathematical approaches as discussed briefly in the following:

Step 1: Pre-emphasis

This step processes the passing of signal through a filter which emphasizes higher frequencies. This process will increase the energy of signal at higher frequency.[26]

$$Y[n] = X[n] – 0.95\ X\ [n-1]$$

Lets consider a = 0.95, which make 95% of any one sample is presumed to originate from previous sample.

Step 2: Framing

The process of segmenting the speech samples obtained from analog to digital conversion (ADC) into a small frame with the length within the range of 20 to 40 msec. The voice signal is divided into frames of N samples. Adjacent frames are being separated by M (M<N).

Typical values used are M = 100 and N= 256.

Step 3: Hamming windowing

Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines.

Step 4: Fast Fourier Transform

To convert each frame of N samples from time domain into frequency domain. The Fourier Transform is to convert the convolution of the glottal pulse U[n] and the vocal tract impulse response H[n] in the time domain. This statement supports the equation below:

$$Y(w) = FFT[h(t)*X(t)] = H(w)*X(w)$$

If X (w), H (w) and Y (w) are the Fourier Transform of X (t), H (t) and Y(t) respectively.

Step 5: Mel Filter Bank Processing

The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale. Each filter's magnitude frequency response is triangular in shape and equal to unity at the centre frequency and decrease linearly to zero at centre frequency of two adjacent filters [26]. Then, each filter output is the sum of its filtered spectral components. After that the following equation is used to compute the Mel for given frequency f in Hz:

$$F(Mel) = [2595 * \log 10[1+f]700]$$

Step 6: Discrete Cosine Transform

This is the process to convert the log Mel spectrum into time domain using Discrete Cosine Transform (DCT). The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector.

Step 7: Delta Energy and Delta Spectrum

The voice signal and the frames changes, such as the slope of a formant at its transitions. Therefore, there is a need to add features related to the change in cepstral features over time. 13 delta or velocity features (12 cepstral features plus energy), and 39 features a double delta or acceleration feature are added. The energy in a frame for a signal x in a window from time sample t1 to time sample t2, is represented at the equation below:[26]

$$\text{Energy} = \sum X^2[t]$$

C. Speech Samples Matching

All eight (8) voice samples data values are loaded in sound database file sounddatabase1.dat. From Train folder samples are tested one by one to database stored in file sounddatabase1.dat and matching result is given on the basis of minimum distortion distance between the corresponding sound files in both test and train folder which eventually determine precisely the best match and identify the speaker.

The following is a diagram showing the detail description of the steps required at the design stage to accomplish the best matching score.[24]

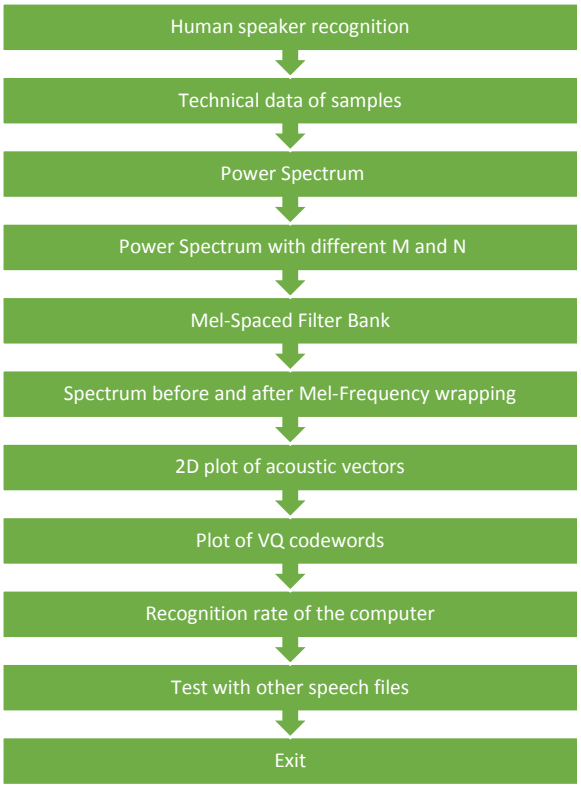

Figure 4: Detail Design Description

IV. RESULTS

The performance rating of speaker recognition method used would recognize the speaker's voice 100% of the time. This system was able to recognize 8 out of 8 speakers. This is an error rate of 0.00%. The recognition rate of this system is much better than the one of a human's recognition rate.

However one must be aware that this test is not really representative of the computer's efficiency to recognize voices because I only tested on 8 persons, with only one training session and with only one word.

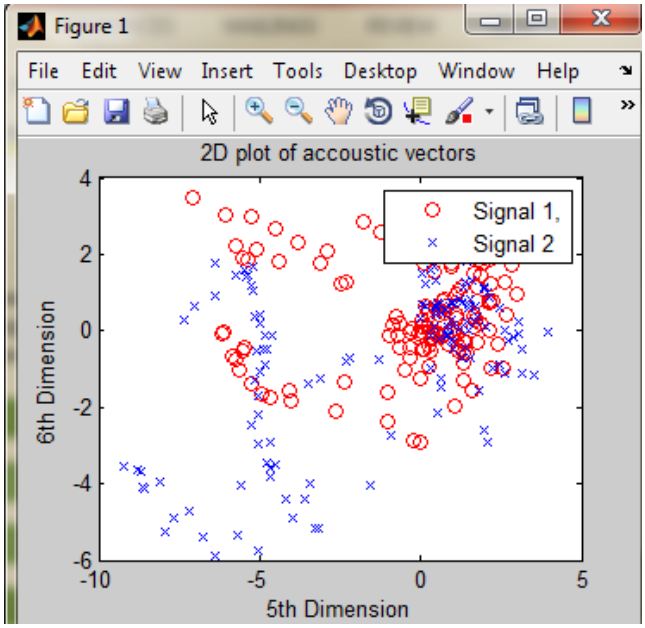

Figure 5: 2D Plot of acoustic vectors for two signal

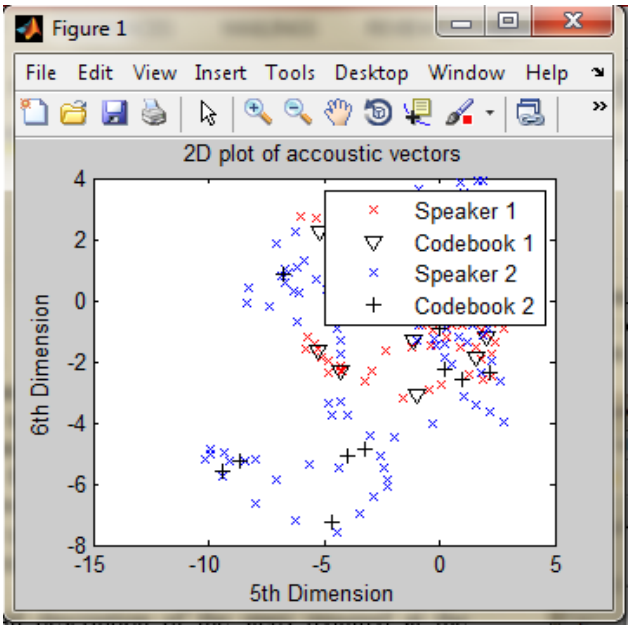

Figure 6: Plot of VQ code words

A. Performance Evaluation Index

The indexes used below are well accepted to determine the recognition rate of voice recognition system. Error that can occur in speaker identification is the false identification of speaker and the errors in speaker verification can be classified into the following two categories:

- **False Rejections**: a true speaker is rejected as an imposter, and
- **False Acceptances:** a false speaker is accepted as a true one [12]

In most systems for speaker recognition, a distance towards stored speaker template is computed and is compared with predetermined threshold. If the computed distance is below the threshold the speaker is verified, otherwise speaker is rejected as an imposter. The decision threshold is located at the point where the probabilities of both the errors are equal.

The same approach was used in the experiment conducted and the identity of the speaker was recognized when the computed distance between the speakers' stored voice print template and predetermined threshold.

V. CONCLUSION

This above implementation was an effort to understand how speaker recognition is used as the best form of biometric to recognize the identity of human voice. It briefly describe all the stages from enrollment of voice samples, power spectrum computation, Mel frequency wrapping to plotting of acoustic vectors and VQ code words which generate the highest percentage of matching score. Different standard techniques were also used at the intermediate stage of the processing.

High percentage verification rate arise due to the ease of developing algorithm to recognize human voice as different data are obtained for voice samples recorded on different occasions.

Also a major challenge is that the system could be deceived by playing a recorded sound of the original speaker to gain access and also the inability of the technique to record directly from the headset plugged to software and the computer inbuilt microphone except for externally connected microphone. It is recommended that future research work should focus on building a more effective voice recognition system for access control whereby a recorded voice would not be able to gain access to services like telephone banking, telephone shopping, database access services, information services, voice mail, and remote access to computers and also achieving connectivity to devices like headset and computer inbuilt microphone to be able to connect with the software for voice recording.

VI. REFERENCES

[1] Alexander, A., Botti, F., Dessimoz, D., Drygajlo, A., "The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications". Forensic Science International 146S, December 2004, pp. 95–99. doi: 10.1016/j.forsciint.2004.09.078

[2] Gonzalez-Rodriguez, J., Garcia-Gomar, D. G.-R. M., Ramos-Castro, D., Ortega-Garcia, J. "Robust likelihood ratio estimation in Bayesian forensic speaker recognition". In: Proc. 8th European Conf. on Speech Communication and Technology (Eurospeech 2003), Geneva, Switzerland, September 2003, pp. 693–696.

[3] Niemi-Laitinen, T., Saastamoinen, J., Kinnunen, T., Fränti, P., 2005. "Applying MFCC-based automatic speaker recognition to GSM and forensic data". In: Proc. Second Baltic Conf. on Human Language Technologies (HLT'2005), Tallinn, Estonia, April 2005, pp. 317–322. doi: 10.1016/j.specom.2009.08.009

[4] Pfister, B., Beutler, R., 2003. "Estimating the weight of evidence in forensic speaker verification". In: Proc. Eighth European Conf. on Speech Communication and Technology (Eurospeech 2003), Geneva, Switzerland, September 2003, pp. 701–704.

[5] Thiruvaran, T., Ambikairajah, E., Epps, J., 2008. "FM features for automatic forensic speaker recognition". In: Proc. Interspeech 2008, Brisbane, Australia, September 2008, pp. 1497–1500.

[6] Palden Lama and Mounika Namburu, "Speech Recognition with Dynamic Time Warping using MATLAB", CS 525, SPRING 2010-PROJECT REPORT

[7] B. H. Juang, L. R. Rabiner, "Automatic Speech Recognition – A Brief History of the Technology Development", Elsevier Encyclopedia of Language and Linguistics (2005)

[8] R. P. Lippmann, "Review of Neural Networks for Speech Recognition, Readings in Speech Recognition", A. Waibel and K. F. Lee, Editors, Morgan Kaufmann Publishers, pp. 374-392, 1990

[9] B.H. Juang, C.H. Lee and Wu Chou, "Minimum Classification Error Rate Methods for Speech Recognition", IEEE Trans. Speech & Audio Processing, T-SA, vo.5, No.3, pp.257-265, May 1997. doi: 10.1109/89.568732

[10] L. R. Bahl, P. F. Brown, P. V. deSouza and L. R. Mercer, "Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition", Proc. ICASSP 86, Tokyo, Japan, pp. 49-52, April 1986. doi: 10.1109/ICASSP.1986.1169179

[11] S. Furui, "Fifty years of progress in speech and speaker recognition", Proc. 148th ASA Meeting, 2004. doi: 10.1121/1.4784967

[12] S. K. Singh, Prof P. C. Pandey, "Features and Techniques for Speaker Recognition", M. Tech. Credit Seminar Report, Electronic Systems Group, EE Dept, IIT Bombay submitted Nov 03.

[13] Davis S., Mermelstein P., "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences". In IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 28 No. 4, pp. 357-366, 1980. doi: 10.1109/TASSP.1980.1163420

[14] "Vector Quantizer Encoder: Blocks(Signal Processing Blockset)", The Mathworks incorporation, 1982-2008

[15] ITU-T Recommendation G.711, "Pulse Code Modulation (PCM) of Voice Frequencies", General Aspects of Digital Transmission Systems; Terminal Equipments, International Telecommunication Union (ITU), 1993.

[16] Beigi, Homayoon (2011). " Fundamentals of Speaker Recognition.".[Online].Available: http://www.wikipedia.org/wiki/speaker_recognition.

[17] Course project (Fall 2009) "Voice Recognition Using MATLAB". California State University Northridge during the semester. [Online] Available: http://www.cnx.org/content/m33347/1.3/module_export?format=zip

[18] 2012, "Article on Human Voice" [Online]. Available: http://www.wikipedia.org/wiki/Human voice.

[19] "Techniques of Voice Recognition System" [Online].Available:http://www.hitl.washington.edu/scllw/EVE/I.D.2.d.VoiceRecognition.htm

[20] "Probability Tutorials on Chebyshevs-Inequality" [Online]. Available: http://www.statistics.about.com/od/probHelpandTutorials/a/Chebyshevs-Inequality.htm

[21] Sangram Bana, "Fingerprint Recognition System using Image Segmentation". International Journal of Advanced Engineering Sciences and technologies Vol No. 5, Issue No. 1, 012 – 023

[22] Kapil Sharma, H.P Sinha & R.K Aggarwal "Comparative study of speech Recognition System using various feature extraction techniques". International Journal of Information Technology and Knowledge Management Vol 3, No2, pp. 695-698.

[23] Mahima Garg, Omar Razi, Supriya Phutela, Vaibhav Kapoor, Varun Chopra, "Voice Recognition and Identification System in MATLAB" Final Project Report [Online]. Available: http://www.youtube.com/watch?v=UgBlJJ83oo0

[24] Mohammed Waleed Kadous ,"Machine Learning Reasearch" [Online]. Available: http://www.cse.unsw.edu.au/~waleed/phd/html/node38.html, downloaded on 3rd March 2010.

[25] Zaidi Razak, Noor Jamilah Ibrahim, Emran Mohd Tamil, Mohd Yamani Idna Idris, Mohd Yaakob Yusoff, "Quranic Verse Recitation Feature Extraction Using Mel Frequency Coestral Coefficient(MFCC)", Universiti Malaya.

[26] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques" Journal of computing, Vol 2, Issue 3, March 2010,

How to cite

Luqman Gbadamosi, "Text Independent Biometric Speaker Recognition System". *International Journal of Research in Computer Science*, 3 (6): pp. 9-15, November 2013. doi: 10.7815/ijorcs.36.2013.073