FINAL REPORT ON DESIGN AND CONSTRUCTION OF CUSTOMER IDENTIFICATION BY THEIR VOICES IN TELECOMMUNICATIONS INDUSTRY

BY

ENGR. MAHMOOD ABDULHAMEED

(Department of Electrical and Electronics Engineering,ATBU Bauchi
amahmood@atbu.edu.ng, 08038139649)

SUBMITTED AS A FINAL REPORT OF THE 2017/2018 NCC RESEARCH GRANT

MARCH 2023

ABSTRACT

This research grant aim at design and construction of customer identification system in telecommunications industry by their voices in an effort to produce a system that can be used by telecommunication industry and other intelligence agencies for searching criminal by their voices over telecommunication system. Initially the project was intended for the development of a hybrid system but later it was changed to software system because it will be better for this area of application. The method of Mel frequency Ceptrum Coefficient (MFCC) and Pitch was used in this research and finally GMM was used for the improvement of the system performance which is part of the novelty of this research because other existing systems have low performance especially for country with many languages. The system was designed on Matlab 2022b and implemented on a Python 3.10. The validation accuracy of the system was 98% with test accuracy of 100% based on the An4 database. The system was then practically tested with a database which contains male and female Igbo, Male and Female Hausa and male and female Yoruba Speakers and still the accuracy of the system was confirmed. The developed system was also tested over a database of size 100 and still the performance was satisfactorily. Based on the result obtained a higher accuracy was achieved and the system can be operated in multilingual country like Nigeria. And finally this system is recommended for telecommunication industries to be used for tracing kidnappers and other militant because once they call for request of money their voices can be used to trace their other previous call records to be able to recognize their true identity or to trace their past record from the telecommunication database.

1. INTRODUCTION

Customer Identification in telecommunication is the process of identifying a customer on the basis of speech alone. This employ the use of speaker identification model, the idea is that it can automatically identify the person speaking given a group of speakers, so in the case of group of people speaking it can help automatically to identify who is speaking (Jain *et al, 2014)*.

The idea of speaker identification service has the ability to recognize customers based on their voices, as voice has a unique characteristics that can be used to recognize and identify a speaker based on the principle that " The physical configuration of Vocal Tract is unique which varies from individual to individual". Speaker Identification mainly involves two modules as follows; (a) feature extraction and (b) feature matching. Feature extraction is the process that extracts the unique identity from the speaker's voice signal that can later be used to represent that speaker, while feature matching involves the actual procedure to identify the unknown speaker by comparing the extracted features from his/her voice input with the ones that are already stored in the speech database (Darshan *et al,2011)* .

In feature extraction Mel Frequency Cepstrum Coefficients (MFCC) are used, which are based on the known variation of the human ear's critical bandwidths with frequency and these, are vector quantized (VQ) by Linde Buzo Gray algorithm resulting in the speaker specific codebook. Vector Quantization Identities distortion between the input utterance of an unknown speaker and the codebooks stored in the databas. Based on VQ distortion metric would decide whether to accept/reject the unknown speaker's identity (Darshan *et al,2011)*.

When this model is made to work over telecommunication data it is expected that it will help in war against bad groups because it will permit Telecom industries to search for people by their voices after making calls or voice chats, to search for a Customer there is no need to know his number or internet protocol (IP) address it only need to have his voice saved in the system . It works based on the idea that if I knew your voice then, phone number and IP addresses are not required once customer voice is in the system, the system will automatically check if the voice made any call or voice chat and once detected all the calls he made with corresponding dates, phone number used and from which location he made any of the calls can be traced by the telecommunication company including tower used to originate the call. A research work of this magnitude is valuable to all telecommunication companies not only in Nigeria but also in the entire world because security is the major concern in telecommunication industries due to the fact that we have a lot of security challenges like Boko Haram, Niger Delta Militant, Armed Robbers, Hackers, Corruption, and Kidnapers etc. Pictures of all known wanted Boko Haram are available and most of their voices is also available like Shekau etc. It is believed that all the time most of the people called wanted people do make calls using telecom networks but with unknown phone numbers. Developing this kind of device is of great advantage in war against bad groups. It is also expected that this work is likely to bring an end to kidnapping cases.

2. LITERATURE REVIEW

Speaker Identification is the process of using the voice of speaker to verify their identity for telecommunication security or control access to services such as voice dialing, mobile banking, database access services, voice mail or security control to a secured system. Speaker Identification is the process of using the voice of speaker to verify their identity and control access to services such as voice dialing, mobile banking, database access services, voice mail or security control to a secured system (Geeta *et al.,* 2014). In most cases the idea of speaker recognition focus on security system of controlling the access to secure data from being used by unauthorized person but recently, the attention has turned to voice recognition for telecommunication application. Some researchers have earlier identified voice as important trend in telecommunication research and development (Douglas *et al,.* 2000).

It was a great development that recent findings revealed that speaker recognition model can be applied over telecommunication channel due to the fact that automatic detection of people's identity from their voices is part of modern telecommunication services (Laura, 2014) but from few author cited below it is clear the accuracy and ability of the system to work in a country with many languages is the major concern area of interest and the need to develop a system that will have higher accuracy based on Nigerian languages is very important because this system is intended to be used purposely Nigeria.

Below are relevant work from which the research gap was identified;

Muda *et al*, (2010), developed a voice recognition algorithm using MFCC and Dynamic Time Warping (DTW) techniques. Input voices from a male and female speaker were recorded for the training session and voice recognition process. The results obtained conformed to the reference templates stored in the database .

Kamruzzaman *et al,* (2010), presented a technique for speaker identification using MFCC and Support Vector Machines (SVM). The technique is text-dependent for speaker identification. 20 samples of the text 'zero' were used for 8 different speakers. Results obtained using Chunking and SMO training algorithm in SVM's showed that the success rates were 91.85% and 95% respectively.

Adikane *et al*, (2014), presented a speaker recognition approach using MFCC and Gaussian Mixture Model (GMM) with Expectation Maximization (EM) algorithm. The EM algorithm was incooperated in order to improve the performance of the GMM in building each speaker profile.

Mahboob *et al*, (2015), implemented a speaker identification system using GMM with MFCC. The system processing included feature extraction, training and matching. The system was designed based on the methodology of incremental model for designing, implementing, integrating and testing. Two samples each for eight speakers were collected for the testing. The obtained system efficiency was 87.5% with an error rate of 12.5%.

Ruhbami *et al*, (2014), presented a security system based on speaker identification using MFCC. The LBG-VQ algorithm was implemented to identify a speaker. A database of twenty one speakers was used for testing the system. The results showed that the identification rate increases as the codebook increases. Also, a combination of MFCC with Hamming window gave best performance.

Kekre *et al*, (2010), presented a novel method for speaker identification based on VQ. The system consists of two phases; traning and testing. Code vectors were generated for two scenerios; with overlap and without overlap. Sample speakers were tested for both textdependent and text-independent identification. Tests showed that code vectors with overlap gave better results.

Samudre *et al*, (2012), presented a text-independent speaker identification system based on VQ. The system incooperates mapping of speech signals from an unknown speaker to a database of known speakers. Based on a sample of eight speakers, the system was able to recognize seven of them correctly with an error rate of 12.5%.

Jhavan *et al*, (2014), presented a speaker Recognition System (SRS) using MFCC and VQ with Voice Activity Detection (VAD), which discriminates between silence and voice activity with aim of improving the performance of the SRS under noise condition. The results showed that increased number of centroids improves identification rate of the system, where also the VAD gives 5% reduction in error rate thus obtaining 95% recognition efficiency.

3.0 DESIGN AND SYSTEM IMPLEMENTATION

3.1 Database: The database for this project was initially an open source Database which contains the Census (AN4) database audio files and the design was tested on that Database but Based on Request by NCC that the Database must contain our Nigeria Local speaker at least male and female Hausa, Yoruba and Igbo. Then the database was finally change at the implementation stage to contain six speakers of which two are Hausa male and Female, two are Yoruba Male and female and two are Igbo male and female. The voices were recorded in wav format and each of length at most 5 second. Enough sample were taken from each of the speaker to enable us have enough voices for training, testing.

3.2 Feature Matching

Feature is the process of assigning speech signals of every customer a different class based on its feature. Features are taken from known samples and then unknown samples are compared with those known samples. Different techniques such as Mel frequency Cepstral, Neural Networks, Vector quantization Technique, Minimum distance classifier, Quadratic classifier, Bayesian classifier,, Correlation are used to achieve feature matching. For this research, MFCC was used to obtain a solution.

3.3 MFCC

Human voice contains large amount of information about a speaker such as his emotions, identity, language as well as the message communicated. Automatic speaker recognition (ASR) is a process of extracting the important voice features of an individual for the purpose of identification by analysis of his speech. The Mel Frequency Cepstral Coefficient (MFCC) introduced by Davis and Mermelstein in the 1980's is one of the most popular techniques used in speech signal processing applications for extraction of voice features. The principle of MFCC is based on the human hearing perception i.e. known variation of the human ear's critical bandwidth with frequency which makes it not to perceive frequencies above 1kHz. It consists of two filters, one spaced linearly at frequency below 1000Hz, while the other is spaced logarithmically above 1000Hz on the mel frequency scale. The block diagram of MFCC processor is depicted as in figure 1.0 below

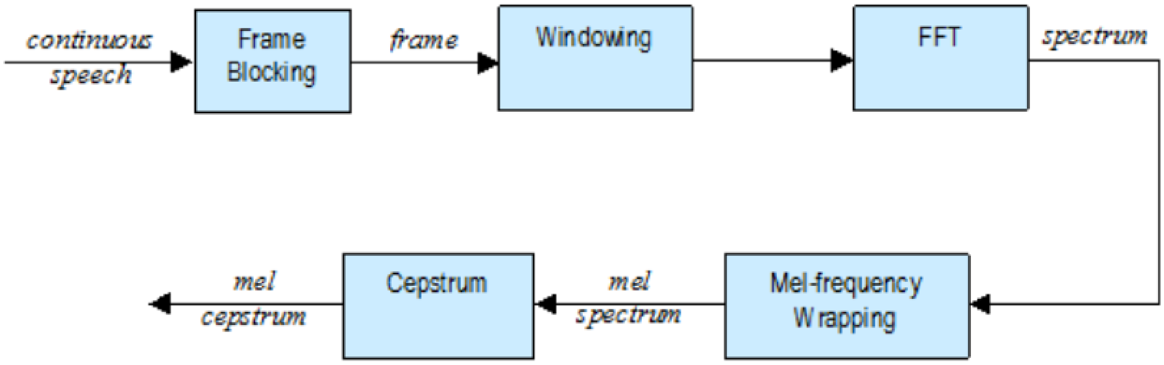


Figure 1.0: Block Diagram of MFCC Processor

An additional pre-processing step (Pre-Emphasis) was added to improve the performance of the MFCC.

3.31 Pre- Emphasis

This initial processing step involves passing the signal (i.e. voice) through a high pass filter. This tends to increase the amplitude of high frequency signal while decreasing the amplitude of low frequency signal thereby increasing the energy of the high frequency signal which is the most desired signal for extraction.

The signal obtained was represented as;

$Y(n) = X(n) \quad aX(n \quad 1)$ …………………………..Eq1

Where a lies between 0.95 – 1.0, meaning 95% of the samples are assumed to originate from previous samples

3.32 Framing;

This is the process of segmenting the speech samples into smaller frames of N samples with lengths within the limits of 20 to 40ms so as to make the signals feasible for Fast Fourier Transform (FFT). Adjacent frames are separated by M samples with M<N. Typical values used are M = 100 and N =256.

3.33 Windowing;

This signal processing step seeks to minimize signal discontinuities at the beginning and at the end of each frame i.e. minimizing spectral distortions by using a window to converge the signal to zero at the beginning and end of each frame. For a window W(n) defined at $0 \leq n \geq E\ 1$, where N is the number of samples in each frame, the windowing of a signal X(n) results to;

$$Y(n) = X(n)W(n), \qquad 0 \leq n \leq N$$ ……………………Eq2

The Hamming window is typically used because it provides a better frequency resolution by minimizing the signal boundaries to the closest side lobe.

Mathematically it was represented in equation 3 below

$$W(n) = 0.56 \quad 0.46\cos\left(\frac{2\pi n}{N-1}\right), \qquad 0 \leq n \leq N$$

……………Eq3

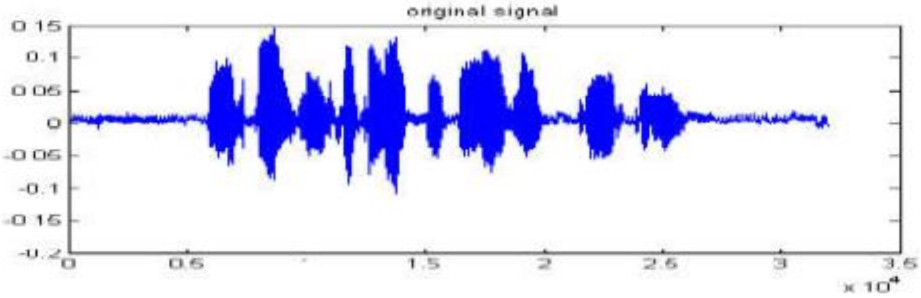


Figure 1.2: Speech Signal before Windowing

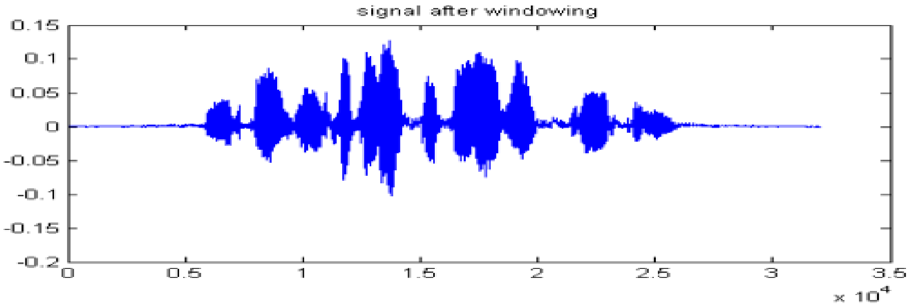


Figure 1.3: Speech Signal after Windowing

3.34 Fast Fourier Transform (FFT)

The Fast Fourier Transform (FFT) converts a signal in time domain to an appropriate representation in frequency domain so as to obtain the magnitude frequency response of frames which are assumed to be periodic and continuous.

FFT’s are fast algorithms implementing discrete Fourier transform (DFT) defined on N sample sets {x$_n$ } as;

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j\frac{2\pi KN}{N}} , \qquad k = 0,1,2 \ldots \ldots N$$ ……..Eq4

X $_k$ are mostly complex, therefore only the absolute values (frequency magnitudes) are considered.

The output sequence { x $_k$ } are interpreted as;

$$0 \le f < \frac{F_s}{2} \quad corresponding\ to \qquad 0 \le n \le \frac{N}{2}$$ ……………….Eq5

And

$$\frac{F_2}{2} < f < 0 \ corresponding\ to \qquad \frac{N}{2} + 1 \le \ n \le N$$ ………………..….Eq6

For positive and negative frequencies respectively.

These outputs of processing step (FFT) are called a spectrum or periodogram

3.35 Mel Frequency Wrapping

Based on psychological survey, the human notion of frequency contents of sound does not follow a linear scale. Therefore, for each tone with an actual frequency *f,* measured in Hz, the Mel scale is used to measure a subjective pitch. The Mel frequency scale is a linear frequency scale spacing below 1000Hz and logarithmically spaced above 1000Hz. A Mel is defined as one thousands of the pitch of 1KHz tone.

In order to simulate a subjective spectrum, a filter bank is used with one filter for each desired Mel frequency component to mimic the human ear.The filter bank has triangular bandpass frequency response with spacing and bandwidth determined by a constant Mel frequency interval k which has a typical value of 20.

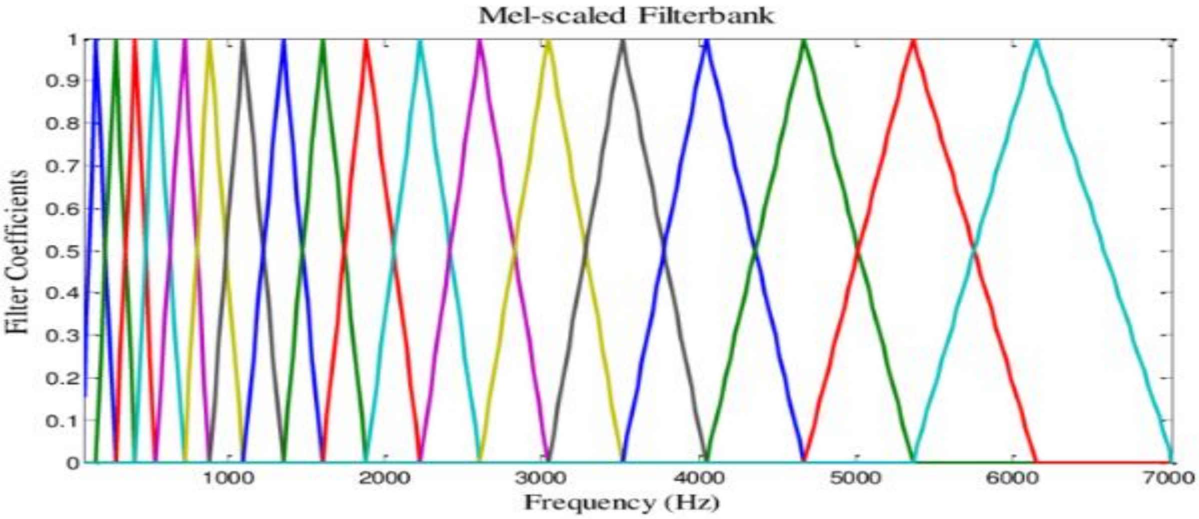


Figure 1.4 Mel sketch

The filter bank is equal to unity at the center frequency and linearly decreases to zero at the center frequency of two adjacent filters. The output of each filter is the sum of its filtered spectral components.

To compute the Mels for a particular frequency, the equation used is given as;

$$Mels(f) = [2596 \quad \log_{10}(1 + \frac{F}{7\ 00})]$$ ……………………Eq7

The Mel cepstrum estimates the spectral envelope of the output filter bank.

Speaker Model

After extracting features by MFCC, The speaker model was created using statistical model called GMM statistical model. The Gaussian distribution, is undoubtedly one of the most well-known and useful distributions in statistics, playing a predominant role in many areas of applications and hence it is a good candidate due to the enormous research effort in the past.

3.4 Gausian Model

A Gaussian mixture density is a weighted sum of M component densities given by:

$$p(x|\lambda) \sum_{i=1}^{M} pi\, bi(\vec{x})$$ ………………………………………Eq8

With the actual meaning of b_i as presented in equation

$$b_{i\ \vec{x}} = \frac{1}{(2\pi)\hat{}D/2|\sum i|\hat{}1/2} exp\left\{-\frac{1}{2}(\vec{x} - \overline{\mu i})' \sum_{i}^{-1} (\vec{x} - \overline{\mu i})\right\}$$ …..Eq9

Where b_i = density

pi= mixture weight

μ_i = mean vector

$\sum i$ = covariance matrix

In the GMM model the mixture weight must satisfy the constraint of equation 10 below. And also the Gaussian mixture density must be parameterized by mean vector, covariance Matrices and mixture weights from all component densities as presented in equation 11 below.

$$\sum_{i=1}^{M} p_{i=1}$$ ………………………………………..….Eq10

$$\lambda = \left\{ p_i \vec{\mu}_i \sum i \right\}$$ ………………………………..Eq11

This Eq 11 run from 1 to M as presented in figure 1.5 below and each speaker is represented a GMM which is referred to the speaker model both female and male and represented by model λ

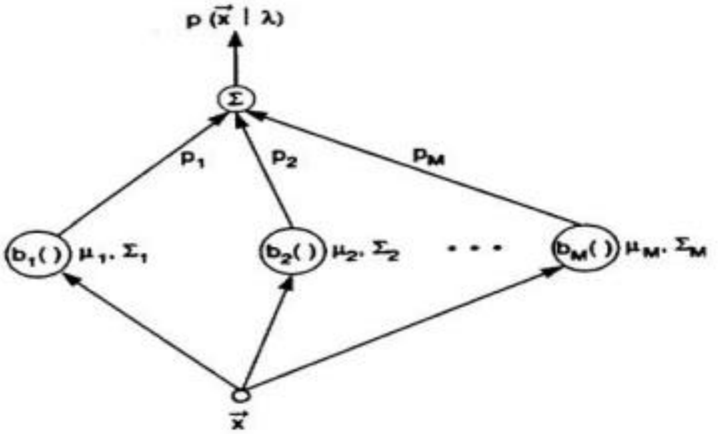

$p(\vec{x} \mid \lambda)$
Σ
p_1
p_2
p_M
$b_1(\,)$ μ_1, Σ_1
$b_2(\,)$ μ_2, Σ_2
• • •
$b_M(\,)$ μ_M, Σ_M
$\vec{x}$

Figure 1.5: Gaussian Mixture model as weighted sum of Gaussian Densities

3.41 Maximum Likelihood Estimation Model

The goal of the speaker model training is to estimate the parameter of the GMM λ, and the most popular method for training GMM is a maximum likelihood estimation which is presented below in Eq12 which was used for estimation of the model parameter which normally maximize the likelihood of the GMM given the training data.

$$P({}^{x}/_{\lambda}) = \prod_{t=1}^{T} P({}^{\overrightarrow{xt}}/_{\lambda})$$

…………………………..…Eq12

Maximization of the quantity in (12) is accomplished through running the expectation-maximization algorithm. The idea is *beginning with an initial model λ, to* estimate a new *model λ* satisfying **p(X/λ*)* >=(X/λ)**. The new model then becomes the initial model for the next iteration and the process is repeated until some convergence threshold is reached. Following formulas are used on each EM iteration.

Weighted mixture in in Eq13, Mean in Eq14 and Variance in Eq 15 below respectively.

$$\overrightarrow{p\iota} = \frac{1}{T}\sum_{t=1}^{T} Pi(i/\overrightarrow{x\iota}, \lambda)$$

……………..……Eq13

$$\overrightarrow{\mu\iota} = \frac{\sum_{t=1}^{T} P(i/\vec{x}, \lambda)\overrightarrow{xt}}{\sum_{t=1}^{T} P(i/\vec{x}, \lambda)}$$

………………………..Eq14

$$\overrightarrow{\sigma\iota^{2}} = \frac{\sum_{t=1}^{T} P(i/\vec{x}, \lambda)\overrightarrow{xt\hat{}2}}{\sum_{t=1}^{T} i/\vec{x}, \lambda)} - \mu\hat{}2_{i}$$

……………..Eq15

While the probability for acoustic class is given in Eq 16 below

$$P\left(i\middle|\overline{\vec{xt}},\lambda\right) = \frac{Pibi\,(\vec{x}t)}{\sum_{k=1}^{M} P_k\, b_k\,(\vec{x})}$$ ……………Eq16

3.42 Speaker Identification

The Identification or detection of speaker is the process of identifying the target speaker based on his identity and it was achieved by using Equation 17 to 19.

$$\hat{S} = Arg\max P(\lambda k \,/\, X) = Argmag\frac{P(X \,/\, \lambda k)P(\lambda K)}{p(X)}$$ 17

$$\hat{S} = Arg\max\sum_{t=1}^{T}\log P(xt \,/\, \lambda k)$$ …..........................18

$$\hat{S} = Arg\max\sum_{t=1}^{T}\log P(xt \,/\, \lambda k)$$ 19

The performance error was evaluated using equation 20 below

IER (%) = $\frac{Number\ of\ incorrect\ Vector}{Total\ Number\ of\ Vector}$*100…………………Eq20

And finally the speaker Identification/Recognition was achieved using figure 1.6 below. The system was modeled to be able to detect whether the speaker is male or female so that the detection will be gender based as that improves the accuracy of the system. The complete implementation is as presented in figure 1.6 below.

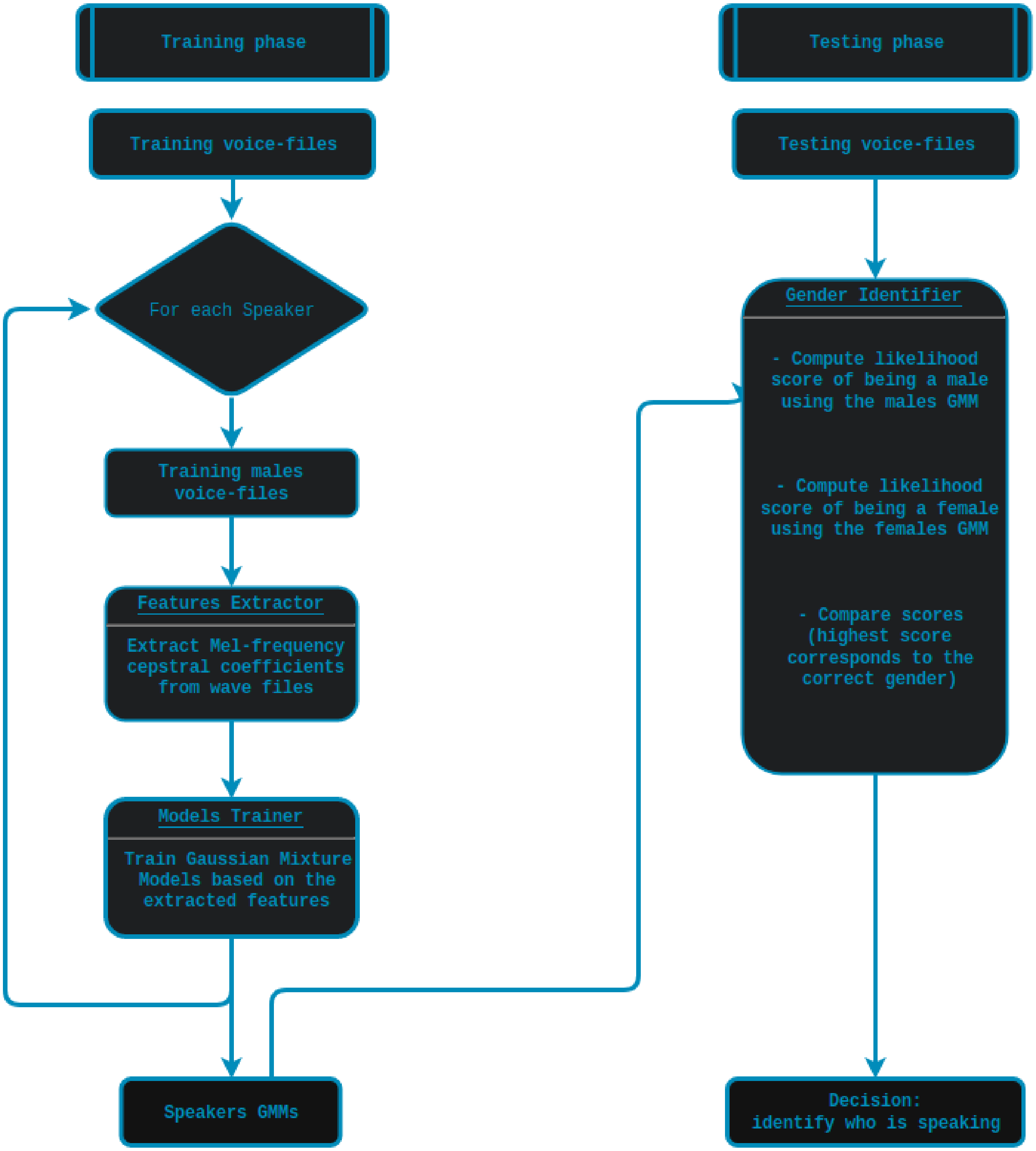

Training phase
Training voice-files
For each Speaker
Training males voice-files
Features Extractor
Extract Mel-frequency cepstral coefficients from wave files
Models Trainer
Train Gaussian Mixture Models based on the extracted features
Speakers GMMs
Testing phase
Testing voice-files
Gender Identifier
- Compute likelihood score of being a male using the males GMM
- Compute likelihood score of being a female using the females GMM
- Compare scores (highest score corresponds to the correct gender)
Decision: identify who is speaking

Figure 1.6 Speaker Recognition Implementation

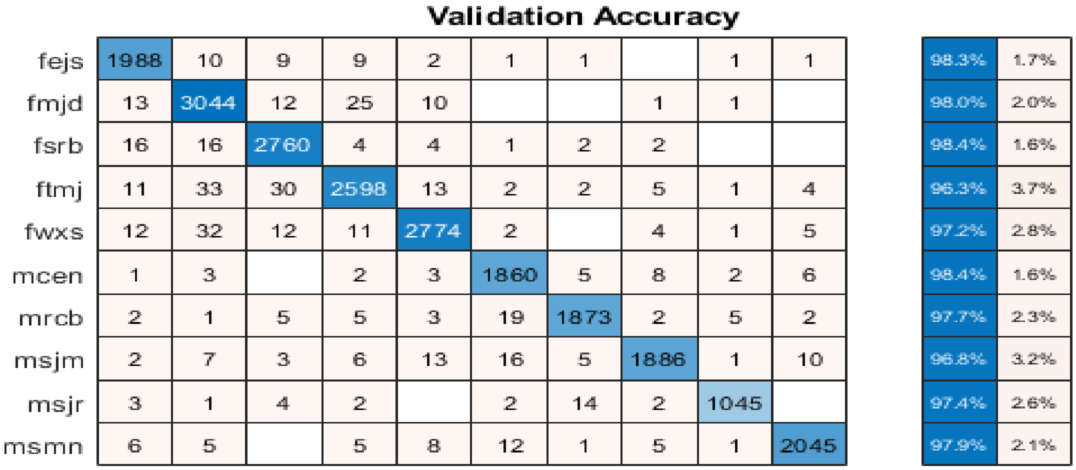
3.5 TEST AND VALIDATION OF RESULT

The result was tested and validated on Matlab from An4 database and the result was also confirmed from the sample of at least two speakers male and female in each of the following Nigerian languages; Hausa, Yoruba and Igbo as directed by NCC.

Validation Accuracy from An4 on Matlab

Speakers name in the An4 database were; fejs, fmjd, fsrb, ftmj, fwxs, mcen, mrcb, msjm, msjr and msmn

A validation accuracy of 98% was obtained using confusion matrix as can be seen in figure 1.7 below

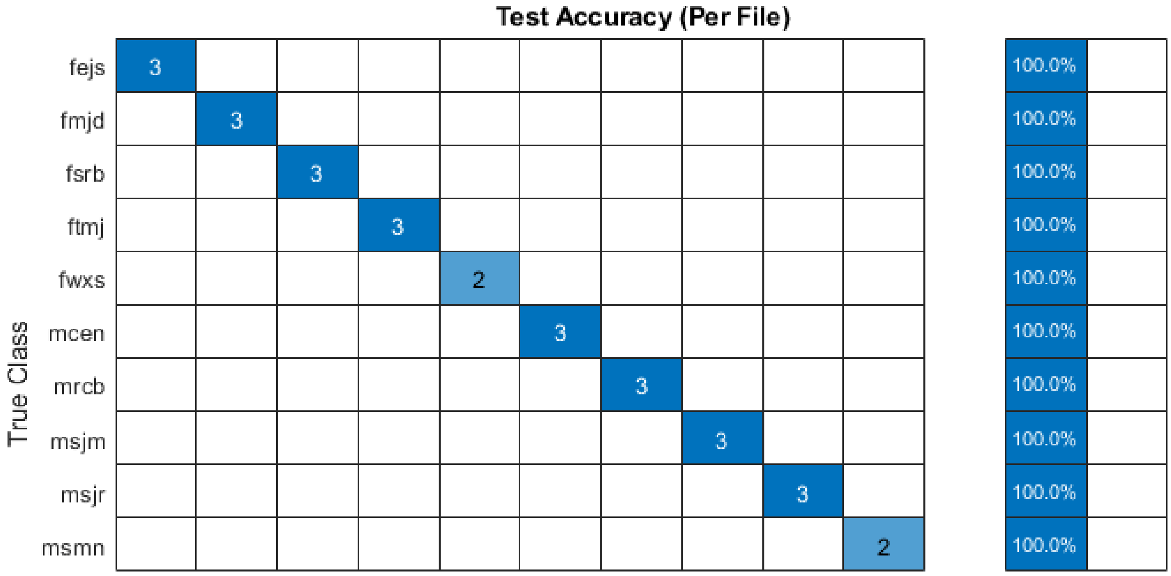


Validation Accuracy

fejs	1988	10	9	9	2	1	1		1	1	98.3%	1.7%
fmjd	13	3044	12	25	10			1	1		98.0%	2.0%
fsrb	16	16	2760	4	4	1	2	2			98.4%	1.6%
ftmj	11	33	30	2598	13	2	2	5	1	4	96.3%	3.7%
fwxs	12	32	12	11	2774	2		4	1	5	97.2%	2.8%
mcen	1	3		2	3	1860	5	8	2	6	98.4%	1.6%
mrcb	2	1	5	5	3	19	1873	2	5	2	97.7%	2.3%
msjm	2	7	3	6	13	16	5	1886	1	10	96.8%	3.2%
msjr	3	1	4	2		2	14	2	1045		97.4%	2.6%
msmn	6	5		5	8	12	1	5	1	2045	97.9%	2.1%

Figure 1.6 Confusion Matrix for Validation of Result

And also the test accuracy was presented in Figure 1.8 below which confirm the test accuracy of 100%



fejs	3										100.0%	
fmjd		3									100.0%	
fsrb			3								100.0%	
ftmj				3							100.0%	
fwxs					2						100.0%	
mcen						3					100.0%	
mrcb							3				100.0%	
msjm								3			100.0%	
msjr									3		100.0%	
msmn										2	100.0%	

Figure 1.7 Confusion Matrix for Test Accuracy

But all these tests result does not make any sense because the aim of the project grant was not to end up with a laboratory result but instead is to end of with a physical system that can be used for speaker Identification. Being this research was conducted in Matlab and being a Matlab is mainly a laboratory software not implementation software therefore the codes was set to work on python 3.10 so that the commission will justify the functionally of the developed project in reality.

5.1 Physical Implementation Test.

Block diagram of Figure 1.6 was implemented on python and we were able to test the code based on reality. The voices of Yoruba, Igbo and Hausa generated on the database were then used to train the model and we were able to recognize the voice of any speaker from the stored speaker model. The claimed laboratory test accuracy was reconfirmed. The system is available for further test.

CONCLUSION

At the end of this research the aim of the project was achieved because speaker identification was developed and the system can be used for customer identification by their voices in telecommunication industry with higher accuracy.

RECOMMENDATION

At the end of this project the research team recommended that NCC should be able to test this project on telecommunication database and on larger database. It was also recommended that NCC should check the limitation of this project based on telecommunication data and finally we recommended that for larger database, higher specification Computer should be used especially in term of processor speed.

REFERENCE

Darshan M., Pravin G., and Rachna S., (2011). " Speaker Recognition using MFCC and Vector quantization model" Department of Electronics and Communication Engineering, NirmaInstitute of Technology Ahmedabad.

Kishore P., (2008). "Speech Technology: A Practical Introduction", Carnegie Mellon University and International Institute of Information Technology Hyderabad.

Laura L., (2014). "Human and Automatic Speaker Recognition over Telecommunication Channels" PhD Thesis Canberra University Australia.

Douglas A. Reynold (M.I.T. Lincoln Laboratory) and Larry P. Heck (Nuance Communications) " Autimatic Speaker Recognition Current Application and Future Trends" presented at AAAS meeting : Humans, Computer and Speech Symposium 19thFebuary, 2000.

Muda L., Begam M. and Elamvazuthi I., (2010). "Voice Recognition Algorithms using MelFrequency Cepstral Coefficient (MFCC) andDynamic Time Warping (DTW) Techniques",*Journal of Computing,* VOL. 2,ISSN\2151-9617.

Kamruzzaman S. M., Rezaul K., Saiful Islam M. D.and Emdadul Haque M. D., (2010). "Speaker Identification using MFCC-Domain Support Vector Machine", Department of Information and Communication Engineering University of Rajshahi, Bangladesh.

ABDUL SYAFIQ B. A., (2012). "Speaker Identification System using MFCC Procedure and Noise Reduction Method", Master of Electrical Engineering Report, Faculty of Electric and Electronic Engineering University Tun Hussein Onn Malaysia.

Adikane A., Moon M., Dehankar P., Borkar S. and Desai S., (2014). "Speaker Recognition Using MFCC and GMM with EM"*International Journal of Engineering Research and Applications (IJERA), ISSN: 2248-9622.*

Mahboob T., Khanum M., Hayat Khiyal M. S. and Bibi R., (2015). "Speaker Identification Using GMM with MFCC" *International Journal of Computer Science(IJCSI),* Vol.12, ISSN (Print): 1694-0814.

Yadav K. andMukhedkar M., (2014)."MFCC Based Speaker Recognition using Matlab", *International Journal of VLSI and Embedded Systems (IJVES),* Vol. 5ISSN: 2249 – 6556.21

Sajjad A., Shirazi A., Tabassum N., Saquib M. and Sheikh N., (2017). "Speaker Identification and Verification Using MFCC and SVM", *International Research Journal of Engineering and Technology (IRJET),* Vol.4,ISSN: 2395-0072.