IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEE)
e-ISSN: 2278-1676,p-ISSN: 2320-3331, Volume 13, Issue 5 Ver. I (Sep. – Oct. 2018), PP 62-70
www.iosrjournals.org

Nigerian Language Simulated Speaker Verification System Using Back-Propagation Neural Network

Shakiru Olajide KASSIM[1], Paul InuwaADAMU[2], Hamza ABBA[3], Mohammed ABDUL-FATAU[4]

[1 & 2]Department of Computer Engineering Technology
The Federal polytechnic Damaturu, Yobe State.
[3]Department of Electrical/ElectronicEngineering Technology.
The Federal polytechnic Damaturu, Yobe State.
[4]Department of Electrical/Electronic Technology.
The Federal College of Education (Technical) Potiskum, Yobe State.

Abstract: *This paper presents the use of Back propagation neural network to implement speaker verification simulated with Nigeria Languages. The focus is to verify voice patterns for different people speaking different Nigeria languagesso as to recognize (verify) their speech electronically.The voice samples of the people utilized were captured and then processed using the sound forge 9.0 software.The frequencies of each voice signals were used to train a backpropagation neural network, which inturn verifies the speakerthrough the voice patterns. Six neural networks (K_{1-1}, K_{2-1}, Y_{1-1}, Y_{2-1}, F_{1-1}, and F_{2-1})were developed for training, testing, and validation ofthree selected Nigeria languages words from nine (9) people (three for each language). Discrete Fourier Transform (DFT) was usedto extract features from the voice samples for the backpropagation neural network (BPNN) training, testing, and validation. The network's performance analysis results as deduced from regression analysis showsan averageoverall R-value of 0.9371 for acceptance, and an average of 0.3414 for rejection. The results obtained shows that each network verified the speaker it was trained for adequately. The results obtained from this work can be generalized to cater for larger vocabularies and for continuous speaker verification processes.*

Keywords: *Backpropagation neural network (BPNN), speaker verification, Discrete Fourier Transform (DFT), feature extraction, regression analysis.*

Date of Submission: 10-10-2018 Date of acceptance: 26-10-2018

I. Introduction

To communicate with each other, Speech is probably the most efficient way. It is possible to use speech as a useful interface to interact with machines(Ali, Mijanur, Uzzal, & Farukuzzaman, 2013).Voice recognition technology is used to verifya speaker's identity or determine an unknown speaker's identity. Speaker verification and speaker identification are both common types of voice recognition(Kassim & Anene, 2015).Fig. 1 shows the block diagram of a typical voice recognition system.

Speaker verification is the process of analyzing an individual's voice pattern with the aim of confirming the identity of the speaker. In other words, Speaker verification is the process of using a person's voice to verify that he is who he say he is. (Kassim & Anene, 2015). While, Speaker identification is the process of determining an unknown speaker's identity. It is the process of using recorded speech in an attempt to identify the individual speaking. This form of technology is most commonly used in criminal investigations and is often carried out in secret.Unlike speaker verification, speaker identification is usually covert and done without the user's knowledge. The system can help to identify individuals who may have undergone physical surgery to alter outward appearances(Patricia, Jerica, Daniel, Miguel, Miguel, & Oscar, 2006).

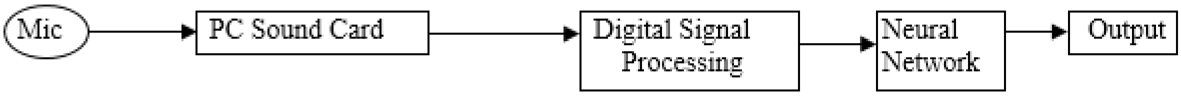


Figure 1: Block Diagram of Voice Recognition System(Kassim & Anene, 2015)

Speaker recognition methods can also be divided into text-dependent and text-independent methods. The former require the speaker to say key words or sentences having the same text for both training and recognition trials. Whereas the latter does not rely on a specific text being spoken, that is, it is not necessarily the same training text that will be used for the recognition session, (Ibikunle & Katende, 2013). Both systems perform the following tasks: feature extraction, similarity analysis and selection (pattern matching)(Kassim & Anene, 2015).

A Neural Network is an information Processing Paradigm and it is stimulated based on the way biological nervous system works, i.e. like the way the brain process information. Simple computational elements operating in parallel are included in neural networks. The network function is determined largely by the connections between elements. A neural network can be trained so that a particularinput guides to a specific target output (Ali, Mijanur, Uzzal, & Farukuzzaman, 2013). Neural Network applications are numerous, examples include in speech recognition, optical character recognition (OCR), modelling human behaviour, classification of patterns, loan risk analysis, music generation, image analysis, creating new art forms, stock market prediction, etc. In this work, Multilayer Feed-forward Network with Back-propagation algorithm is used to verify text dependent speech for Nigeria languages.

The basic problem with Speech/Voice recognition is identification of proper features for the recognition task, and the strategy to extract these features from the signal. Feature extraction involves analysis of signal. Broadly, feature extraction techniques are classified as temporal analysis and spectral analysis technique. In temporal analysis the waveform itself is used for analysis, while in spectral analysis, spectral representation of the signal is used for analysis.

The goal of the work is to recognize users from their voice signal patterns (samples). Using the procedure where all of them(users) utters the same Nigeria Language word (text-dependent)for verification purposes. The effect of age and sex was also considered to see how it affects the recognition processes. According to *(Desmond & Graham, 2007)*, an increase in energy at the fundamental of a voice signal will result in a corresponding high increase in the spectrum. These indirectly is influenced by age and sex. The speech signals corresponding to a text phrase of the users was recorded as voice files on a computer system using sound recording software. The time valued information is converted to their frequency components using digital signal processing techniques (DFT of the MATLAB environment) andthen used to train a backpropagation neural network whose output is expected to verify of the user (either accepted or rejected) depending the regression value.

II. Methodology

In general,the methodologyinvolves the following steps:

i. Voice capturing and processing;
ii. Signal pre-processing (feature extraction);
iii. Training the neural network (pattern matching) and;
iv. Testing the Neural Network with trained and untrained voice samples (output).

2.1 Voice Capturing and Processing

The voice signal of the users was captured using the in-built microphone in a personal computer through the record pad sound recording software, and was edited using the wave pad sound editor software, and then saved as a sound file(at fixed time of recording for signals homogeneity).

The human hearing and voice range is about 20 Hz to 20 kHz, and the most sensitive is within the range of 2 to 4 kHz(John & Dimitris, 1996). According to the Nyquist rate, the sampling rate, $\boldsymbol{f_s}$, should be twice the maximum frequency, F_{max}, (i.e. $f_s = 2\ F_{max}$). Thus, after capturing the sound signals, the voice signals are digitized at 8 kHz sampling frequency, yielding a signal with 8192 sample points after pre-processing. The sampled voice signal of one of the speaker is given in Fig. 2.

2.2 Signal Pre-processing (FeaturesExtraction)

The information content of voice signals is contained in the amplitudes, frequencies and phases of the various frequency components, but the detailed knowledge of the characteristics of such signals is not available prior to obtaining the signals. In fact, the purpose of the pre-processing is to extract this detailed information, hence the need for signal pre-processing(John & Dimitris, 1996).

Human speech can sensibly be interpreted using frequency energy interpretation such as a spectrogram. Frequency-energy interpretations and power spectral densities can be used to differentiate between the speakers (Kassim & Anene, 2015).

To achieve the extractionof the features in the sample voice signals, the MATLAB signal processing toolbox was duly employed. The two major steps followed are as outlined:

2.2.1 Determination of the Fourier transforms of the voice samples

Generally, the Fourier transform of discrete signal is given by the relationship:

$$X(mW_s) = \sum_{k=0}^{N-1} x_k e^{-jm\,w_s k} \;\; ; \; w_s \triangleq \frac{2\pi}{N} \qquad \ldots (1)$$

The discrete Fourier Transform (DFT) when computed shows the spectral peak present in the voice samples. The MATLAB ***fft function***is used to perform the computation of the DFT efficiently. This converts the vector x in the time domain samples (sampled voice signals), to a vector X of samples in the frequency domain(Andrew, 2000).

2.2.2.1 Determination of the absolute values of Fourier transforms of the voice samples.

The absolute values of Fourier transforms of the voice samples was determined by taking the absolute value of the vector X of samples in the frequency domain.According to (Furui, 1986), theabsolute values contain unique features attributed to an individuals. These are presented to the pattern matching network for the verification required.

2.3Training the neural network (pattern matching)

A two layer feed-forward neural network with a sigmoidal layer followed by a linear layer was employed for the pattern matching. The neural network was trained using a supervised resilience back propagation algorithm with a momentum term, because of its versatility in handling data for pattern recognitionand its ability to achieve a faster global convergence (Kassim & Anene, 2015).

2.4Data and Network Pre-Processing for Training

The training of the neural network requires proper processing of the voice samples and selection of suitable parameters for the networks to be used.

2.4.1Data Division for Optimal NN Training

For optimal NN training, each of the data (voice samples) for training of the BPNN are divided into three subsets:***training set*** for computing the gradient and updating the network weights and biases;***validation set*** for basicallymonitoringerror on the validation set during the training process; and the ***test set***for comparing different models and also useful to plot the test set error during the training process. The ***"dividerand"*** function for dividing each voice signal with 8192 samples, as a consequence, there are5734 training samples, 1229 samples each for testing and validation.

2.4.2Data categorization

The voice samples captured from different speakers are categorized in accordance to their ages (old, young and teenage) with special consideration of sex. These are then grouped into the input data and target data for training. Table 1 shows the categorization of the sampled voice signals.

2.4.3 Network creation for NN training

The networks used for the training process of the NN were created from the recorded voice signals (data), two (2) each (making a total of six (6)) was created for the three Nigerian languages:"Yoruba", "Kanuri" and "Fulani" with spoken words "LA'AKAYE", "KAWU'SKE" and "ARDUNGAL" respectively. Table 2 shows data K_1 and K_2; K_3 and K_4; Y_1 and Y_2; Y_3 and Y_4: F_1 and F_2; and, F_3 and F_4which are signals from same speaker. But K_5, Y_5 and F_5 are from different **speakers**. Each network created has input data and target data. The input data is the actual data presented to the network, while target data define the desired network outputs. The response of the network to its inputs relative to the target data gives the output data.

Table 1: Categorization of the Sampled Voice Signals

Voice samples	Category of speaker sampled (in terms of age and sex)
E_1 and E_2	Old Man above 50 years
E_3 and E_4	Old Woman above 50 years
E_5 and E_6	Young man between 20 and 40 years
E_7	Young Lady between 20 and 40 years
K_1 and K_2	Young man between 20 and 40 years
K_3 and K_4	Young Lady between 20 and 40 years
K_5	Teenage Girl between 10 and 20 years
Y_1 and Y_2	Old Man above 50 years
Y_3 and Y_4	Old Woman above 50 years
Y_5	Young man between 20 and 40 years
F_1 and F_2	Teenage Girl between 10 and 20 years
F_3 and F_4	Teenage Boy between 10 and 20 years
F_5	Young Lady between 20 and 40 years

Table 2: Network Created for NN Training

S/No.	Network	Target data	Input Data
1.	K_{1-1}	K_2	K_1
			K_3
			K_5
2.	K_{2-1}	K_4	K_1
			K_3
			K_5
3.	Y_{1-1}	Y_2	Y_1
			Y_3
			Y_5
4.	Y_{2-1}	Y_4	Y_1
			Y_3
			Y_5
5.	F_{1-1}	F_2	F_1
			F_3
			F_5
9.	F_{2-1}	F_4	F_1
			F_3
			F_5

2.5 Network Modelling

In this work, a three layer (multi-layer perceptron) feed-forward NN was used for the verification processes. This consists of an input layer, hidden layers and an output layer of neurons. The actual modelled networks for NN training is shown in Fig. 2; L_1, L_3, & L_5 represent the inputs with any of the three Nigerian languages used.

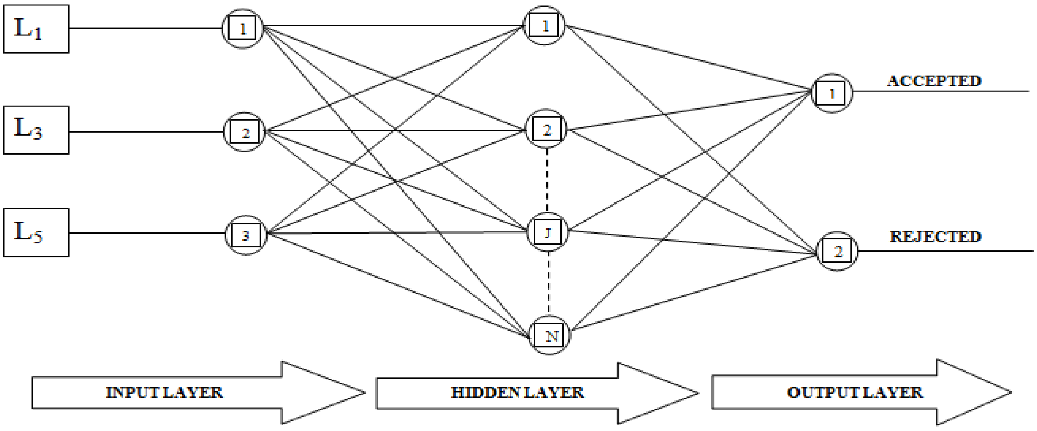


Figure 2: Modelled NN TrainingNetwork for the Three Nigerian Languages.

2.5 Network Performance and Training Analysis

Most researchers uses the mean square error (mse) to determine theirnetwork performance; and was reported to be a good tool for determining the network performance(Ali, Mijanur, Uzzal, & Farukuzzaman, 2013; Ibikunle & Katende, 2013; Kassim & Anene, 2015; Patricia, Jerica, Daniel, Miguel, Miguel, & Oscar, 2006; Kaustubh & Vijay, 2015). Hence, it was used in this work. Themse is the average squared error between the network outputs, **a**, and the target outputs, which is defined as follows:

$$F = mse = \frac{1}{N}\sum_{i=1}^{N}(e_i)^2 = \frac{1}{N}\sum_{i=1}^{N}(t_i - a_i)^2 \qquad \text{... (2)}$$

The performance of a trained network can be measured to some extent by the errors on the training, validation, and test sets(Hagan, Demuth, && Beale, 2014). In order to investigate the network response, regression analysis which investigates the relationship between the targets and the expected network output responses in relation to the inputs presented to network for training was also considered. The regression plot in the neural network training platform in MATLABwas used to generate the regression coefficients and corresponding plots. From the plot, the R-value (regression coefficients), slope and intercept of the best linear regression which relates the targets to the network outputs are given. The R-value is bounded between 0 and 1 (i.e. 0 ≤ R ≤ 1). R = 1 implies a perfect acceptance (good recognition), while R = 0 indicates total rejection (poor recognition).

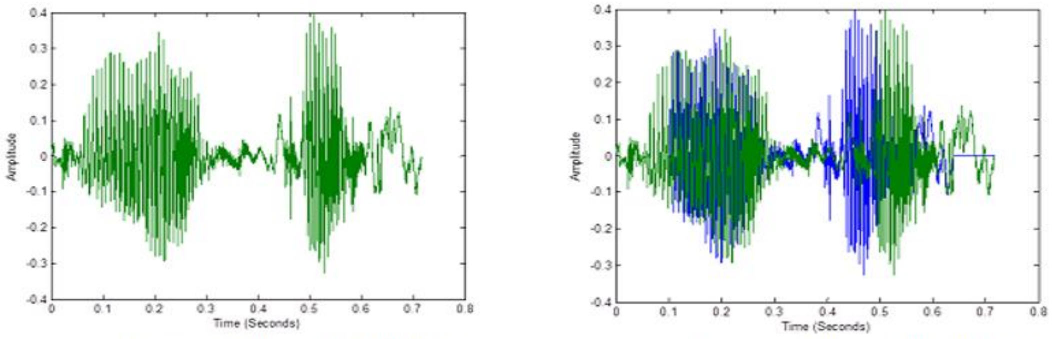
III. Results and Discussions

3.1 Results from Voice Capturing and Processing

Fig. 3 – 5 shows some of the resulting plots of the processed(using the sound forge 9.0 computer software) voice samples of the speakers using the MATLAB command:

```
>> l = data;                    representing the recorded voice sample
          >>fs = 8000;                    sampling frequency
          >> t = (0:1:length (l) - 1)/fs;                  specification of the time axis division
          >>plot (t, l);                  signal plot of the voice sample.
          >>title ('Sampled Voice signal of Speaker L');        plot title
          >>xlabel ('Time(Seconds)'); x-axis labeling
          >>ylabel ('Amplitude');   y-axis labeling
```

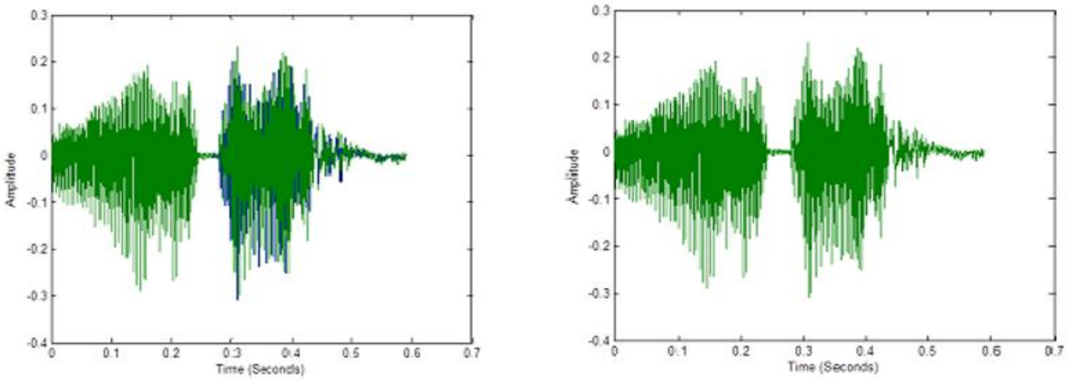
Where: l = represents the voice samples for the languages.
fs = 1/Δ which is the sampling frequency.



Speaker K_1 uttering the word "KAWUSKE"

Speaker K_2 uttering the word "KAWUSKE"

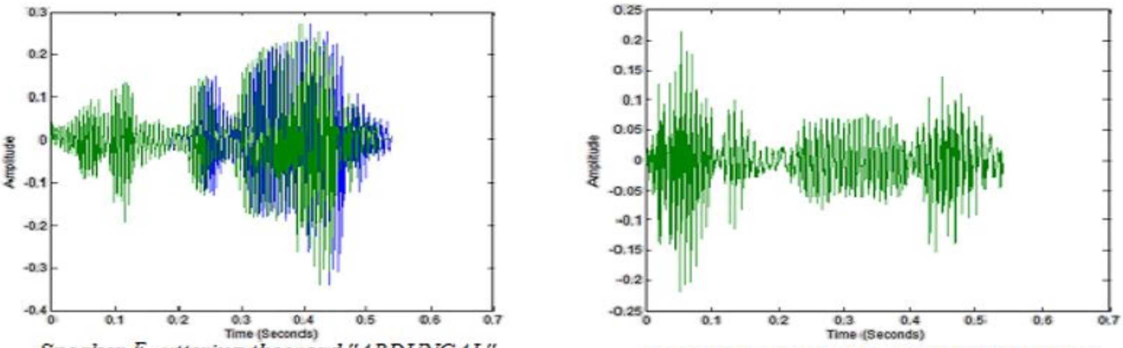
Figure 3: Sampled Voice Signal for Kanuri Speakers.



Speaker Y_3 uttering the word "LA'AKAYE"

Speaker Y_4 uttering the word "LA'AKAYE"

Figure 4: Sampled Voice Signal for Yoruba Speakers.



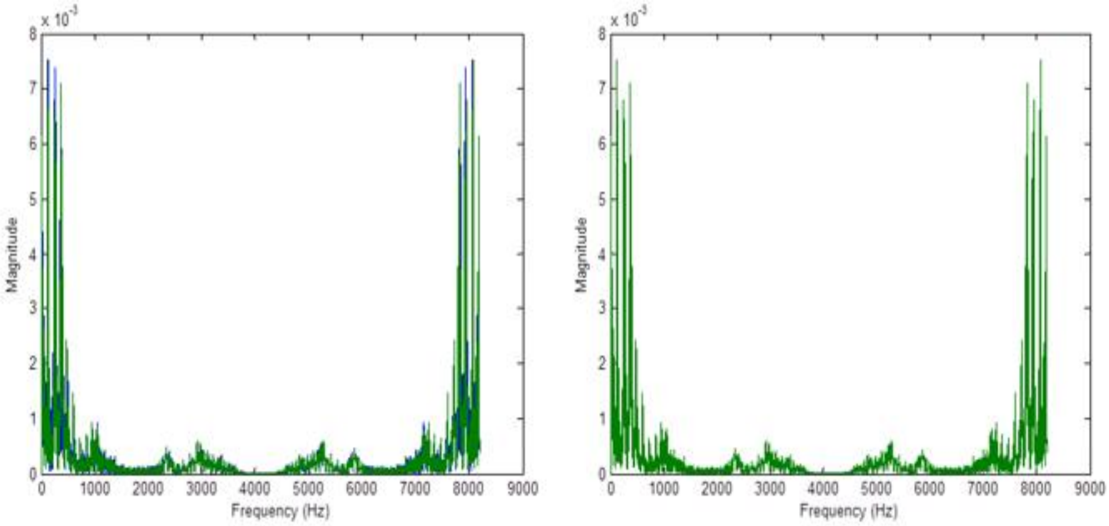
Speaker F₂ uttering the word "ARDUNGAL"   *Speaker F₃ uttering the word "ARDUNGAL"*

Figure 5: Sampled Voice Signal for Fulani Speakers.

3.2 Signal Pre-Processing Results

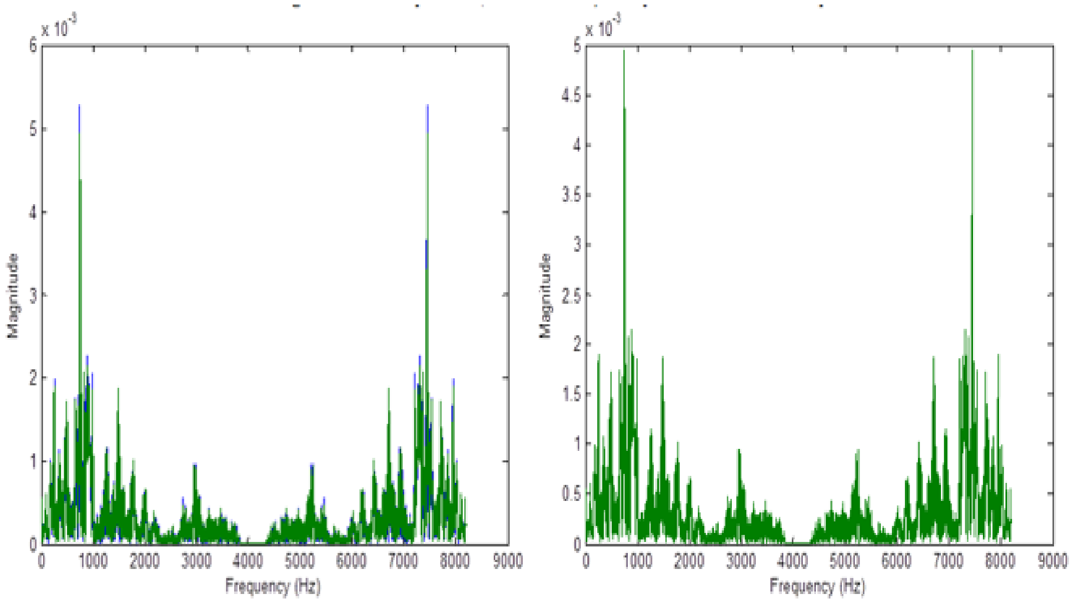
Fig. 6 – 8 shows the plots of the absolute values of DFT computations for the sample voice signalsin Fig. 3 - 5 respectively, using the MATLAB commands given below.

```
>> l = data;                    representing the recorded voice sample
        >>fs = 8000;                    sampling frequency
        >>N = fs;                       specifies the signal length
        >>NFFT = 2^nextpow2 (N);        next power of 2 from length of signal
        >> X = fft (l, NFFT)/N;         Discrete Fourier Transform of the vector x
        >> plot (abs(L));plot the absolute value of DFT.
        >>title ('DFT Computation (Absolute Value)');       plot title
        >>xlabel ('Frequency (Hz)'); x-axis labeling
        >>ylabel ('magnitude');   y-axis labeling
```



DFT Computation (Absolute Values) Plots for the Signals K₁   *DFT Computation (Absolute Values) Plots for the Signals K₂*

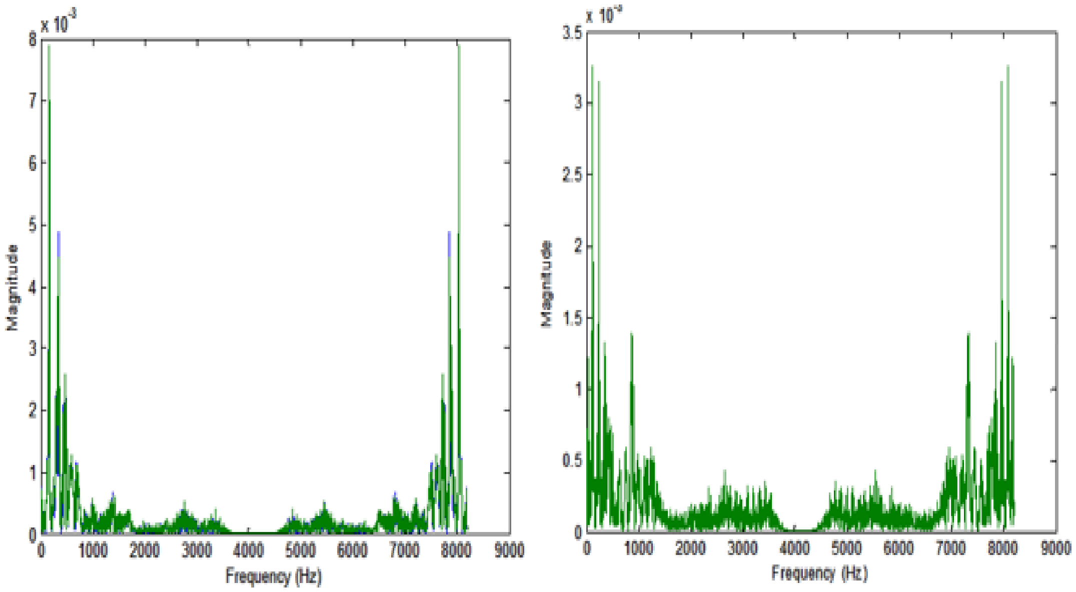
Figure 6: DFT Computation (Absolute Values) plots for Kanuri Voice Samples.



DFT Computation (Absolute Values) Plots for the Signals Y_3

DFT Computation (Absolute Values) Plots for the Signals Y_4

Figure 7: DFT Computation (Absolute Values) plots for Yoruba Voice Samples.



DFT Computation (Absolute Values) Plots for the Signals F_2

DFT Computation (Absolute Values) Plots for the Signals F_5

Figure 8: DFT Computation (Absolute Values) plots for Fulani Voice Samples.

3.3 Networks training results

The results obtained from the networks training are summarised in three different tables. The mean square errors (mse) and the regression coefficients for the training, testing and validation data are contained in the Tables 3, 4, and 5.

Table 3 shows the training results of voice signals samples for the word “KAWUSKE”.

Table 4 shows the training results of voice signals samples for the word “LA’AKAYE”.

Table 5 shows the training results of voice signals samples for the word “ARDUNGAL”.

These results are deduced from the trained network performance parameters which include the plot for the performance progress (which is the ***MSE***) and the regression coefficient plots.

Table 9: Training Results of Voice Samples (Kanuri Word)

.	Mean Square Error (MSE)			No. of Epoch	Regression			
	Training	Test	Validation		Training	Test	Validation	Overall
K_{1-1}	2.467 e-8	2.123 e-8	2.525 e-8	34	0.9780	0.9848	0.9793	0.9793
	1.051 e-7	9.833 e-8	1.033 e-7	40	0.5391	0.5128	0.5507	0.5371
	4.672 e-7	4.598 e-7	3.937 e-7	46	0.4640	0.5254	0.5069	0.4799

K_{2-1}	4.690 e-7	5.097 e-7	3.828 e-7	73	0.4799	0.3599	0.5182	0.4674
	3.488 e-8	3.358 e-8	2.425 e-8	12	0.8769	0.8820	0.8989	0.8804
	1.059 e-7	1.110 e-7	1.073 e-7	58	0.5201	0.5155	0.5122	0.5178

Table 10: Training Results of Voice Samples (Yoruba Word)

Networks	Mean square error (MSE)			No. of epoch	Regression			
	Training	Test	Validation		Training	Test	Validation	Overall
Y_{1-1}	4.179 e-9	4.439 e-9	3.848 e-9	10	0.9530	0.9580	0.9484	0.9534
	5.836 e-8	5.158 e-8	4.335 e-8	21	0.4955	0.5216	0.5505	0.5064
	3.881 e-8	3.740 e-8	4.009 e-8	8	0.3918	0.3766	0.3958	0.3901
Y_{2-1}	3.329 e-8	4.010 e-8	4.131 e-8	5	0.4844	0.4678	0.4304	0.4730
	8.304 e-12	1.351 e-11	2.274 e-11	12	0.9999	0.9999	0.9998	0.9999
	6.416 e-8	6.391 e-8	6.769 e-8	5	0.3677	0.3441	0.3263	0.3579

Table 11: Training Results of Voice Samples (Fulani Word)

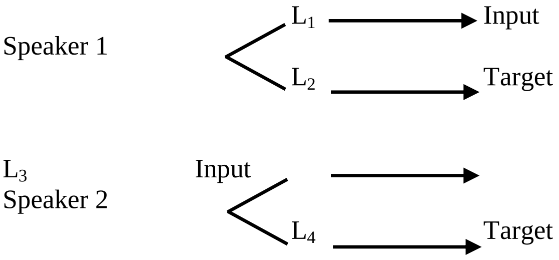
Networks	Mean square error (MSE)			No. of Epoch	Regression			
	Training	Test	Validation		Training	Test	Validation	Overall
F_{1-1}	1.085 e-8	1.269 e-8	1.051 e-8	14	0.9778	0.9794	0.9779	0.9780
	3.120 e-7	3.628 e-7	3.303 e-7	60	4.4643	0.4340	0.5025	0.4666
	2.299 e-7	3.082 e-7	1.402 e-7	27	0.3160	0.2811	0.3930	0.3185
F_{2-1}	2.078 e-7	2.717 e-7	2.634 e-7	13	0.3455	0.2677	0.3485	0.3330
	3.401 e-8	3.268 e-8	3.744 e-8	11	0.9569	0.9587	0.9474	0.9555
	2.907 e-7	3.159 e-7	2.935 e-7	53	0.5353	0.5140	0.5444	0.5334

Results Discussion

From the results obtained from entire system implementation, discussions are presented on the deductions made at each stage with respect to the response of the system.

3.4.1 *Network performance*

For the Nigerian language networks, two networks are developed for each of the three languages used. In each of the languages, two of the three speakers uttered the same word twice, for the inputs and target signals as follows:



The remaining speaker, L_5 utters the word once for the purpose of control.

For the Kanuri language, the developed networks; $K_{1-1}$and K_{2-1} are expected to recognize the speakers they were specifically trained for (K_1 and K_3 respectively).For each network, three inputs (K_1, K_3, and K_5) were presented. Results obtained (as in table 3) shows that network K_{1-1} has an overall R-value of 0.9793 for K_1, 0.5371 for K_3 and 0.4799 for K_5. This indicates that K_1 was verified because it has a high R-value. Similarly, network K_{2-1} has an overall R-value of 0.4674 for K_1, 0.8804 for $K_3$and 0.5178 for K_5, indicating that K_3 was verified.

For the Yoruba language, the networks developed; $Y_{1-1}$and $Y_{2-1}$are expected to recognize the speakers Y_2 and Y_4 respectively for which they were specifically trained for. For each network, three inputs (Y_1, Y_3, and Y_5) were presented. From table 4, results obtained shows that network Y_{1-1} has an overall R-value of 0.9534 for Y_1, 0.5064 for Y_3 and 0.3901 for Y_5. This indicates that Y_1 was verified because it has a high R-value. Also, network Y_{2-1} has an overall R-value of 0.4730 for Y_1, 0.9999 for Y_3 and 0.3579 for Y_5,indicating that Y_3 was verified.

Similarly the networks developed; $F_{1-1}$and F_{2-1}(for the Fulani word) are expected to recognize the speakers F_2 and F_4 respectively for which they were specifically trained for.For each network, three inputs (F_1, F_3, and F_5) were presented. Results obtained as given in table 5 shows that network F_{1-1} has an overall R-value of 0.9780 for F_1, 0.4666 for F_3 and 0.3185 for F_5. This indicates that F_1 was verified because it has a high R-value. Also, network F_{2-1} has an overall R-value of 0.3330 for F_1, 0.9555 for F_3 and 0.5334 for F_5,indicating that F_3 was verified.

From results outlined, the R-values that falls between 0.25 and 0.60 indicates that there is no any significant relationship between the targets and the inputs signals presented to the networks, which implies that the input signals are not verified. While, where overall R-values are above 0.85, it indicates that there isrelationship between the targets and the inputs signals presented to the networks, which implies that the input signals are verified.

IV. Conclusion

The results obtained from the system performance gives credence to the possibility of using the system for voice recognition (verification). The network performance in this work shows acceptance for voice signals from the same speaker with reasonable level of accuracy, and a total rejection for voice signal from different speakers. These deductions can be clearly seen in the regression coefficients of the various training network given in tables 3 – 5. In the tables, the approximate values of the regression coefficients for the trained voice signals from the same speakers are 0.85 (acceptance). And the approximate values of the regression coefficients for trained voice signals from different speakers are 0.4 (rejection).

Thus, considering the overall system performance (both network performance and regression coefficients), the system performance can be rated to an approximate recognition rate of 100 %.

References

[1]. Ali, H., Mijanur, R., Uzzal, K. P., & Farukuzzaman, K. (2013). Implementation Of Back-Propagation Neural Network For Isolated Bangla Speech Recognition. International Journal of Information Sciences and Techniques (IJIST) , 3 (4), 1-9.

[2]. Andrew, K. (2000). Basic of Matlab and beyond. London: Chapman and Hall/CRC Press.

[3]. Desmond, S., & Graham, F. W. (2007). Age-related changes in long-term average spectra of children voices. Journal of Voice. , 5 (3), 127-135.

[4]. Furui, S. (1986). Research on individuality features in speech waves and automatic speaker recognition techniques. 5 (2), 183 - 197.

[5]. Hagan, M. T., Demuth, H. B., && Beale, M. H. (2014). MATLAB Neural Network ToolboxUser's Guide. united kingdom: The Mathworks Inc.

[6]. Ibikunle, F., & Katende, J. (2013). Recognition of Nigerian Major Languages Using Neural Networks. Journal of Computer Networks. , 1 (2), 32-37.

[7]. John, G. P., & Dimitris, G. M. (1996). Digital Signal Processing: Principles, Algorithms, and Applications (3rd ed.). Upper Saddle River, New Jersey 07458: Prentice-Hall, Inc. Simon & Schuster Viacom Company.

[8]. Kassim, S. O., & Anene, E. C. (2015). Text-Dependent Speaker Verification System Using Neural Network. International Journal of Emerging Technology and Advanced Engineering , 5 (5), 43-49.

[9]. Kaustubh, B. J., & Vijay, V. P. (2015). Text-dependent Speaker Recognition and Verification using Mel Frequency Cepstral Coefficient and Dynamic Time Warping. International Journal of Electronics & Communication Technology , 6 (3), 150-154.

[10]. Patricia, M., Jerica, U., Daniel, S., Miguel, S., Miguel, L., & Oscar, C. (2006). voice recognition with neural network, type-2 fuzzy logic and generic algorithms. 13(2). mexico: mexican researcg council (CONACYT).