Neural Network Based Hausa Language Speech Recognition

Matthew K. Luka
Department of Electrical and Information Engineering, Covenant University, Ota, Nigeria

Ibikunle A. Frank
Department of Electrical and Information Engineering, Covenant University, Ota, Nigeria

Gregory Onwodi
National Open University of Nigeria
Lagos, Nigeria

***Abstract*--Speech recognition is a key element of diverse applications in communication systems, medical transcription systems, security systems etc. However, there has been very little research in the domain of speech processing for African languages, thus, the need to extend the frontier of research in order to port in, the diverse applications based on speech recognition. Hausa language is an important indigenous lingua franca in west and central Africa, spoken as a first or second language by about fifty million people. Speech recognition of Hausa Language is presented in this paper. A pattern recognition neural network was used for developing the system.**

Keywords--Artificial neural network; Hausa language; pattern recognition; Speech recognition; speech processing

I. INTRODUCTION

Speech is a natural mode of communication for people because it is the most efficient modality for communication. This is coupled with the fact that the human brain has an impressive superiority at speech recognition as with other cognitive skills. Automatic speech recognition (ASR) is a process by which a machine identifies speech [1]. Speech recognition is used in different domain of life such as in telecommunications, mobile telephony, multimedia interaction, transcription, video games, home automation [2, 3].

The accuracy of a speech recognition system is affected by a number of factors. Generally, the error associated with discriminating words increases as the vocabulary size grows. Even a small vocabulary can be difficult to recognize if it contains confusable words. Also, speaker independence is difficult to achieve because system's parameters become tuned to the speaker(s) that it was trained on, and these parameters tend to be highly speaker-specific [4]. Other factors of interest are: continuity of speech, task and language constraint, and adverse conditions.

Generally, there are three methods widely used in speech recognition: Dynamic Time Warping (DWT), Hidden Markov Model (HMM) and ANNs [5]. Dynamic Time Warping algorithm is used to recognize an isolated word sample by comparing it against a number of stored word templates to determine the one that best matches it. This goal is complicated by a number of factors.

First, different samples of a given word will have somewhat different durations. This problem can be eliminated by simply normalizing the templates and the unknown speech so that they all have an equal duration. Dynamic Time Warping (DTW) is an efficient method for finding optimal nonlinear alignment between a template and the speech sample. The main problem of systems based on DTW is the little amount of learning words, high computation rate and large memory requirements [6]. A Hidden Markov Model is a statistical Markov Model in which the system being modelled is assumed to be a Markov process with unidentified (hidden) states. For speech recognition, the use of HMM is subject to the following constraint:

1) must be based on a first order Markov chain; 2) must have stationary states transitions; 3) observations-independence and 4) probability constraints.

Because speech recognition is basically a pattern recognition problem [4], neural networks, which are good at pattern recognition, can be used for speech recognition. Many early researchers naturally tried applying neural networks to speech recognition. The earliest attempts involved highly simplified tasks, e.g., classifying speech segments as voiced/unvoiced, or nasal/fricative/plosive. Success in these experiments encouraged researchers to move on to phoneme classification; this task became a proving ground for neural networks as they quickly achieved world-class results.

Speech recognition has been applied to most western and Asian languages. However, there is limited work reported on African languages. This research paper provides the background to the application of ANNs for speech processing of Hausa language. A pattern recognition network; which is a feed-forward network that can be trained to classify inputs according to target classes was used for recognition of Hausa language isolated words.

ANN is an information processing system that modelled on the performance characteristics of biological neural networks. ANNs are developed as generalizations of mathematical models of neural biology or human cognition based on the following assumptions [7]:

- Processing of Information occurs at many simple nodes (nodes are also called cells, units or neurons).
- Connection links are used for passing signals between the nodes.
- Each connection link is associated with a weight, which is a number that multiplies the signals.

- Each cell applies an activation function (usually nonlinear) to the weighted sum of its input to produce an output.

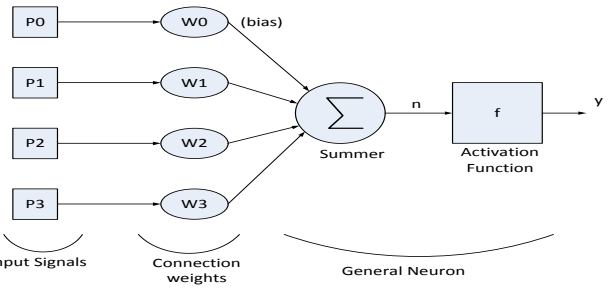

Figure 1. Mathematical Model of a basic neural network

The mathematical model of a three-input neuron is given in Fig. 1. The inputs are modified by the weights (synapses in biological neural network). A positive weight represents an excitatory connection, which a negative weight designates an inhibitory connection. A fictitious input known as the bias is normally used for training. The weighted inputs are summed in a linear fashion. Lastly, an activation function is applied to the weighted sum to determine the range and characteristic of the output. Mathematically,

$$n = \left(\sum_{i=1}^{3} w_j p_i + w_o p_o \right) \quad (1)$$

$$y = f(n) \quad (2)$$

There are several activation functions available. The four most commonly used ones include: (1) the threshold function, (2) the piecewise-linear function, (3) the sigmoid function and (4) the Gaussian (bell shaped) activation function. Based on architecture (connection patterns), artificial neural networks can be grouped into two groups; feed-forward networks and recurrent (feedback) networks. Feed-forward networks involve the one way flow of data from the input to the output. Examples of feed-forward networks include: single layer perceptron, multilayer perceptron and radial basis function networks. On the other hand, recurrent networks contain feedback connections which make their dynamical network properties important. Classical examples of feedback networks include: Competitive networks, Kohonen's Self-Organizing Maps, Hopfield networks and Adaptive Resonance Theory Models. Learning (or training) in the artificial neural network context is the task of updating network architecture and connection weights so that a network can efficiently perform a function [8]. The major difference between learning in ANN and experts systems is that in the latter, learning is based on a set of rules specified by human experts whereas in ANN, learning process is automatic (e.g. from input-output relationships). There are three main learning paradigms: Supervised, Unsupervised and Hybrid. Also, there are four fundamental learning rules (how connections are updated): Hebbian, error-correction, Boltzmann and competitive learning. The manner in which this rules are used for training a specific architecture is referred to as the learning algorithm.

II. Overview of Speech Recognition

Speech recognition deals with the analysis of the linguistic content of a speech signal. Speech recognition is one of the areas of research in speech processing and it is the study of speech signals and processing methods. Other aspects of speech processing include: speaker recognition, speech enhancement, speech coding, and voice analysis and speech synthesis. Based on the type of utterance to be recognized, speech recognition systems can be classified into four categories [9]: Isolated words system, connected words, Continuous speech and spontaneous speech systems. Human speech recognition starts with receiving speech waveform through the ear. It is based on "Partial recognition" of information across frequency, probably in the form of speech features that are local in frequency (e.g formants) [10]. The partial recognition (extracted features) is then integrated into sound units (phonemes), and the phonemes are then grouped into syllables, then words and so forth.

In order to analyse linguistic content of speech, the acoustic signal is first converted into an analogue signal that can be processed by an electronic device. The analogue signal is sampled and digitized for storage and further processing. In this work, PRAAT- a flexible program was used for the analysis and reconstruction of acoustic speech signals. PRAAT offers a wide range of procedures, including spectrographic analysis, pitch analysis, intensity analysis, articular synthesis and neural networks.

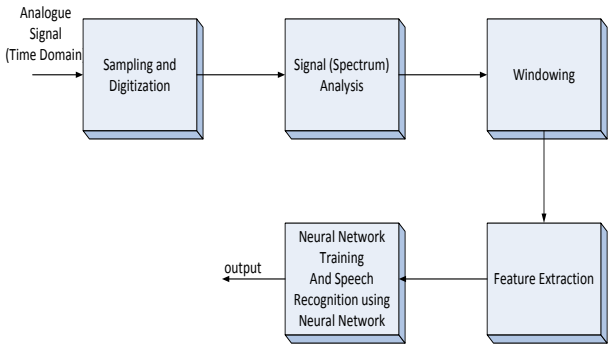

Figure 2. Connectionist based speech recognition model

Speech recognition may be considered to involve five important processes which are:

- Input Acquisition
- pre-processing (pre-emphasis and windowing)
- Feature Extraction
- Modelling
- Model Validation (testing)

A. Input Acquisition

The acoustic signals were captured in a low noise environment using a good quality microphone. The analogue output signal of the microphone is recorded using PRAAT at a sampling rate of 12 KHz. To facilitate ease of editing and manipulation, the digital data is saved as a .wav file.

B. Pre-Processing

Pre-Processing is the separation of the voiced region from the silence/unvoiced portion of the captured signal [11]. Pre-processing is necessary because most of the speaker or speech specific information are present in the voiced part of the speech signal [12]. Pre-processing also helps significantly in reducing the computational complexity of later stage [13]. Two vital pre-processing steps are pre-emphasis and windowing.

Pre-emphasis: In general, the digitized speech waveform has a high dynamic range and suffers from additive noise to reduce this range pre-emphasis is applied. By pre-emphasis, we imply the application of a high pass filter, which is usually a first-order FIR of the form:

$$H(z) = \sum_{k=0}^{N} \alpha(k) z^{-k} \qquad (3)$$

Normally, a single coefficient filter digital filter known as pre-emphasis filter is used:

$$H(z) = 1 - az^{-1} \qquad (4)$$

Where the pre-emphasis factor α is computed as:

$$\alpha = \exp(-2\pi F \Delta t) \qquad (5)$$

Where F is the spectral slope will increase by 6dB/octave and Δt is the sampling period of the sound. The pre-emphasis factor is chosen as a trade-off between vowel and consonants discrimination capability [14].The new sound y is then computed as:

$$y_i = x_i - \alpha x_{i-1} \qquad (6)$$

Hearing is more sensitive above the 1 kHz. The pre-emphasis filter amplifies this region of the spectrum. This assists the spectral analysis algorithm in modelling the perceptually important aspects of the speech spectrum [15]. In this work, a pre-emphasis factor of 0.9742 was used.

Windowing: To minimize the discontinuity of signal at the beginning and end of each frame, the signal should be tapered to zero or near zero, and hence reduce the mismatch. This can be achieved by windowing each frame to increase the correlation of the Mel Frequency Cepstrum Coefficients (MFCC), Spectral estimates between consecutive frames [16]. To the chosen 12 Mel-Frequency coefficients, and for time 0.005 seconds, a window length of 0.015 is selected using the PRAAT Object software tool.

A. Feature Extraction

Feature extraction involves computing representations of the speech signal that are robust to acoustic variation but sensitive to linguistic content [17]. There are different ways of representing speech. Some feature extraction techniques include: Linear Discriminate Analysis, Linear Predictive Coding, Cepstral Analysis, Mel-frequency scale analysis, principal component analysis and Mel-frequency Cepstrum.

The feature extracted depends on the situation and the kind of speech information to be represented. A waveform, which is the variation of speech amplitude in time, is the most general way to represent a signal. However, a waveform contains too much irrelevant data to use it directly for pattern recognition. The spectrogram offers a better three dimensional representation of speech signals. The patterns represented by spectrogram tend to vary significantly, which makes it unsuitable for pattern recognition.

The Mel Filter Bank is a set of triangular filter banks used to approximate the frequency resolution of the human ear. The filter function depends on three parameters: the lower frequency, the central frequency and higher frequency. On a Mel scale the distances between the lower and the central frequencies, and that of the higher and the central frequencies are equal. The filter functions are:

$$H(f) = 0 \; \forall \; f \leq f_l \; and \; f \geq f_h \qquad (7)$$

$$H(f) = \frac{f - f_l}{f_c - f_l} \quad \forall \; f_l \leq f \leq f_c \qquad (8)$$

$$\mathrm{H(f)} = \frac{\mathrm{f_h - f}}{\mathrm{f_h - f_c}} \quad \forall \; \mathrm{f_c} \leq \mathrm{f} \leq \mathrm{f_h} \qquad (9)$$

The Mel frequency cepstral coefficients are found from the Discrete Cosine Transform of the Filter bank spectrum by using the formula given by Davis and Mermelstein [18].

$$C_i = \sum_{j=1}^{N} P_j \cos(i\pi / N(j - 0.5)) \qquad (10)$$

Where P_j denotes the power in dB in the jth filter and Ndenotes number of samples. 12 Mel frequency coefficients are considered for windowing. Mel-Frequency analysis of speech is based on human perception experiments.

The signal is sampled at 12 kHz. The sampled speech data is applied to the Mel filter and the filtered signal is trained. The number of frames for each utterance is obtained from the frequency coefficients by using PRAAT object software tool.

III. Implementation of the System

The number of frame for each word is used as inputs for the neural network.The vocabulary set is composed of ten Hausa words. Each word is spoken eight times by four speakers (two male and two female).

Thus the database is composed of 320 words. Frames obtained for each utterance of the speaker form Mel-Frequency Cepstral Coefficients are as shown in Table 1.

The database was divided randomly into training, testing and validation sets. A respective division ratio of 70%, 15% and 15% was adopted in splitting the data. The training subset is used for computing the gradient and updating the network weights and biases. The validation subset is used to prevent model overfitting.

The error on the validation set is monitored during the training process. Normally, the validation error decrease during the starting phase of the training as does the training set error.

However, the validation set will typically begin to rise when the network begins to over fit. If the validation error continues to increase for a given number of epochs, the training will be stopped. The testing subset is not used during training, but it is used to compare different models.

A Multilayered Neural Network is used for speech recognition task. The network is made up of eight (8) inputs corresponding to an utterance of each word for every speaker. A hidden layer consisting of ten neurons and an output layer of ten neurons was used as shon in Figure 3. The scaled conjugate gradient back-propagation algorithm was used for training because it converges faster which helps improve generalization.

TABLE I. FRAMES OBTAINED FOR MFCC FEATURE

	Words (utterance)									
	Daya	***Biyu***	***Uku***	***Hudu***	***Biyar***	***Shida***	***Bakwai***	***takwas***	***tara***	***Goma***
Speaker1										
	353	438	379	344	353	353	430	353	404	327
	319	438	464	336	344	361	319	361	310	361
	285	430	344	387	336	344	336	421	361	370
	276	575	404	310	361	319	353	353	344	319
	285	430	438	370	413	319	336	327	344	310
	268	472	387	336	336	336	353	361	370	310
	225	430	370	327	327	344	336	413	404	344
	302	558	396	319	310	387	344	566	302	310
Speaker2										
	413	447	344	379	327	361	387	310	353	336
	370	404	268	353	319	293	276	327	336	293
	455	489	507	327	319	302	379	293	302	319
	361	336	421	413	319	276	276	336	455	302
	361	319	361	293	370	293	379	319	276	285
	498	438	626	310	302	310	396	396	259	310
	353	336	421	566	507	344	327	285	310	310
	319	336	327	327	455	293	327	327	404	319
Speaker3										
	430	370	379	319	379	387	524	379	396	396
	438	361	327	361	379	353	319	319	353	361
	498	353	336	319	387	396	344	549	370	361
	447	447	413	396	336	336	396	438	481	353
	310	379	361	327	310	438	310	464	370	327
	319	387	361	379	353	344	319	379	438	455
	327	319	361	481	387	327	327	455	353	327
	455	600	361	379	353	293	336	464	353	387
Speaker4										
	293	293	379	370	302	293	353	344	319	293
	327	276	285	524	276	233	319	404	507	327
	276	293	319	310	293	302	302	285	293	276
	310	242	319	404	285	464	302	276	336	268
	293	268	293	344	293	225	293	310	507	268
	276	242	285	370	251	251	327	285	276	293
	285	276	285	259	336	293	242	370	268	302
	300	635	276	498	276	327	293	319	302	319

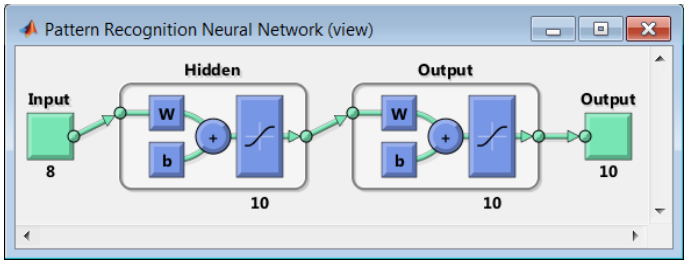

Figure 3. Multi-layered Pattern Recognition Neural Network

IV. SIMULATION RESULTS AND DISCUSSION

The experimental results for the four speakers are presented in Figures 4-7. The performance of the system depends on a number of factors. Adequate pre-processing is a very vital element to ensuring good results in a speech recognition system.

The choice of adequate training parameters helps in arriving at good generalization. From the results, the speech model performance for speaker 2 and speaker 4 was better than that of speakers 1 and 3. The results obtained for the overall data of all the speakers combined is given in Figure 8. When the data was combined, a better generalization was obtained than that of individual speakers.

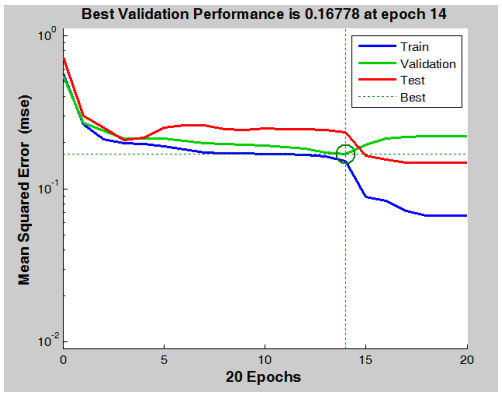

Figure 4. Simulation results for speaker 1 speech recognition

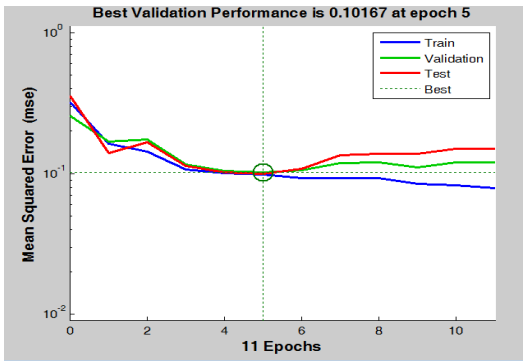

Figure 5. Simulation results for speaker 2 speech recognition

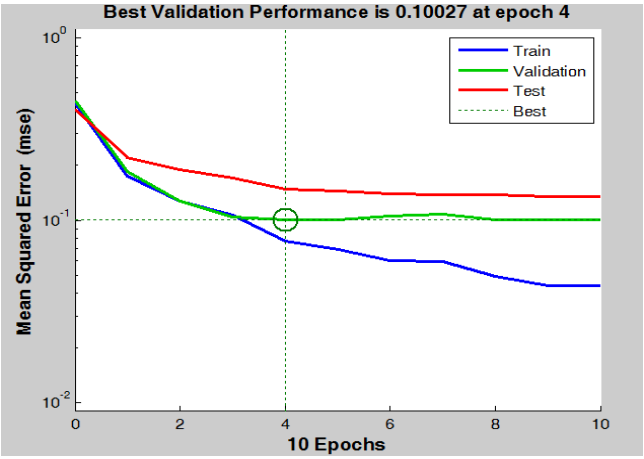

Figure 6. Simulation results for speaker 3 speech recognition

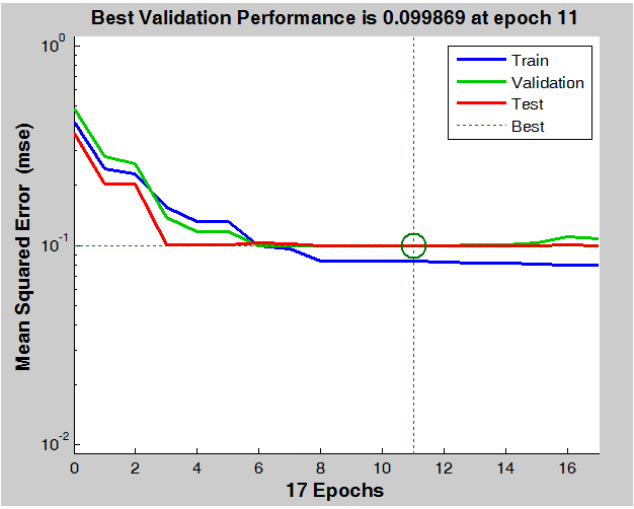

Figure 7. Simulation results for speaker 4 speech recognition

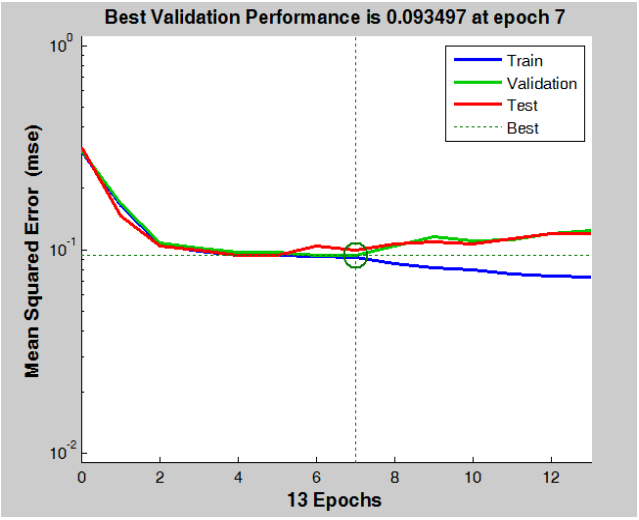

Figure 8. Hausa speech recognition for all the speakers.

V. Conclusion

This paper presents speech recognition of Hausa language using pattern recognition artificial neural network. The captured data was first pre-processed then the Mel-Frequency Cepstral Coefficients (MFCC) was extracted. These extracted features were used to train the neural network. The choice of pre-emphasis filter factor was carefully chosen because Hausa is a tone language, in which syllable-based pitch differences add as much to the meaning of words as do consonants and vowels. A better generalization was achieved by adopting early stopping achieved using model validation. The generalization power of the neural network increase when the size of the database increases. This work is a primer for more extensive research in speech processing of African languages (Hausa language in particular).

Future work will focus on increasing the database and expanding the loci of research to other domains of speech processing such as speech enhancement, speech coding, and voice analysis and speech synthesis.

References

[1] R.L.K.Venkateswarlu, R. Vasantha Kumari, G.VaniJayaSri, "Speech Recognition by Using Recurrent NeuralNetworks", International Journal of Scientific & Engineering Research Volume 2, Issue 6, June-2011 1ISSN 2229-5518

[2] Lawrence R. Rabiner, "Applications of Speech Recognition in the Area of Telecommunications", retrieved on March 20th, 2012, from http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/341_telecom%20applications.pdf

[3] Wikipedia Encyclopaedia, "Speech Recognition", Retrieved March 20th, 2012, from http://en.wikipedia.org/wiki/speech

[4] DanielRios, "Speech Recognition ", *NeuroAI: Algorithms and applications*, Retrieved March 20th , 2012, from www.learnartificialneuralnetworks.com/speechrecognition.htm

[5] Song Yang, MengJooEr and Yang Gao. " A High Performance Neural-Network-Based Speech Recognition System". IEEE trans. PP 1527, 2001

[6] Meysam Mohamad Pour and Fardad Farokhi." An Advanced Method for Speech Recognition". World Academy of Science and Technology, pp. 995-1000, 2009

[7] Laurene Fausett, "Fundamentals of Neural Networks: Architectures, Algorithms and Applications", 1st edit., (pg. 3) Prentice hall, 1993

[8] Anil K. Jain and Jianchang Mao, "Artificial Neural Networks: A Tutorial" *IEEE Computer* (pg.31-44), 1996.

[9] SantoshK.Gaikwad, BhartiW.Gawali and PravinYannawar, "A Review on Speech Recognition Technique", International Journal of Computer Applications (0975 – 8887), Vol., 10, No.3, Nov. 2010

[10] Jont B. Allen. "How do Humans Process and Recognize Speech". *IEEE trans.*, pp. 567-577, Vol.2 No.4, Oct., 1994

[11] AyazKeerio et al, "On Pre-processing of Speech Signals", *World Academy of Science, Engineering and Technology* 47 (pg. 317-323) 2008.

[12] Saha. G., Chakroborty. S., and Senapati. S, "A new SilenceRemoval and End Point Detection Algorithm for Speech andSpeaker Recognition Applications", in Proc. of Eleventh National Conference on Communications (NCC), IITKharagpur,India, January 28-30, 2005, pp. 291-295.

[13] Mitra. A., Chatterjee. B., Mitra. B. K., "Identification ofPrimitive Speech Signals using TMS320C54X DSP Processor",in Proc. of Eleventh National Conference on Communications (NCC), IIT-Kharagpur, India, January 28-30, 2005, pp. 286-290.

[14] Kuldip K. Paliwal, "Effect of Pre-emphasis on Vowel Recognition Performance", *speech communication 3* (pg.101-106), North-Holland, 1984.

[15] J. W. Picone. "Signal modelling technique in speech recognition". IEEE Proc., vol. 81, no.9, pp. 1215-1247, Sep. 1993.

[16] Picton, P, "Neural Networks," Palgrave, NY 2000.

[17] R.L.K. Venkateswarlu and R. Vasanthakumari. "Neuro Based Approach For Speech Recognition By Using Mel-Frequency Cepstral Coefficients" IJSC, Vol. 2, No. 1, January-June 2011, pp. 53– 57

[18] Davis S. B and Mermelstein P., "Comparison of Parametric Representations for Mono-symbolic word Recognition in continuously spoken sentences", *IEEE trans. on Acoustic Speech Signal Process.*, Vol. ASSP-28, No. 4, 1980, pg. 357-366

AUTHORS PROFILE

Matthew K. Luka is a postgraduate student of Covenant University. His research Interest include: Artificial Intelligence, Wireless Communications.

Frank A. Ibikunle (PhD) is a senior lecturer with Covenant University, Ota. He is a seasoned lecturer with extensive industry experience in wireless communications and internet systems. He also has novel publications on various application of artificial intelligence.

Gregory Onwodi is with the National Open University of Nigeria