Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

Creation of a Nigerian Voice Corpus for Indigenous Speaker Recognition

To cite this article: Adekunle A. Akinrinmade *et al* 2019 *J. Phys.: Conf. Ser.* **1378** 032011

Creation of a Nigerian Voice Corpus for Indigenous Speaker Recognition

Adekunle A. Akinrinmade[1], Emmanuel Adetiba[1,2], Joke A. Badejo[1], Aderemi A. Atayero[1]
[1]Department of Electrical and Information Engineering, Covenant University, P.M.B 1023, Ota, Nigeria.
[2]HRA, Institute for Systems Science, Durban University of Technology, P.O. Box 1334, Durban, South Africa.
Corresponding author: adekunleakinrinmade@gmail.com, +2348032004087

Abstract

One of the goals of Word Bank's Identification for Development (ID4D) is the realization of robust digital identification systems as a means of sustainable development priority. ID4D's most recent report shows about 1.1 billion of the world's population are yet to be identified for development. Africa represents about half of that number while Nigeria represents about a quarter of Africa's share. Biometrics is the state-of-the-art approach for identification using human behavioral and/or physiological digitally calibrated traits and one such trait is the voice. The backbone of biometric research is the database employed in the design of biometric systems. Although many voice databases are publicly available such as the THCHS-30 for Chinese and Microsoft Indian language Speech Corpus for Indians, none is currently publicly available or free for Nigerians. The creation of such an indigenous database (or corpus) can open doors to Nigerian automatic speaker recognition as well as for indigenous language, ethnicity, gender, age group and emotion classification amongst others. This work is a first step in the direction of creating a Nigerian Voice Corpus (NVC) to aid indigenous voice biometric research. A voice corpus of popular Nigerians was created by curation of audio samples of 14 women and 23 men from YouTube. The corpus contains 10 different samples of 5 seconds duration for each individual resulting in a total of 370 samples. The created corpus was used to carry out speaker recognition experiment by dividing the audio samples into 25ms non-overlapping frame durations. Silent frames were excluded using short-term spectral energy threshold for Voice Activity Detection (VAD). This was followed by extraction of Mel Frequency Cepstral Coefficient (MFCC) as descriptors to discriminate different speakers using Support Vector Machine (SVM) with median Gaussian function. An overall recognition accuracy of 93.24% was achieved demonstrating the feasibility and research potential in this direction.

Keywords: ID4D, MFCC, Nigerian Voice Corpus (NVC), Speaker Recognition, SVM, VAD

1. Introduction

Voice recognition is identification based on voice characteristics of speakers regardless of spoken words or language [1]. Voice offers the biometric advantage of requiring no physical contact and allows control of other devices while the arms are engaged with other activities [2]. Biometric databases are important to research because they serve as tools to evaluate the performance of various recognition algorithms and to compare results amongst them. Without biometric databases, researchers have no efficient means of testing developed algorithms. For example, Mahmoodi et al. [3] made use of FARSDAT database containing Farsi speeches from 304 Iranian speakers to perform age estimation by speech with experiments using MFCC and Perceptual Linear Prediction (PLP) features and SVM classifier they obtained an accuracy of 94.11% for MFCC features and 91.16% for PLP features. Liu *et al.* [4] used the Mandarin corpus MTDSR2015 containing audio recordings from 181 subjects in their voiceprint-based identity research. They obtained an Equal Error Rate (EER) of 1.17% using MFCC, i-vectors and Cosine distance metrics. The Korean word corpus consisting of 30 speakers (7 females and 23 males) was used in [2] to train and test a keyword-based speaker specific voice trigger system for mobile devices using MFCC features with Hidden Markov Model (HMM) for classification achieving an accuracy of 90.50%.

In the work of Bahari et al carried out to estimate age based on voice characteristics, the telephone speech segment of the NIST (National Institute of Standards and Technology) 2010 and 2008 SRE (Speaker Recognition Evaluation) databases were used to evaluate their methodology which made use of MFCC for features, i-vectors for modeling and Support Vector Regression (SVR) for estimation [5]. They achieved a Mean Average Error (MAE) of 7.63 and 7.61 for male and female respectively in their estimations. Fedorova et al applied exactly the same methodology using the same databases but instead of SVR they explored Artificial Neural network (ANN) for age estimation based of voice features obtaining MAE of 6.35 and 5.49 for the male and female counterparts respectively showing that ANN gave a better performance over SVR [6]. McCool et al applied the same methodology to person authentication on a Nokia N900 mobile device using voice and face traits substituting Probabilistic Linear Discriminant Analysis (PLDA) in place of SVR. They achieved an Equal Error Rate (EER) performance of 10.9% and 10.5% for male and female respectively on the Mobile Biometry (MOBIO) database [7].

Some researchers on the other hand have had to create their own database for evaluation of algorithms for voice biometrics to suit their needs, for example, Dhinesh et al [8], in order to develop both speaker and word recognition of low complexity for resource-limited mobile devices created a closed-loop database consisting of 5 words from 5 speakers. Their algorithm which employed MFCC for features, Gaussian Mixture Model (GMM) for modeling and classification was able to yield a recognition accuracy of 91.6% for both the pronounced words and the speakers. Their algorithm was suitable for applications in real-time because it performed identification in fractions of a second. Andrei et al [9], in order to achieve a design that consumes low processing power and memory space for text-dependent speaker recognition in real-time developed a database of 23 speakers consisting of 12 females and 11 males with age ranges within 20 and 60 years. Using MFCC features and Dynamic Time Warping (DTW) approach for classification, they achieved a recognition accuracy of 96.42%. Moreover, since the identification for development problem majorly affect Africans, developing robust biometric systems to meet this goal are better using indigenous traits, hence the need for biometric databases unique to African nations, that is, Nigeria in the study at hand.

2. Materials and Methods

2.1. Curating the Nigerian Voice Corpus

This voice corpus contains audio information of 37 speakers curated from YouTube who are majorly religious leaders, government officials, entertainers and so on. Ten audio samples each of 5 seconds duration were curated for each speaker in indoor and outdoor environments during interviews, campaigns or public speaking. The recordings in the database are as such expected to be plagued with some form of background and channel noises owing to recordings at different environments and with different microphones.

2.2. Preprocessing

The basic preprocessing activity in this work was windowing or framing of the voice signals. The frame sizes used were 25ms durations with no overlap, each frame was first processed to detect and remove silence using a number of Voice Activity Detection (VAD) techniques such as short term energy threshold in the time-domain, Zero Crossing Rate (ZCR) and short term spectral energy which was eventually adopted having yielded best result. This parameter was then normalized and frames having threshold values less than 0.01 were excluded from further processing.

2.3. Feature Extraction

Voice features refer to the properties that are unique to individual speakers that could be used for the purpose of recognition. Such features could be as simple as the pitch, short-term time-domain energy, energy entropy, spectral entropy and spectral centroid. They could also be more sophisticated like the Linear Prediction Coefficient (LPC), Linear Prediction Cepstral Coefficient (LPCC) [10], Linear Frequency Cepstral Coefficient (LFCC), MFCC [11] [12] and others. The feature used in this work was the MFCC. The MFCC computation generally involves Discrete Fourier Transformation (DFT) of the preprocessed audio signal, which is passed through a filter-bank followed by logarithmic compression and a final Discrete Cosine Transform (DCT) to obtain the MFCC vectors. MFCCs are features which models the vocal tract using a number of coefficients based on the understanding of how speech is perceived by the human ear. In this work, frame durations of 25ms with no overlap were used for MFCC computation using 10 equally spaced triangular filters converted to the Mel frequency scale afterwards. Only the first 13 coefficients were retained as feature vectors with the first of these coefficients replaced by the log energy values of the corresponding frames. Figures 1 and 2 illustrate the procedure for MFCC computation.

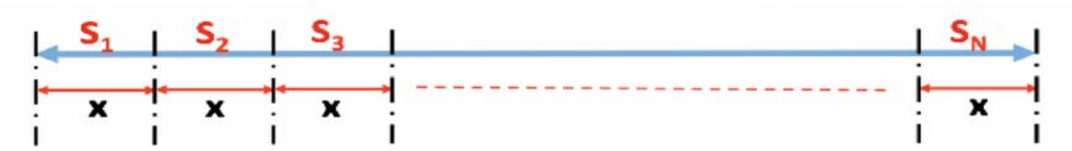

Figure 1: Division of audio sample from video into segments (S1,S2, …,SN)

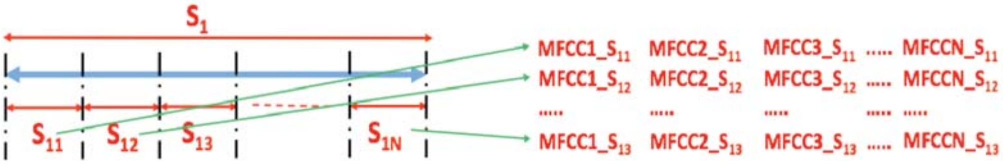
S_1
S_{11}
S_{12}
S_{13}
S_{1N}
MFCC1_S_{11} MFCC2_S_{11} MFCC3_S_{11} ….. MFCCN_S_{11}
MFCC1_S_{12} MFCC2_S_{12} MFCC3_S_{12} ….. MFCCN_S_{12}
….. ….. ….. ….. …..
MFCC1_S_{13} MFCC2_S_{13} MFCC3_S_{13} ….. MFCCN_S_{13}

Figure 2: Division of audio into frames to obtain frame features (MFCC)

2.4. Classification

Classification is the process of placing a given data point into the right category. Machine learning algorithms that exist for classification of features are broadly divided into deep and shallow techniques. Since there are only 37 unique speakers in this work, shallow classification algorithms were utilized. The two classifiers used in this work were K Nearest Neighbors (KNN) and SVM [13] [14]. The SVM algorithm determines the class a given data point falls into amongst different classes by extending the margins between them as wide apart as possible while KNN performs classification using the K nearest training examples in the provided feature space considering the one with majority of points to determine which class a given data point falls into. The database was split into two categories, the first category was the training set which made use of the first 8 of the 10 samples of the speakers in training the SVM and KNN.

3. Result and Discussions

Figure 3 shows the results obtained from using different VAD algorithms to preprocess the first voice sample of subject 1 while Figure 4 shows the scatter plot of the features extracted for the 37 speakers.

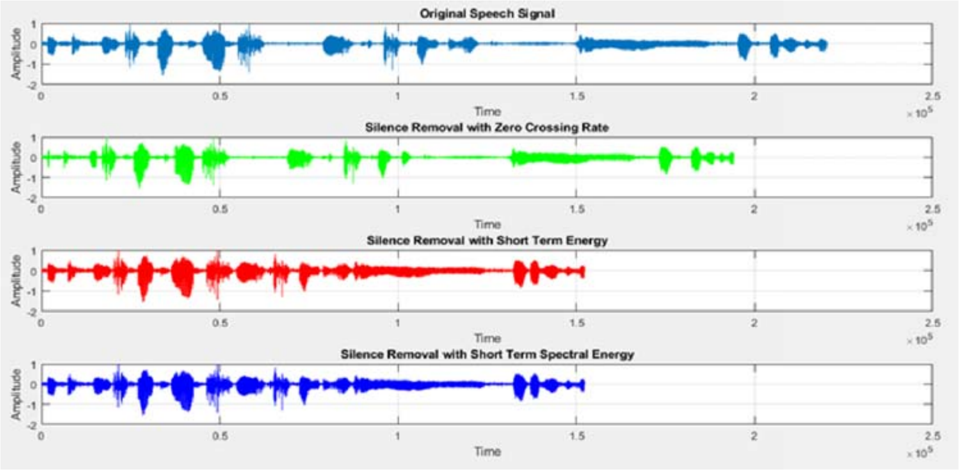

Figure 3: Result of preprocessing with different VAD algorithms for voice sample 1 of Subject 1

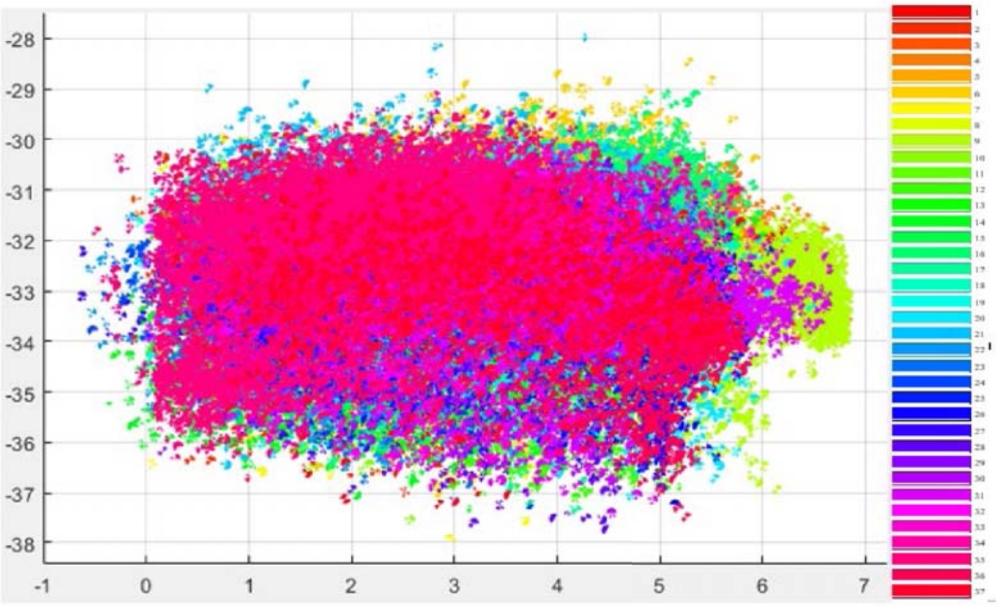

Figure 4. Scatter plot of extracted features from 37 subjects

Notably, SVM gave better testing accuracy than KNN. Figure 5 and Figure 6 show bar charts of the prediction results with the SVM classifier for tests with 9th and 10th audio samples of speakers respectively. In 69 of 74 tests (93.24%), the predictions were correct while the predictions were wrong in 5 cases. For example, the 8th speaker in both tests (having prediction votes of 11 and 22) was falsely predicted as the 35th speaker (having prediction votes of 33 and 55). Associated with the accuracy of performance is another parameter, which measures the confidence level of predictions. For a particular speaker, it is computed as the percentage of the number of frames correctly predicted to the total number of frames processed for that speaker. The bar chart shows the distributions of these proportions amongst the different speakers. The prediction confidence level is an indication of how well the predictions could be trusted, for example, the prediction confidence level for the second sample of the 5th speaker was 103/199 (51.76%) while it was 153/199 (76.88%)for the first sample of the 9th speaker. A number of experiments could be carried out to further improve accuracies like using longer audio durations (since short durations are unlikely to capture enough discriminative information) [12], more robust preprocessing of audio frames to remove effect of noise and exploiting fusion of two or more features or classifier scores.

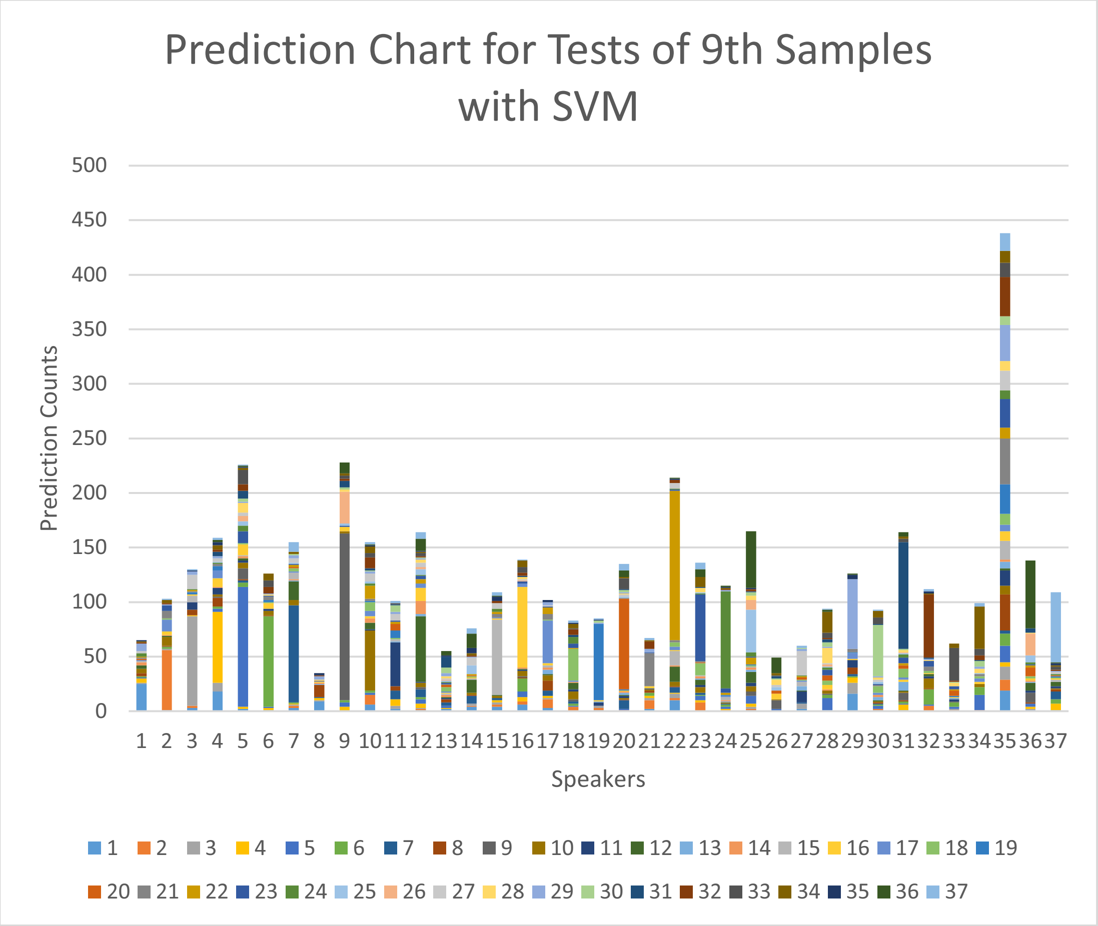

Figure 5. Bar chart of speaker prediction tests of 9th Samples with SVM

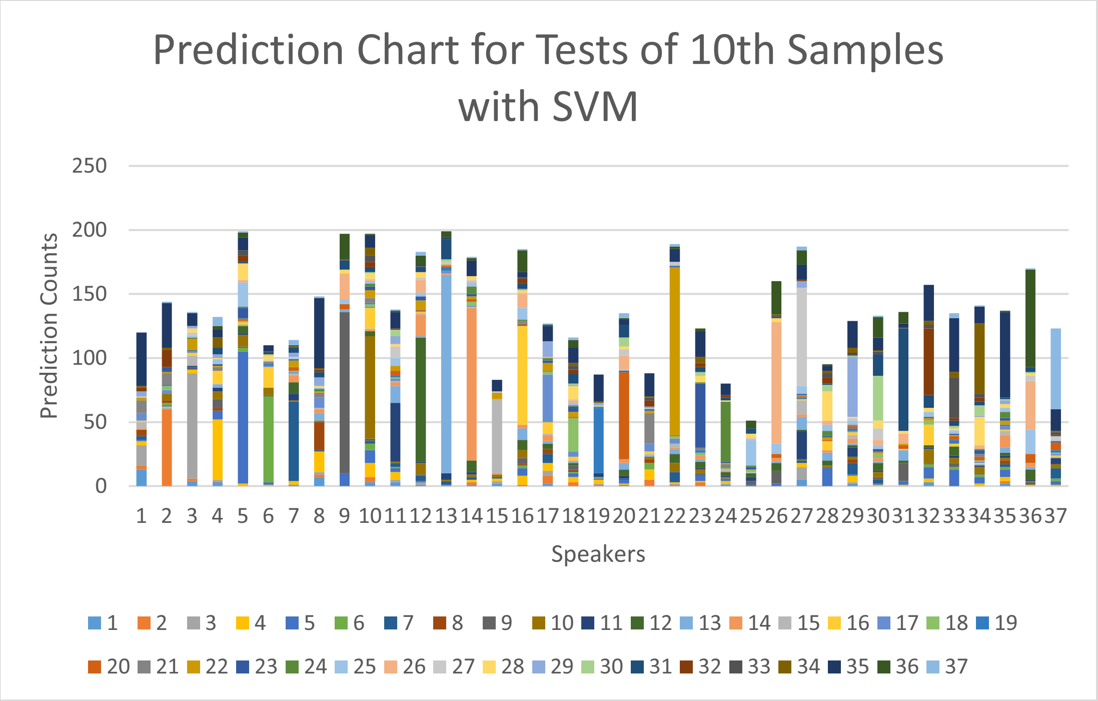

Figure 6. Bar chart of speaker prediction tests of 10th Samples with SVM

4. Conclusion

A Nigeria voice corpus has been created to foster indigenous biometric research. It contains a total of 370 audio samples from 14 female and 23 male Nigerian speakers. Each audio sample is of 5 seconds duration and there are 10 samples per speaker. The voice corpus was created by curation of popular Nigerian voices from YouTube. The voice samples were preprocessed to exclude silent portions then divided into non-overlapping 25 ms frames after which MFCC features were extracted per frame to perform speaker recognition experimenting with KNN and SVM as classifiers. The system identified speakers based on the majority of frames predicted by the classifiers and can in this way also specify prediction level confidence for the speakers. In the future, we hope to improve on the study at hand by extending the corpus to a large scale voice-face multimodal biometric database and apply deep learning for the entire detection pipeline.

Acknowledgement

The publication of this work is fully sponsored by Covenant University Centre for Research, Innovation and Development (CUCRID), Covenant University, Ota, Nigeria.

References

[1] Farjo, J., Aoun, M. H. J., Hamad, M., Kassem, A., & Hamouche, M. (2012). *Speaker identification on compactRIO.* Paper presented at the 2012 16th IEEE Mediterranean Electrotechnical Conference.

[2] Lee, H., Chang, S., Yook, D., & Kim, Y. (2009). A voice trigger system using keyword and speaker recognition for mobile devices. *IEEE Transactions on Consumer Electronics, 55*(4).

[3] Mahmoodi, D., Marvi, H., Taghizadeh, M., Soleimani, A., Razzazi, F., & Mahmoodi, M. (2011). *Age estimation based on speech features and support vector machine.* Paper presented at the Computer Science and Electronic Engineering Conference (CEEC), 2011 3rd.

[4] Liu, J., Zou, Y., & Huang, Y. (2016). *An effective voiceprint based identity authentication system for Mandarin smartphone users.* Paper presented at the 2016 23rd International Conference on Pattern Recognition (ICPR).

[5] Bahari, M. H., McLaren, M., & Van Leeuwen, D. (2012). Age estimation from telephone speech using i-vectors.

[6] Fedorova, A., Glembek, O., Kinnunen, T., & Matějka, P. (2015). *Exploring ANN back-ends for i-vector based speaker age estimation.* Paper presented at the Sixteenth Annual Conference of the International Speech Communication Association.

[7] McCool, C., Marcel, S., Hadid, A., Pietikäinen, M., Matejka, P., Cernocký, J., . . . Levy, C. (2012). *Bi-modal person recognition on a mobile phone: using mobile phone data.* Paper presented at the Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on.

[8] Dhinesh, G. R., Jagadeesh, G. R., & Srikanthan, T. (2011). *A Low-Complexity Speaker-and-Word Recognition Application for Resource-Constrained Devices.* Paper presented at the Electronic System Design (ISED), 2011 International Symposium on.

[9] Andrei, V., Paleologu, C., & Burileanu, C. (2011). *Implementation of a real-time text dependent speaker identification system.* Paper presented at the Speech Technology and Human-Computer Dialogue (SpeD), 2011 6th Conference on.

[10] Frewat, G., Baroud, C., Sammour, R., Kassem, A., & Hamad, M. (2016). *Android voice recognition application with multi speaker feature.* Paper presented at the Electrotechnical Conference (MELECON), 2016 18th Mediterranean.

[11] Prithvi, P., & Kumar, T. K. (2016). Comparative Analysis of MFCC, LFCC, RASTA-PLP. *International Journal of Scientific Engineering and Research, 4*(5), 1-4.

[12] Thakur, S., Adetiba, E., Olugbara, O. O., & Millham, R. (2015). Experimentation using short-term spectral features for secure mobile internet voting authentication. *Mathematical Problems in Engineering, 2015*.

[13] Oyewole, S., Olugbara, O., Adetiba, E., & Nepal, T. (2015). *Classification of product images in different color models with customized kernel for support vector machine.* Paper presented at the 2015 3rd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS).

[14] Adeyemo, J. O., Olugbara, O. O., & Adetiba, E. (2016). *Smart city technology based architecture for refuse disposal management.* Paper presented at the 2016 IST-Africa Week Conference.