Yorùbá Gender Recognition from Speech using Neural Networks

Tshephisho Joseph Sefara
Next Generation Enterprises and Institutions
Council for Scientific and Industrial Research
Pretoria, South Africa
tsefara@csir.co.za

Abiodun Modupe
Next Generation Enterprises and Institutions
Council for Scientific and Industrial Research
Pretoria, South Africa
abiodunmodupe@gmail.com

Abstract—**The impressive improvement in performance obtained using neural networks for automatic speech recognition (ASR) have motivated the application of neural networks to other speech technologies such as speaker, emotion, language, and gender recognition. Prior work has shown significant improvement in gender recognition from images and videos. This paper uses speech to build a gender recognition system based on neural networks. Three types of neural networks are investigated to find the best model for gender recognition system using Yorùbá, namely, feed-forward artificial neural networks (Multilayer Perceptrons), Recurrent neural networks (long short-term memory), and Convolutional neural networks. All the classifier models obtained the state-of-the-art performance in speech-based gender recognition with 99% in accuracy and F$_1$ score.**

Keywords-gender recognition, under-resourced languages, neural network, Yorùbá

I. Introduction

Gender classification is undoubtedly a simple problem for human beings, however, it is still an open research problem that attracts the attention of researchers in different domains such as computer vision and machine learning [1], [2]. Gender recognition is an important topic in signal processing and has a variety of applications, especially in mobile healthcare system [3], facial recognition [2], and age classification [4]. Applications of gender recognition system include verifying a customer when making a telephone bank transaction, security measure when retrieving confidential information, surveillance, forensics, and blog authorship [5]. Such systems are available for well-resourced languages such as English but for African languages like Yorùbá which is an official language in Nigeria, there is limited work being done [6].

Yorùbá is a tonal language and well-known example from the Niger-Congo family, words are distinguished based on two or three distinct level tones realized on a syllable. In a three-tone system, these tones are classified as high, medium and low, mostly relying on changes in pitch between successive syllables. Although, Yorùbá is regarded as an under-resourced language [7], few systems for under-resourced African languages have been developed [8]-[16]. According to our knowledge, this is the first development of a gender recognition system for Yorùbá. While the development of speech-based systems for Yorùbá is an open research, it is important to develop a Yorùbá gender recognition system that may later help as a pre-processing step for other systems such as speech, speaker, and emotion recognition systems. This paper also investigates the effects of changing the threshold on model accuracy. We release the code in support for recognition of under-resourced languages[1]. The contributions of this paper are as follows: (i) We introduce the *What-If tool* [2] to make advanced comparison of the models. (ii) We introduce a Python library[3] to extract speech features.

The rest of the paper is organized as follows: Section II gives the literature review on gender recognition. Section III details the features, learning models, and evaluation methods. Section IV discusses the experimental results, and the paper is concluded in Section V.

II. Literature Review

Gender recognition can be solved using text [5] images [1], [2], [17], videos [18], [19], accelerometers [20], wearables [21], and speech to train machine learning models or neural networks for classification. The improvement in performance obtained using neural networks for automatic speech recognition (ASR) [6] has motivated the application of neural networks to other speech technologies such as speaker recognition [11], emotion recognition [10], [22], and gender recognition [22].

Hwang et al. [2] state gender as one of the important factors for recognizing appearance of the human faces. Authors use gender information to build a face recognition system since most studies do not apply gender information in vision-based face recognition tasks. Moreover, Chen et al. [4] use gender information to apply a two-stage classification using support vector machines (SVMs) where in the first stage: gender information is extracted, and in the second stage: age classification is performed. Authors realized this method outperforms the direct classification scheme using the accuracy rate. Zhan et al. [23] create an automated speaker recognition robot to collect speech signals for evaluating the performance of the system for gender and age classification. Using real-life speech data, authors observe the performance of their method consistently outperforms the individual classifiers. Alhussein et al. [3] stated that the

[1] https://github.com/SefaraTJ/yoruba-gender-recognition
[2] https://github.com/SefaraTJ/what-if-tool
[3] https://github.com/tyiannak/pyAudioAnalysis

contribution of the vocal folds is important in a human voice production system. The length of the vocal folds depends on gender. A female speaker has shorter vocal folds than a male speaker. Hence, the male's voice is heavy and has more voice intensity. Alhussein et al. [3] use this concept to create a time-domain acoustic feature that measures voice intensity. Good results are obtained by training SVMs for gender recognition. In [1], SVMs are investigated for visual gender classification with low resolution. Authors use human subjects to test the performance of the SVMs and SVMs performed better than humans.

Aggarwal et al. [22] create emotion and gender recognition application by training SVMs and Naive Bayes on four speech features: shimmer, jitter, energy, and pitch. Authors observed SVMs outperformed Naive Bayes with 65% accuracy in emotion recognition and 70% accuracy in gender recognition. Their application can be enhanced by using features such as Mel frequency cepstral coefficients (MFCCs) which has vital acoustic information in any recognition system [10], [11]. A voice-based gender recognition system is proposed in [24]. Authors investigate spectral features with a different number of dimensions on gender recognition systems for two datasets. An accuracy of 80% and 65% is obtained using MFCC60 for a cross-corpus experiment of data set A to B and reverse respectively.

III. Methodology

The architecture of a gender recognition system is shown in Fig. 1. The system contains two major components: the training and prediction phases. In the training phase, the speech signal is inputted to the system where features are extracted. Then a machine learning model is trained on the labelled features. In the prediction phase, a speech signal is inputted to the system without a gender. The model predicts and outputs the gender of the signal.

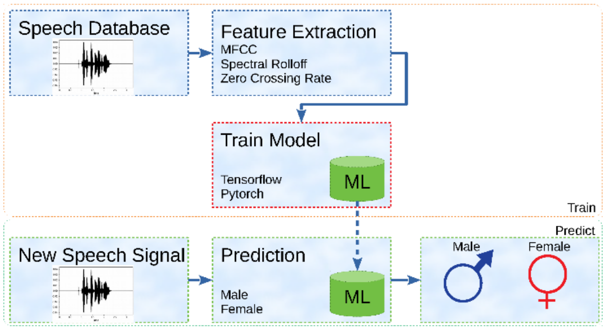


Fig. 1: Architecture of a gender recognition system.

A. Data

We acquired speech corpus from [25] used in [8], where recordings consists of 33 native speakers (16 female and 17 male) of Yorùbá. The recordings were recorded in Lagos, Nigeria. About 130 utterances (approximately 5 minutes of audio) were read from short texts selected for phonetic coverage. The total number of utterances is 4316 and the total length of audio is 165 minutes. Recordings are of 16-bit PCM samples at 16 kHz sampling rate. The dataset is split into 90% for training and 10% for testing. The model is trained for 100 epochs and involved 3884 samples for training and 432 samples for validation.

Fig. 2 shows the dataset scaled to two dimension using Principal Component Analysis (PCA) [26] [27] and the centers are shown using k-means [27] with *k=2*. We observe the data can be separated into males and females. This will simplify the learning of the models.

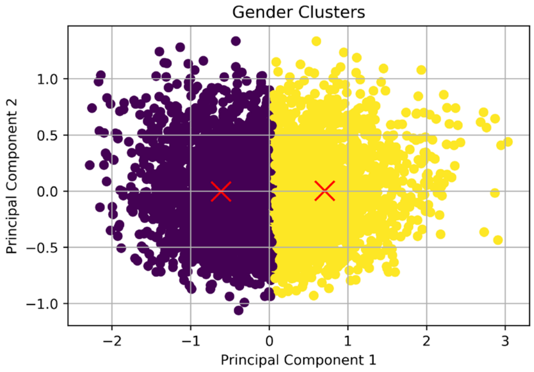


Fig. 2: Dataset scaled to 2-dimensional using PCA. Males in purple and females in yellow.

B. Feature Extraction

To discriminate gender from recorded speech sound, meaningful acoustic features must be extracted from the waveform of the sound. Audio feature extraction methods have been presented for different audio recognition systems. For speech, we use Time, Frequency, and Cepstral-domain features listed in Fig. 3 also defined in [28].

Feature ID	**Feature Name**	**Description**
1	Zero Crossing Rate	The rate of sign-changes of the signal during the duration of a particular frame.
2	Energy	The sum of squares of the signal values, normalized by the respective frame length.
3	Entropy of Energy	The entropy of sub-frames' normalized energies. It can be interpreted as a measure of abrupt changes.
4	Spectral Centroid	The center of gravity of the spectrum.
5	Spectral Spread	The second central moment of the spectrum.
6	Spectral Entropy	Entropy of the normalized spectral energies for a set of sub-frames.
7	Spectral Flux	The squared difference between the normalized magnitudes of the spectra of the two successive frames.
8	Spectral Rolloff	The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.
9-21	MFCCs	MFCCs form a cepstral representation where the frequency bands are distributed according to the mel-scale.
22-33	Chroma Vector	A 12-element representation of the spectral energy where the bins represent the 12 equal-tempered pitch classes of western-type music (semitone spacing).
34	Chroma Deviation	The standard deviation of the 12 chroma coefficients.

Fig. 3: Acoustic features [28].

a) Time-domain features: (features 1-3) refers to how the signals change over time, they include Energy [23], Entropy of Energy, and Zero Crossing Rate (ZCR) [4], [23], that are extracted directly from the raw audio samples.

b) Frequency-domain features: (features 4-34, excluding the MFCCs) based on the magnitude of the Discrete Fourier Transform (DFT), these include Spectral Spread [4], Spectral Centroid [4], Spectral Flux, Spectral Entropy, Spectral Rolloff, Chroma Deviation and Chroma Vector.

c) Cepstral-domain features: include Mel Frequency Cepstral Coefficients (MFCCs) that result after the inverse DFT is applied on the logarithmic spectrum. MFCCs are popular audio features extracted from speech signals for use in speech, emotion, and speaker recognition [29], [6], [10], [11]. MFCCs are determined with the help of a psychoacoustically motivated filter bank, followed by logarithmic compression and discrete cosine transform.

C. Models

We train three types of neural networks: feed forward, convolutional and recurrent neural networks.

a) Multilayer Perceptron: is a class of feed-forward artificial neural network. MLP has an input layer, a hidden layer and an output layer. For the hidden layers, we use rectified linear unit as an activation function for each node and the output layer is activated by sigmoid. Tensorflow is used to implement MLP with dropout regularization is set to 50% to prevent overfitting. The model has a total of 25,602 parameters. Fig. 4a shows the architecture of the model with 3 layers containing 128, 128, and 2 units respectively.

b) Convolutional Neural Networks: are powerful regularized versions of MLPs popularly used in computer vision. CNNs use deep learning to perform both descriptive and generative tasks. Tensorflow is used to implement CNN with dropout regularization set at 50% to prevent overfitting. The model has a total of 141,838 parameters. The architecture of the model is depicted in Fig. 4b with four hidden connected layers of CNN followed by a pooling layer. The four CNN hidden layers contain 68, 32, 16, 8 units respectively. The output layer is activated by sigmoid.

c) LSTMs are recurrent neural networks with multiple hidden layers. This structure allows LSTM models to capture temporal information. Tensorflow is used to implement LSTM with dropout regularization set at 50% to prevent overfitting. The model has a total of 170,546 parameters. Fig. 4c shows the architecture of the model with three hidden connected layers of LSTM followed by a pooling layer. The three LSTM hidden layers contain 68, 32, 16 units respectively. The output layer is activated by sigmoid.

D. Evaluation

This section discusses the performance metrics used to assess the quality of the neural network models. The quality of the speech signal, the size of the training data, and most importantly the type of learning algorithm affects the performance of the model. We use the following evaluation metrics:

Accuracy represents all correctly identified examples from all the examples given. It is calculated as follows:

$$Accuracy = \frac{tpe + tne}{tpe + tne + fpe + fne} \tag{1}$$

Binary cross-entropy is a Sigmoid activation plus a Cross-Entropy loss. We use binary cross-entropy loss function since the data is categorical. It is calculated as follows:

$$-(y \log(p) + (1 - y) \times \log(1 - p)) \tag{2}$$

where *p* is the probability predicted by the model.

Precision is the total number of positively predicted examples that are relevant. It is calculated as follows:

$$Precision = \frac{tpe}{tpe + fpe} \tag{3}$$

Recall measures how well a model is at predicting the positives. It is calculated as follows:

$$Recall = \frac{tpe}{tpe + fne} \tag{4}$$

F_1 score is the harmonic mean of precision and recall. It is calculated as follows:

$$F_1 score = 2 \times \frac{precision \times recall}{precision + recall} \tag{5}$$

where:

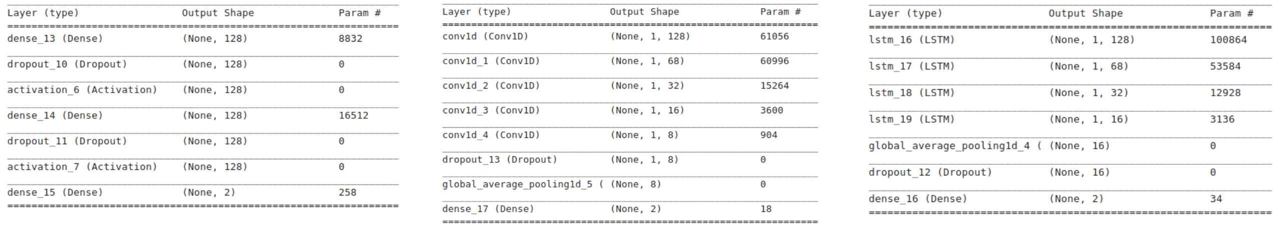
- *tpe* (true positive) is the number of males that are predicted as males.
- *tne* (true negative) is the number of females that are predicted as females.
- *fpe* (false positive) is the number of females examples that are predicted as males.
- *fne* (false negative) is the number of males examples that are predicted as females.

IV. Results and Discussions

This section discusses model performance results based on accuracy, F_1 score and binary cross-entropy. We extract a total of 34 short-term acoustic features (frame size is set to 50ms at a rate of 25ms with the Hamming window) using an open-source comprehensive library called pyAudioAnalysis [28]. The feature vector contains the mean and standard deviation of the 34 short-term acoustic features, totaling to a vector of length 68.

A. Performance

Table I shows the testing results after evaluating the models. We observe CNN struggling to discriminate training data compared to other models with training accuracy of around 90% in Fig. 5. But when testing CNN, the accuracy went high to 98.8%. The MLP and LSTM obtained testing accuracy of 99%. We observe in Table I all the models obtained F_1score of 99%. Even though the data is not balanced, the accuracy and F_1score are almost the same.



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_13 (Dense)             (None, 128)               8832
_________________________________________________________________
dropout_10 (Dropout)         (None, 128)               0
_________________________________________________________________
activation_6 (Activation)    (None, 128)               0
_________________________________________________________________
dense_14 (Dense)             (None, 128)               16512
_________________________________________________________________
dropout_11 (Dropout)         (None, 128)               0
_________________________________________________________________
activation_7 (Activation)    (None, 128)               0
_________________________________________________________________
dense_15 (Dense)             (None, 2)                 258
=================================================================
```

(a) MLP

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d (Conv1D)              (None, 1, 128)            61056
_________________________________________________________________
conv1d_1 (Conv1D)            (None, 1, 68)             60996
_________________________________________________________________
conv1d_2 (Conv1D)            (None, 1, 32)             15264
_________________________________________________________________
conv1d_3 (Conv1D)            (None, 1, 16)             3600
_________________________________________________________________
conv1d_4 (Conv1D)            (None, 1, 8)              904
_________________________________________________________________
dropout_13 (Dropout)         (None, 1, 8)              0
_________________________________________________________________
global_average_pooling1d_5 ( (None, 8)                 0
_________________________________________________________________
dense_17 (Dense)             (None, 2)                 18
=================================================================
```

(b) CNN

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_16 (LSTM)               (None, 1, 128)            100864
_________________________________________________________________
lstm_17 (LSTM)               (None, 1, 68)             53584
_________________________________________________________________
lstm_18 (LSTM)               (None, 1, 32)             12928
_________________________________________________________________
lstm_19 (LSTM)               (None, 1, 16)             3136
_________________________________________________________________
global_average_pooling1d_4 ( (None, 16)                0
_________________________________________________________________
dropout_12 (Dropout)         (None, 16)                0
_________________________________________________________________
dense_16 (Dense)             (None, 2)                 34
=================================================================
```

(c) LSTM

Fig. 4: Architecture of the models.

Thus, this shows that neural networks are among the best models for gender recognition using speech only.

TABLE I
RESULTS AFTER TESTING THE MODELS.

Model	Accuracy	F_1 score
MLP	0.991	0.99
CNN	0.988	0.99
LSTM	0.991	0.99

B. *Overfitting*

In investigating overfitting, learning curves are illustrated in Fig. 5 for MLP, CNN, and LSTM. The curves for testing are above the curves for training and did not decrease. Hence, the models are not underfit. Moreover, Fig. 6 shows the binary cross-entropy loss function. The loss function kept decaying for all the models but for CNN, the loss function is decaying at a lower rate. We observe CNN is around 0.2 while other models are further below 0.1. Hence, all the models did not overfit.

In Fig. 7 we investigate if changing the threshold will affect accuracy. We use What-If toolkit [4] to load the probability predictions of the models on the test data. Generally, for binary classification 0.5 is used as a threshold to indicate the class, that is, the class is 1 when the probability is greater than threshold and 0 when the probability is less than the threshold. We observe in Fig 7a and Fig. 7b that 100% accuracy can be achieved when changing a threshold to divide the two labels. For MLP, a threshold of 0.7 gives 100% accuracy, and for CNN, a threshold of 0.3 gives 100% accuracy. For LSTM in Fig. 7c, 100% accuracy cannot be achieved even though 4 samples were mislabeled, this can be reduced to 1 sample to be mislabeled by increasing threshold to 0.58. As a result, LSTM accuracy is now increased to 99.8%.

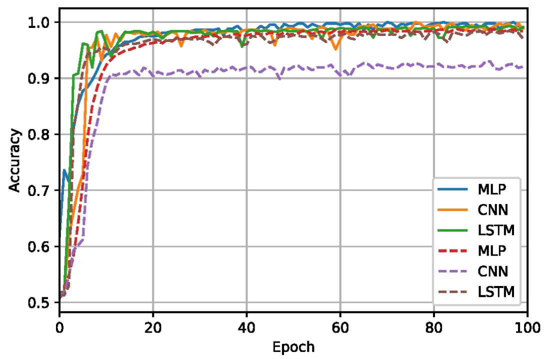


Fig. 5: Accuracy for MLP, CNN, and LSTM.

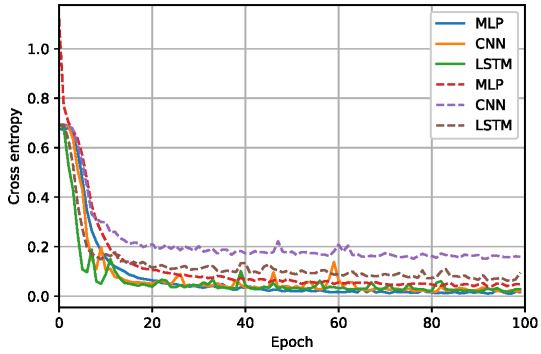


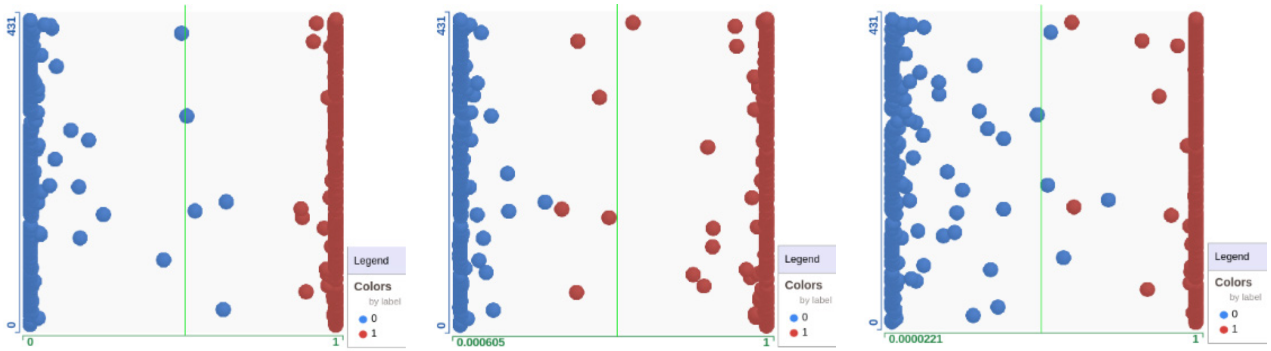
Fig. 6: Binary cross entropy for MLP, CNN, and LSTM.

V. CONCLUSION

This paper presented a Yorùbá gender recognition from speech using neural networks. We presented the literature on gender recognition. The acoustic features were explained. We explained the learning algorithms. We observed neural networks achieving the state-of-the-art accuracy of 99% for a low-resourced language.

The future work will focus on investigating what made the models to make wrong predictions for gender recognition.

[4] https://pair-code.github.io/what-if-tool/



(a) MLP (b) CNN (c) LSTM

Fig. 7: Model prediction probabilities of each test sample after adjusting the threshold.

Acknowledgment

This work was facilitated at the Council for Scientific and Industrial Research and we thank Department of Science and Technology Inter-Programme Bursary Support (IBS) for the funding support.

References

[1] B. Moghaddam and M.-H. Yang, "Gender classification with support vector machines," in *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, 2000.

[2] W. Hwang, H. Ren, H. Kim, S. Kee and J. Kim, "Face recognition using gender information," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, 2009.

[3] M. Alhussein, Z. Ali, M. Imran and W. Abdul, "Automatic gender detection based on characteristics of vocal folds for mobile healthcare system," *Mobile Information Systems,* vol. 2016, 2016.

[4] C. Chen, P. Lu, M. Hsia, J. Ke and O. T. Chen, "Gender-to-Age hierarchical recognition for speech," in *2011 IEEE 54th International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2011.

[5] A. Mukherjee and B. Liu, "Improving gender classification of blog authors," in *Proceedings of the 2010 conference on Empirical Methods in natural Language Processing*, Grenoble, France, 2010.

[6] A. Atanda, S. Yusof and M. Hariharan, "Yorùbá automatic speech recognition: A review," in *Rural ICT Development (RICTD) International Conference*, 2013.

[7] L. Besacier, E. Barnard, A. Karpov and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication,* vol. 56, pp. 85-100, 2014.

[8] D. van Niekerk and E. Barnard, "Tone realisation in a Yorùbá speech recognition corpus," in *Third Workshop on Spoken Language Technologies for Under-resourced Languages*, Cape Town, South Africa, 2012.

[9] T. I. Modipa and M. H. Davel, "Predicting vowel substitution in code-switched speech," in *2015 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, 2015.

[10] P. J. Manamela, M. J. Manamela, T. I. Modipa, T. J. Sefara and T. B. Mokgonyane, "The Automatic Recognition of Sepedi Speech Emotions Based on Machine Learning Algorithms," in *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*, 2018.

[11] T. B. Mokgonyane, T. J. Sefara, T. I. Modipa, M. M. Mogale, M. J. Manamela and P. J. Manamela, "Automatic Speaker Recognition System based on Machine Learning Algorithms," in *2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)*, 2019.

[12] T. Sefara, P. Malatji and M. Manamela, "Speech Synthesis Applied to Basic Mathematics as a Language," in *South Africa International Conference on Educational Technologies*, 2016.

[13] T. J. Sefara, T. B. Mokgonyane, M. J. Manamela and T. I. Modipa, "HMM-based Speech Synthesis System incorporated with Language Identification for Low-resourced Languages," in *International Conference on Advances in Big Data, Computing and Data Communication Systems*, 2019.

[14] T. J. Sefara, M. J. Manamela and T. I. Modipa, "Web-based Automatic Pronunciation Assistant," in *Southern Africa Telecommunication Networks and Applications Conference (SATNAC)*, 2017.

[15] T. J. Sefara and M. J. Manamela, "The development of local synthetic voices for an automatic pronunciation assistant," in *Southern Africa Telecommunication Networks and Applications Conference (SATNAC)*, 2016.

[16] T. J. Sefara, M. J. Manamela and P. T. Malatji, "Text-based language identification for some of the under-resourced languages of South Africa," in *2016 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, 2016.

[17] S. Kumar, S. Singh and J. Kumar, "Gender Classification Using Machine Learning with Multi-Feature Method," in *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*, 2019.

[18] Z. Ding and Y. Ma, "Manifold-based face gender recognition for video," in *Proceedings of 2011 International Conference on Computer Science and Network Technology*, 2011.

[19] J. Chen, S. Liu and Z. Chen, "Gender classification in live videos," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017.

[20] D. Bales, P. A. Tarazaga, M. Kasarda, D. Batra, A. G. Woolard, J. D. Poston and V. Malladi, "Gender Classification of Walkers via Underfloor Accelerometer Measurements," *IEEE Internet of Things Journal,* vol. 3, no. 6, pp. 1259-1266, Dec 2016.

[21] A. Gümüşçü, K. Karadağ, M. Çalişkan, M. E. Tenekecı and D. Akaslan, "Gender classification via wearable gait analysis sensor," in *2018 26th Signal Processing and Communications Applications Conference (SIU)*, Izmir, Turkey, 2018.

[22] G. Aggarwal and R. Vig, "Acoustic Methodologies for Classifying Gender and Emotions using Machine Learning Algorithms," in *2019 Amity International Conference on Artificial Intelligence (AICAI)*, 2019.

[23] Y. Zhan, H. Leung, K. Kwak and H. Yoon, "Automated Speaker Recognition for Home Service Robots Using Genetic Algorithm and Dempster-Shafer Fusion Technique," *IEEE Transactions on Instrumentation and Measurement,* vol. 58, no. 9, pp. 3058-3068, Sep. 2009.

[24] I. Kanani, H. Shah and S. H. Mankad, On the Performance of Cepstral Features for Voice-Based Gender Recognition, Springer, 2019, pp. 327-333.

[25] D. van Niekerk, E. Barnard, O. Giwa and A. Sosimi, Lagos-NWU Yoruba Speech Corpus, North-West University, 2015.

[26] B. Moore, "Principal component analysis in linear systems: Controllability, observability, and model reduction," *IEEE Transactions on Automatic Control,* vol. 26, no. 1, pp. 17-32, February 1981.

[27] C. Ding, X. He and ACM, "K-means clustering via principal component analysis," in *Proceedings of the twenty-first international conference on Machine learning*, 2004.

[28] T. Giannakopoulos, "pyaudioanalysis: An open-source Python library for audio signal analysis," *PloS one,* vol. 10, no. 12, 2015.

[29] V. Tiwari, "MFCC and its applications in speaker recognition," *International journal on emerging technologies,* vol. 1, no. 1, pp. 19-22, 2010.