Yorùbá Gender Recognition from Speech using Attention-based BiLSTM

Ibukunola A. Modupe
Department of ICT
Vaal University of Technology
ibukunolam@vut.ac.za

Tshephisho J. Sefara
Next Generation Enterprises and Institutions
Council for Scientific and Industrial Research
tsefara@csir.co.za

Sunday O. Ojo
Department of Computer Science
Tshwane University of Technology
ojoso@tut.ac.za

Abstract

Gender recognition in speech processing is one of the most challenging tasks. While many studies rely on extracting features and designing enhancement classifiers, classification accuracy is still not satisfactory. The remarkable improvement in performance achieved through the use of neural networks for automatic speech recognition has encouraged the use of deep neural networks in other voice techniques such as speech, emotion, language and gender recognition. An earlier study showed a significant improvement in the gender recognition of pictures and videos. In this paper, speech is used to create a gender recognition scheme based on neural networks. Attention-based BiLSTM architecture is proposed to discover the best approach for gender identification in Yorùbá. Acoustic features, including time, frequency, and cepstral features are extracted to train the model. The model obtained the state-of-the-art performance in speech-based gender recognition with 99% accuracy and F_1 score.

1 Introduction

Gender recognition is an important topic in signal processing and can be applied in mobile healthcare system (Alhussein et al., 2016), facial recognition (Hwang et al., 2009), and age classification (Chen et al., 2011). Applications of gender recognition system includes tasks such as (Mukherjee and Liu, 2010): (i) Verifying a customer when making telephone bank transaction, (ii) Security measure when retrieving confidential information, (iii) Forensic, (iv) Surveillance, (v) and Blog authorship. Recognition of gender from the speech is a challenging task with these increasing number of systems in real-life. Recent hardware and software development allowed new techniques and methods to be explored to improve the efficiency of gender recognition systems. Gender classification systems from speech signal are affected by the performance of the recording tools, the language of the speaker, and noisy recording settings. As a result, to obtain adequate classification results, gender recognition from speech signals requires valid classifiers and feature extractors. In the areas of machine learning and computer vision, deep neural networks (DNNs) have shown notable achievements (Moghaddam and Ming-Hsuan Yang, 2000; Hwang et al., 2009). Deep neural networks, after thorough training, can effectively extract and classify different feature sets. DNNs are most effective when the training set contains a complicated feature space that needs high-level representation. In this paper, deep recurrent neural networks (DRNNs) are used as classifiers and gender-recognition extractors. Bidirectional long-short term memory (BiLSTM) is combined with an attention mechanism to learn the features. Because gender recognition is a binary classification, a sigmoid activation function has been used to classify the gender.

1.1 Motivation

Gender recognition systems for well-resourced languages like English are available, but for African languages like Yorùbá are not available. Yorùbá is a Niger-Congo language related to Igala, Edo, Ishan, and Igbo amongst others. It is one of the official languages of Nigeria and spoken in a couple of countries on the West African coast. An estimated 20+ million people speak Yorùbá as their first language in southwestern Nigeria and more in the Republics of Benin and Togo. Yorùbá is also spoken by diaspora communities of traders in Cote d'Ivoire, Ghana, Senegal and the Gambia, and it used to be a vibrant language in Freetown, Sierra Leone. Outside West Africa, millions of people have Yorùbá language and culture as part of their heritage; Yorùbá religion being one of the

means of survival in Cuba during the obnoxious slave trade. Many who did not have Yorùbá as their heritage bought into Yorùbá identity through religious transformation. Yorùbá language, culture and religion survived since then until now in Brazil and various other New World countries (Atanda et al., 2013; Pulleyblank et al., 2017). Yorùbá is identified as one of the under-resourced languages (Besacier et al., 2014), few systems for under-resourced African languages has been developed (Sefara et al., 2016; Sefara et al., 2019; Sefara et al., 2017; Sefara and Manamela, 2016; Sefara et al., 2016; Van Niekerk and Barnard, 2012; Modipa and Davel, 2015; Manamela et al., 2018; Mokgonyane et al., 2019). While the development of speech-based systems for Yorùbá is an open research, it is essential to continue to create a Yorùbá gender recognition system that may later help other researchers and to strengthen the cultural identify of the language.

The main contributions of this paper can be listed as below.

- A new classifier architecture is proposed. A BiLSTM architecture with attention mechanism is used.
- Acoustic features such as Time, Frequency, and Cepstral-domain features are used for gender recognition.
- We release the code[1] used in this paper.

The rest of the paper is organized as follows: Section 2 gives the literature review on gender recognition. Section 3 details the features, learning models, and evaluation methods. Section 4 discusses the experimental results, and the paper is concluded in Section 5.

2 Literature Review

Gender recognition can be approached from text (Mukherjee and Liu, 2010), images (Moghaddam and Ming-Hsuan Yang, 2000; Hwang et al., 2009; Kumar et al., 2019; Qawaqneh et al., 2017a), videos (Ding and Ma, 2011; Chen et al., 2017), accelerometers (Bales et al., 2016), wearables (Gümüşçü et al., 2018), and speech (Harb and Chen, 2003; Azghadi et al., 2007; Meena et al., 2013) to train machine learning models and neural networks for classification. Meena et al. (2013) proposed a novel gender classification technique in speech processing using neural network and fuzzy logic. Authors used acoustic features such as short time energy, zero crossing rate and energy entropy. Their work can be expanded by not only using time domain features but also to include frequency and cepstral domain features. An example of cepstral-domain features are Mel Frequency Cepstral Coefficients (MFCCs). Qawaqneh et al. (2017a) used MFCCs, fundamental frequency (F0) and the shifted delta cepstral coefficients (SDC) to train a jointly fine-tuned deep neural networks. Their model obtained accuracy of 64%. Conversely, Harb and Chen (2003) did not use MFCCs but used Mel Frequency Spectral Coefficients (MFSC) to train a gender identification system using neural networks. Authors showed that smoothing improves the accuracy of the model and MFSC features were better than MFCC features. Azghadi et al. (2007) used acoustic features and pitch features to train a gender classification system based on feed-forward back-propagation neural network. Their model obtained an accuracy of 96%. Qawaqneh et al. (2017b) introduced shared class labels among misclassified labels to regularize the DNN weights and to generate transformed MFCCs feature set using Backus-Naur Form (BNF). Authors used DNN and i-vector models to build age and gender classification system. The BNF-DNN obtained accuracy of 58.98 and BNF-I-vector obtained 56.13

Machine learning algorithm are used for gender recognition. Chaudhary and Sharma (2018) used support vector machines (SVMs) to train a gender identification system based on voice signal by extracting the features such as pitch, energy and MFCC. Their model obtained accuracy of 96.45%. Gaussian mixture models (GMMs) and multilayer perceptrons (MLPs) are used in (Djemili et al., 2012) to create a gender identification system. The models obtained accuracy of 96.4% using MFCCs as features. Jadav (2018) proposed a voice-based gender identification using machine learning. Author extracted acoustic features to train a SVM which obtained testing accuracy of 97%.

3 Methodology

The architecture of a gender recognition system is shown in Figure 1. The system consists of the training and prediction phases.

[1] https://github.com/SefaraTJ/yoruba-gender-recognition/

- In the training phase, the speech signal is inputted to the system, and pre-processing occurs (noise removal, dimensionality reduction). Acoustic features are extracted. Then a machine learning model is built and trained on the extracted features.

- In recognition phase, an unlabelled or unknown speech signal is inputted to the system. The model predicts and outputs the gender of the inputted signal.

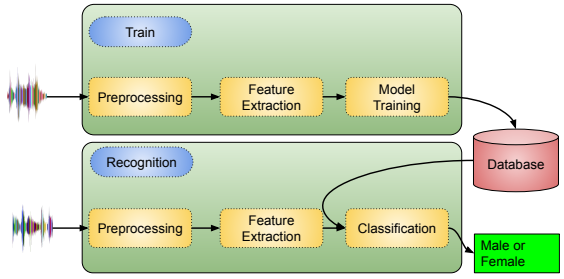


Figure 1: Architecture of a gender recognition system.

3.1 Data

We obtained speech database from (van Niekerk et al., 2015) used in (Van Niekerk and Barnard, 2012), where recordings consist of 16 female and 17 male recordings in Yorùbá. About 130 utterances were read from short texts for each speaker. The length of the recordings is 165 minutes. The audios are 16 bit PCM at 16kHz sampling rate.

We use Principal Component Analysis (PCA) (Moore, 1981; Ding and He, 2004) to explore the data in Figure 2 by scaling to 2 dimension. The centers are illustrated using k-means (Ding and He, 2004) with $k = 2$. We observe the data can be separated into males and females. This will simplify the learning of the models.

3.2 Feature Extraction

Feature extraction is the transformation of original data into a dataset that contains the most discriminatory information, with reduced numbers of variables. The 34 acoustic features shown in Figure 3 are extracted from the short-term windows with frame size of 50ms at a Hamming window of 25ms using a library in (Giannakopoulos, 2015). The final feature vector contains the mean and standard deviation which sums to feature size of 68. The features can be grouped into three categories:

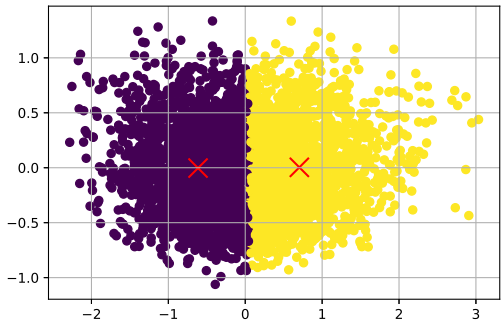


Figure 2: PCA showing gender clusters and k-means showing cluster centres.

- Time-domain features (Zero Crossing Rate, Energy, and Entropy of Energy).

- Frequency-domain features (Spectral Spread, Spectral Centroid, Spectral Flux, Spectral Entropy, Spectral Rolloff, Chroma Deviation, Chroma Vector).

- Cepstral-domain features - includes MFCCs that has an ability to model the vocal tract filter.

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.

Figure 3: Acoustic features (Giannakopoulos, 2015).

3.3 Feature Normalization

Is an crucial step for gender recognition using speech. The goal is to remove speaker and record-

ing variability. We normalize features by removing the mean and scaling to a unit variance using the following normalization equation. For normalized feature $\hat{y}$:

$$\hat{y} = \frac{x - \mu}{\sigma} \tag{1}$$

where σ represents the variance and μ represents the mean for each feature vector x.

3.4 The Classifier Model

This section explains the proposed BiLSTM model. As shown in Figure 4, the first layer is the input layer having the same size of the input vector. Followed by the BiLSTM layer having 128 units. Followed by the attention layer, followed by LSTM layer, followed by 4 dense layers with the last layer activated by the *sigmoid* function.

3.4.1 BiLSTM Layer

For this gender recognition problem, we model the speech signal using recurrent neural network (RNN), specifically BiLSTM. LSTM was introduced by Hochreiter and Schmidhuber (1997), has shown to be stable and accurately model long-time dependencies in different tasks like speech recognition, machine learning, and computer vision (Moghaddam and Ming-Hsuan Yang, 2000; Hwang et al., 2009). BiLSTM trains two LSTMs on the input sequence. The second LSTM is a reverse copy of the first one, the aim is to capture past and future input features for a specific time step.

3.4.2 Attention Layer

Attention is a mechanism allowing neural networks to examine specific areas of the input speech signal in more detail to decrease the task complexity and to exclude irrelevant information. An attention layer is included for determining the contribution of each signal frame to the whole speech signal. The attention mechanism assigns a weight w_i to each frame feature h_i. The hidden state is lastly calculated by a weighted sum function to generate a hidden acoustic feature vector r. Formally:

$$p_j = \tanh\left(W_h h_j + b_h\right), \quad p_j \in [-1, 1] \tag{2}$$

$$w_j = \frac{exp(p_j)}{\Sigma_{t=1}^{N} exp(p_t)}, \quad \Sigma_{j=1}^{N} w_j = 1 \tag{3}$$

$$r = \Sigma_{j=1}^{N} w_j h_j, \quad r \in R^{2L} \tag{4}$$

where W_h and b_h are the weight and bias from the attention layer.

3.4.3 Dense Layer

The attention layer is followed by four dense layers with different sizes of neurons. The output of attention layer is fed into first dense layer with 128 hidden neurons activated by *rectified linear unit*. And to avoid overfitting, we add a dropout layer having probability of 0.5 between the first three dense layers that have 128, 64, and 32 neurons respectively. The last dense layer uses *sigmoid* activation function to create binary classification. The *sigmoid* activation function is defined as follows:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{5}$$

3.5 Evaluation

This section describes the performance measurements used to evaluate model quality. The performance of the model is affected by the speech signal quality, the training data size, and most importantly the optimization of learning algorithm. The following evaluation metrics are applied:

Accuracy represents all correctly predicted samples, calculated as follows:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{6}$$

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

$$-(y \log(p) + (1 - y) \times \log(1 - p)) \tag{7}$$

where p is the probability predicted by the model.

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

$$Precision = \frac{tp}{tp + fp} \tag{8}$$

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

$$Recall = \frac{tp}{tp + fn} \tag{9}$$

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

$$F_1 score = 2 \times \frac{precision \times recall}{precision + recall} \tag{10}$$

where:

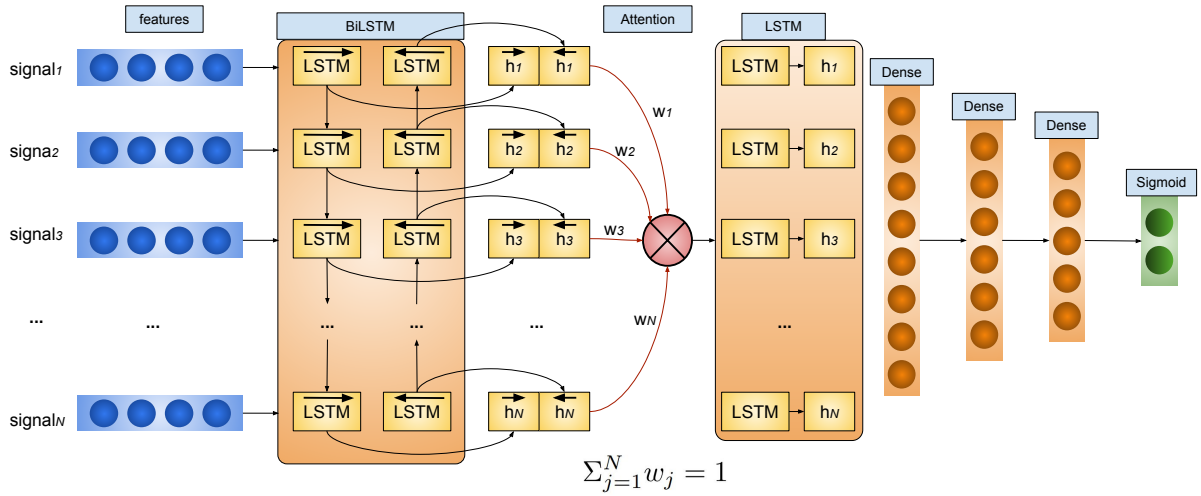


Figure 4: Architecture of the BiLSTM with Attention Mechanism.

- tp (true positive) is the number of males that are predicted as males.
- tn (true negative) is the number of females that are predicted as females.
- fp (false positive) is the number of females examples that are predicted as males.
- fn (false negative) is the number of males examples that are predicted as females.

4 Results and Discussions

This section discusses model performance results based on accuracy, F_1 score and binary cross entropy. The dataset is splitted into 90% for training, 10% for testing. The model is trained for 200 epochs and involved 3884 samples for training and 432 samples for testing.

4.1 Performance

Table 1 shows the testing results after evaluating the model. We observe BiLSTM obtaining high accuracy and F_1 score of 99% after 200 epochs. The BiLSTM outperformed the neural network models in (Harb and Chen, 2003; Azghadi et al., 2007; Meena et al., 2013; Qawaqneh et al., 2017a,b). Even though Qawaqneh et al. (2017a) used both images + audio files, their performance does not beat the BiLSTM. Figure 5a shows the accuracy curve of the BiLSTM model. The accuracy of model increased as the number of epochs increase.

Table 1: Comparison with other models

Model	Accuracy
MLP (Harb and Chen, 2003)	92
MLP (Azghadi et al., 2007)	96
ANN + Fuzy Logic (Meena et al., 2013)	65
DNN (Qawaqneh et al., 2017a)	64
DNN (Qawaqneh et al., 2017b)	59
BiLSTM-Attention	**99**

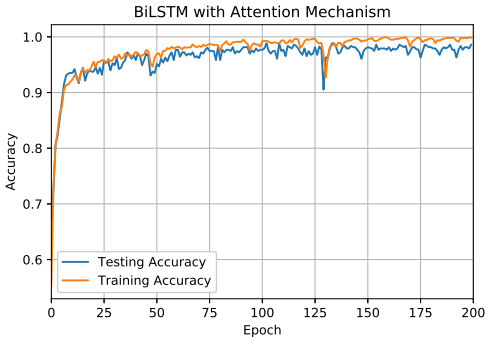
4.2 Overfitting

Overfitting happens when a model attempts to predict a trend in a noisy data. Overfitting is the consequence of a complicated model with excessive parameters. An overfitted model makes incorrect predictions as the trend does not represent the reality of the data. To show that overfitting is avoided, Figure 5b shows the binary cross entropy loss function curve. The loss function kept decreasing as number of training iterations increased. We observe BiLSTM reaching the lowest loss of 0.1 after 200 epochs. Hence, the model did not overfit.

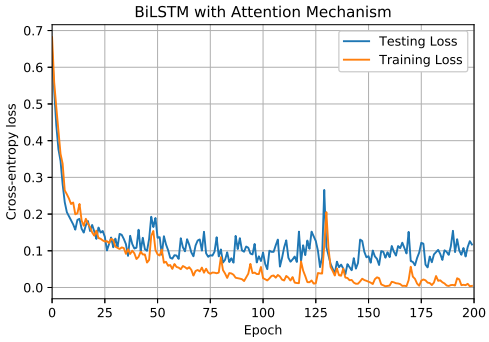
5 Conclusion

This paper presented a Yorùbá gender recognition from speech using BiLSTM with attention mechanism. We discussed the literature on gender recognition. The acoustic features were explained together with normalization method. We explained the architecture of the proposed model. We observed BiLSTM achieving the state-of-the-art accuracy of 99% for a low-resourced language.

The future work will focus on using transformer models for gender recognition.



(a) Accuracy



(b) Binary cross entropy

Figure 5: Model prediction Accuracy and estimated binary cross entropy for BiLSTM.

References

Musaed Alhussein, Zulfiqar Ali, Muhammad Imran, and Wadood Abdul. 2016. Automatic gender detection based on characteristics of vocal folds for mobile healthcare system. *Mobile Information Systems*, 2016.

Abdul Wahab Funsho Atanda, Shahrul Azmi Mohd Yusof, and M Hariharan. 2013. Yorùbá automatic speech recognition: A review. In *Rural ICT Development (RICTD) International Conference*, pages 116–121.

S Mostafa Rahimi Azghadi, M Reza Bonyadi, and Hamed Shahhosseini. 2007. Gender classification based on feedforward backpropagation neural network. In *Artificial Intelligence and Innovations 2007: from Theory to Applications*, pages 299–304, Boston, MA. Springer US.

D. Bales, P. A. Tarazaga, M. Kasarda, D. Batra, A. G. Woolard, J. D. Poston, and V. V. N. S. Malladi. 2016. Gender classification of walkers via underfloor accelerometer measurements. *IEEE Internet of Things Journal*, 3(6):1259–1266.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

S. Chaudhary and D. K. Sharma. 2018. Gender identification based on voice signal characteristics. In *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pages 869–874, Greater Noida (UP), India.

C. Chen, P. Lu, M. Hsia, J. Ke, and O. T. . Chen. 2011. Gender-to-age hierarchical recognition for speech. In *2011 IEEE 54th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 1–4.

J. Chen, S. Liu, and Z. Chen. 2017. Gender classification in live videos. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1602–1606.

Chris Ding and Xiaofeng He. 2004. K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning*, page 29. ACM.

Zhengming Ding and Yanjiao Ma. 2011. Manifold-based face gender recognition for video. In *Proceedings of 2011 International Conference on Computer Science and Network Technology*, volume 2, pages 1104–1107.

R. Djemili, H. Bourouba, and M. C. A. Korba. 2012. A speech signal based gender identification system using four classifiers. In *2012 International Conference on Multimedia Computing and Systems*, pages 184–187.

Theodoros Giannakopoulos. 2015. pyaudioanalysis: An open-source Python library for audio signal analysis. *PloS one*, 10(12):1–17.

Abdülkadir Gümüşçü, Kerim Karadağ, Mustafa Çalişkan, Mehmet Emin Tenekecı, and Dursun Akaslan. 2018. Gender classification via wearable gait analysis sensor. In *2018 26th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE.

H. Harb and Liming Chen. 2003. Gender identification using a general audio classifier. In *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, volume 2, pages II–733, Baltimore, MD, USA.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

W. Hwang, H. Ren, H. Kim, S. Kee, and J. Kim. 2009. Face recognition using gender information. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 4129–4132.

S. Jadav. 2018. Voice-based gender identification using machine learning. In *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, pages 1–4.

S. Kumar, S. Singh, and J. Kumar. 2019. Gender classification using machine learning with multi-feature method. In *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0648–0653.

P. J. Manamela, M. J. Manamela, T. I. Modipa, T. J. Sefara, and T. B. Mokgonyane. 2018. 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)*, pages 1–7.

Kunjithapatham Meena, Kulumani Subramaniam, and Muthusamy Gomathy. 2013. Gender classification in speech recognition using fuzzy logic and neural network. *International Arab Journal of Information Technology (IAJIT)*, 10(5).

T. I. Modipa and M. H. Davel. 2015. Predicting vowel substitution in code-switched speech. In *2015 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, pages 154–159.

B. Moghaddam and Ming-Hsuan Yang. 2000. Gender classification with support vector machines. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 306–311.

T. B. Mokgonyane, T. J. Sefara, T. I. Modipa, M. M. Mogale, M. J. Manamela, and P. J. Manamela. 2019. 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)*, pages 141–146.

B. Moore. 1981. Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Transactions on Automatic Control*, 26(1):17–32.

Arjun Mukherjee and Bing Liu. 2010. Improving gender classification of blog authors. In *Proceedings of the 2010 conference on Empirical Methods in natural Language Processing*, pages 207–217. Association for Computational Linguistics.

Daniel van Niekerk, Etienne Barnard, Oluwapelumi Giwa, and Azeez Sosimi. 2015. Lagos-NWU Yoruba speech corpus.

Douglas Pulleyblank et al. 2017. Yoruba. In *The World's Major Languages*, pages 882–898. Routledge.

Zakariya Qawaqneh, Arafat Abu Mallouh, and Buket D. Barkana. 2017a. Age and gender classification from speech and face images by jointly fine-tuned deep neural networks. *Expert Systems with Applications*, 85:76–86.

Zakariya Qawaqneh, Arafat Abu Mallouh, and Buket D. Barkana. 2017b. Deep neural network framework and transformed MFCCs for speaker's age and gender classification. *Knowledge-Based Systems*, 115:5–14.

T. J. Sefara, M. J. Manamela, and P. T. Malatji. 2016. 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)*, pages 303–307.

T. J. Sefara, T. B. Mokgonyane, M. J. Manamela, and T. I. Modipa. 2019. 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*.

Tshephisho Sefara, Promise Malatji, and Madimetja Manamela. 2016. Speech synthesis applied to basic mathematics as a language. In *South Africa International Conference on Educational Technologies*, pages 243–253.

Tshephisho Joseph Sefara and Madimetja Jonas Manamela. 2016. The development of local synthetic voices for an automatic pronunciation assistant. In *Southern Africa Telecommunication Networks and Applications Conference (SATNAC)*.

Tshephisho Joseph Sefara, Madimetja Jonas Manamela, and Thipe Isaiah Modipa. 2017. Web-based automatic pronunciation assistant. In *Southern Africa Telecommunication Networks and Applications Conference (SATNAC)*, pages 112–117.

Daniel Van Niekerk and Etienne Barnard. 2012. Tone realisation in a Yorùbá speech recognition corpus. In *Third Workshop on Spoken Language Technologies for Under-resourced Languages*, pages 54–59, Cape Town, South Africa.