Gender Identification in Sepedi Speech Corpus

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

Tumisho Billson Mokgonyane
Department of Computer Science
University of Limpopo
Polokwane, South Africa
mokgonyanetb@gmail.com

***Abstract*—Gender identification is the task of identifying the gender of the speaker from the audio signal. Most gender identification systems are developed using datasets belonging to well-resourced languages. There has been little focus on creating gender identification systems for under resourced African languages. This paper presents the development of a gender identification system using a Sepedi speech dataset containing a duration of 55.7 hours made of 30776 males and 28337 females. We build a gender identification system using machine learning models that are trained using multilayer Perceptron (MLP), convolutional neural network (CNN), and long short-term memory (LSTM). Mid-term features are extracted from time domain features, frequency domain features and cepstral domain features, and normalised using the Z-score normalisation technique. XGBoost is used as a feature selection method to select important features. MLP achieved the same F-score and an accuracy of 94% for data with seen speakers while LSTM and CNN achieved the same F-score and an accuracy of 97%. We further evaluated the models on data with unseen speakers. All the models achieved good performance in F-score and accuracy.**

***Index Terms*—gender identification, convolutional neural network, Sepedi, XGBoost, feature selection, long short-term memory, multilayer Perceptron**

I. Introduction

Gender identification systems play an important role in the pre-processing step of speech recognition applications where gender identification identifies the gender of the speaker so that speech recognition application can automatically configure their parameters. In digital forensic investigations, gender identification systems help the investigator to identify the gender of the speaker. In voice-based identity recognition and speaker recognition systems [1], knowledge about the gender of the speaker improves such systems where the system restricted to one gender. Applications such as speaker diarisation, emotion recognition, biometrics social robots, and human-computer interaction make use of gender identification.

There are different methods of gender identification. Some are discussed in [2] but the authors only discussed recognition from facial images. Other methods of gender identification involve speech [3] and text [4]. Gender identification using speech involves the extraction of acoustic features such as time domain, frequency domain and cepstral domain features listed in [5], Linear Prediction Coefficients (LPC), Linear Predictive Cepstral Coefficient (LPCC), and spectrograms. Pahwa and Aggarwal [6] used MFCCs and their first and second-order derivatives when proposing a gender recognition system trained using a hybrid model involving support vector machine (SVM) and neural networks for Hindi. Jayasankar et al. [7] used time-domain features to create gender identification in speech recognition.

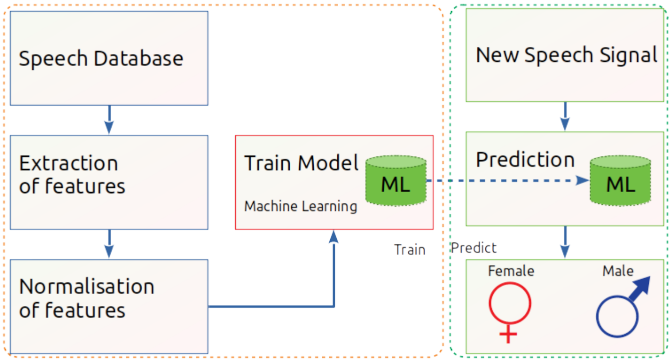


Fig. 1. Design of a gender identification system

Figure 1 shows the architecture of gender identification from speech. The speech database contains speech samples of female and male speakers. Acoustic features are extracted from the speech samples to create feature vectors that are normalised to remove speaker variability. Machine learning models are trained and used to predict the gender of the new speech signal.

The recent development of gender identification applications has focused on the dataset for well-resourced languages such as European languages. However, there has been little focus on low-resourced African languages. Few studies have worked on Sepedi, which is one of the official languages of South Africa [1], [8]. South Africa has a population of 59.62 million[1], of which 4.9 million of the population are native speakers of Sepedi, according to census 2011[2]. Geographically most speakers reside in the northern part of South Africa in Limpopo province.

Hence, we propose the development of a gender identification system using a Sepedi speech corpus trained on mid-term acoustic features selected by XGBoost. We train the model using neural networks such as multilayer Perceptron (MLP), long short-term memory (LSTM) and convolutional

[1] http://www.statssa.gov.za/?p=13453
[2] https://en.wikipedia.org/wiki/Northern_Sotho_language

neural network (CNN). We use the Sepedi NCHLT dataset [9], which has also been used for other speech application, such as speaker identification [1], [10]–[12].

The layout of the paper is organised as follows. The discussion of the literature is given in the second section followed by the discussion of the methodology including the data, feature extraction, normalisation and selection, machine learning models, and evaluation techniques in the third section. The fourth section explains the discussion of the findings, while the fifth section concludes the paper.

II. Literature Review

Neural networks are reported to produce a good performance on the classification of gender from speech. Dat et al. [3] used chroma, zero-crossing rate, Mel spectrogram, Mel-frequency cepstral coefficients, spectral contrast and tonnetz to create a CNN for gender and age identification of adults and children. The authors built a CNN model using TensorFlow where they compared different optimisation algorithms such as RMSProp, stochastic gradient descent, Adam, and Adagrad. The authors obtained the highest accuracy of 92% when using Adam for adult gender recognition.

Nugroho et al. [13] proposed a gender recognition system from speech using neural networks and Singular Value Decomposition (SVD) method of dimensionality reduction of features. The authors extracted used Mel-frequency cepstral coefficient (MFCC) features and applied dimensionality reduction using SVD. The three learning algorithms, support vector machine (SVM), logistic regression (LR), and neural network are trained on the features. The neural network model outperformed SVM and LR.

Sefara et al. [14] proposed a gender recognition system for the Yorùbá language that is an official language of Nigeria. The authors extracted time, frequency, and cepstral-domain features to train MLP, CNN and LSTM models. Their models produced good results as they compare to the literature. Furthermore, the authors in [15] improved their model to build an attention-based Bidirectional LSTM model that produced good results.

Abdulsatar et al. [16] proposed an age and gender recognition system using speech utterances. The authors extracted MFCCs and the first four formant frequencies to train a K-Nearest Neighbor (KNN) model. The authors obtained an accuracy of 50% when combining age and gender, and 66.66% for gender classification.

Although most studies use MFCCs as features, LPC is another feature that is used to train gender recognition models. Yusnita et al. [17] proposed an automatic gender recognition based on LPC. The authors extracted LPC and performed normalisation of the features. Two layer Feed-forward Multi-Layer Perceptron is created and trained using the Levenberg-Marquardt learning algorithm. The authors obtained an accuracy of 93.3%.

Zhong et al. [18] used decision trees to propose a gender recognition system while Nazifa et al. [19] used SVM and KNN to propose a gender recognition based on MFCC, LPC, and LPCC. SVM was trained on both features and their combination to evaluate which SVM kernel performs the best. Polynomial kernel performed better than other kernels and outperforming KNN when using a combination of LPCC and MFCC. On the other hand, Bhukya [20] applied MFCC and LPC to train a gender recognition system that is used to improve the performance of a speech recognition system. The authors observed good results and further noticed that the accent between speakers affects the quality of the speech recognition system.

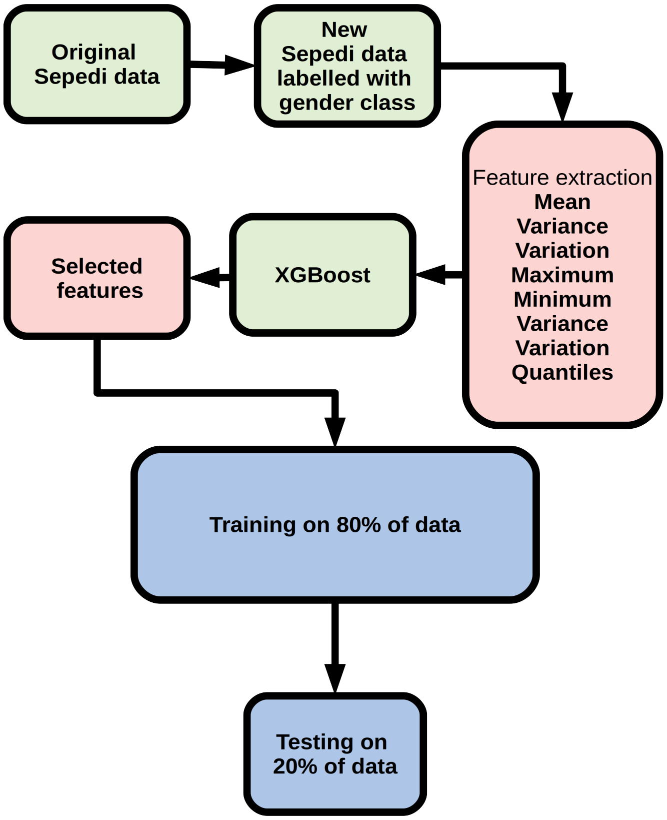


Fig. 2. Proposed system architecture

III. Methodology

In this section, we discuss the acquired data, feature extraction, feature normalisation, feature selection, machine learning methods and evaluation techniques. The proposed system is shown in Figure 2 where we label original Sepedi speech dataset with gender labels. Then we perform feature selection using XGBoost to train MLP, CNN, and LSTM on selected features. The models are trained on 80% of the data and 20% of the data is used for testing the models.

A. Data

The data has been acquired from the South African Centre for Digital Language Resources (SADiLaR)[3] which is a national centre supported by the South African government. SADiLaR contains a database of language technologies for the research and development of indigenous languages. We acquired the Sepedi dataset initially created for speech recognition. The corpus contains the metadata including the gender of the speakers. This enabled us to label the speech samples according to gender. There were 30776 males and 28337 females speech samples. The recordings have an average length of 5 seconds recorded at 16-bit mono PCM sampled at 16kHz.

B. Feature Extraction

We extracted 34 acoustic features listed in Table I fully explained in [5] with their deltas which increases the features to 68. The short-term features are extracted using a frame size of 50ms at a Hamming window size of 25ms for each speech sample to create a series of short-term feature vectors. Then we computed the mid-term features that are represented by statistics of the extracted short-term features. The mid-term features include the maximum, minimum, coefficient of variation, variance, standard deviation, mean, median, and quantiles (1st, 2nd, 3rd). Their equations are explained in Equation 1-10. The total number of the features is 680 which is 68 short-term features for each Equation 1-10.

- Mean of the short-term features

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1}$$

- Standard deviation of the short-term features

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} \tag{2}$$

- Variance of the short-term features

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \tag{3}$$

- Coefficient of Variation of the short-term features

$$CV = \frac{\sigma}{\mu} \tag{4}$$

- Minimum value of the short-term features

$$min(x) = arg_x min(x_i \dots x_N) \tag{5}$$

- Maximum value of the short-term features

$$max(x) = arg_x max(x_i \dots x_N) \tag{6}$$

- Median value of the short-term features

$$\begin{aligned} Median &= \frac{N+1}{2}\ when\ N\ is\ odd \\ &= \frac{\frac{N}{2} + \frac{N+1}{2}}{2}\ when\ N\ is\ even \end{aligned} \tag{7}$$

TABLE I
SHORT-TERM ACOUSTIC FEATURES [5]

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.

- Quantiles of the short-term features

$$Q_1 = \frac{1}{4}(N+1) \tag{8}$$

$$Q_2 = \frac{2}{4}(N+1) \tag{9}$$

$$Q_3 = \frac{3}{4}(N+1) \tag{10}$$

where N is the total number of short-term windows for each speech sample.

C. Feature Normalisation

The mid-term feature vectors contain values that are high and small. This may reduce the quality of the models. Hence, we use feature normalisation to remove the mean and scale to unit variance. This helps to reduce the speaker and recording variability. We use the Z-score normalisation defined as follows.

$$\hat{y} = \frac{x - \mu}{\sigma} \tag{11}$$

where $\hat{y}$ is a normalised feature, μ is a population mean, σ is a population variance for each mid-term feature vector x.

D. Feature Selection

We trained XGBoost model using all the features. XGBoost contains a technique to weight the features according to their importance. A feature is important when XGBoost provides a score that indicates how valuable a feature is when building

[3] https://www.sadilar.org/

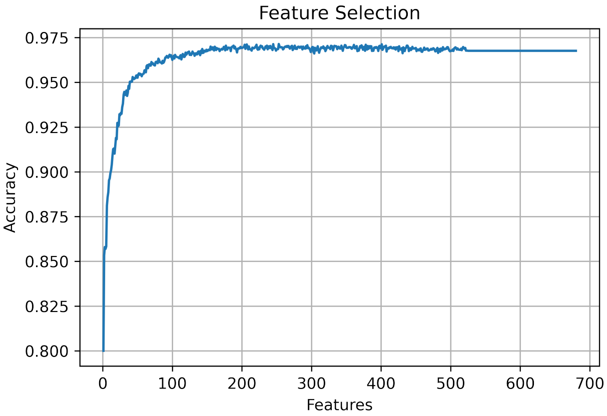


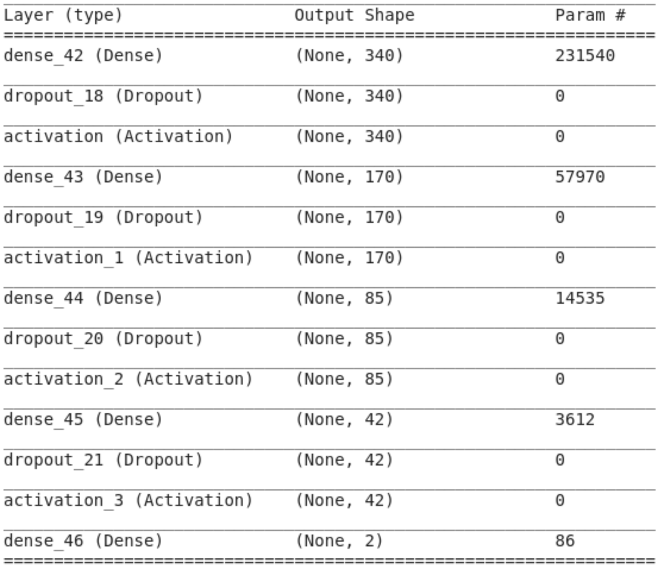
Fig. 3. Feature importance

boosted decision tree within the model. Figure 3 shows the accuracy when increasing number of features according to their importance. We concluded to select all the features as they positively contribute on classification performance.

E. Models

We use Tensorflow[4] sequence-to-sequence models to implement the MLP, CNN and LSTM for gender recognition.

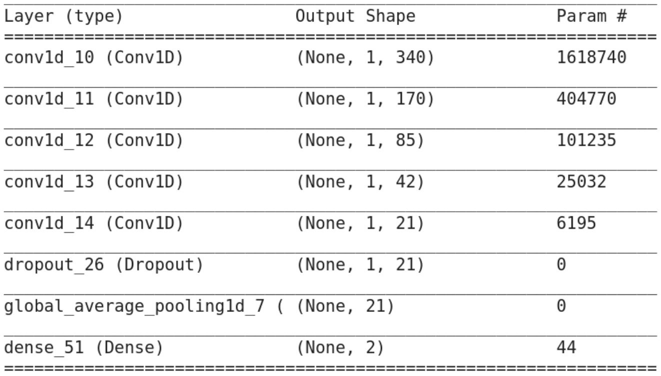
1) MLP: Figure 4 shows the architecture of the MLP. The architecture contains five dense layers where the first four are activated by rectified linear unit (RELU). The output layer is activated by the sigmoid activation function. The MLP model consists of 307,743 parameters.



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_42 (Dense)             (None, 340)               231540
_________________________________________________________________
dropout_18 (Dropout)         (None, 340)               0
_________________________________________________________________
activation (Activation)      (None, 340)               0
_________________________________________________________________
dense_43 (Dense)             (None, 170)               57970
_________________________________________________________________
dropout_19 (Dropout)         (None, 170)               0
_________________________________________________________________
activation_1 (Activation)    (None, 170)               0
_________________________________________________________________
dense_44 (Dense)             (None, 85)                14535
_________________________________________________________________
dropout_20 (Dropout)         (None, 85)                0
_________________________________________________________________
activation_2 (Activation)    (None, 85)                0
_________________________________________________________________
dense_45 (Dense)             (None, 42)                3612
_________________________________________________________________
dropout_21 (Dropout)         (None, 42)                0
_________________________________________________________________
activation_3 (Activation)    (None, 42)                0
_________________________________________________________________
dense_46 (Dense)             (None, 2)                 86
=================================================================
```

Fig. 4. The architecture of MLP model

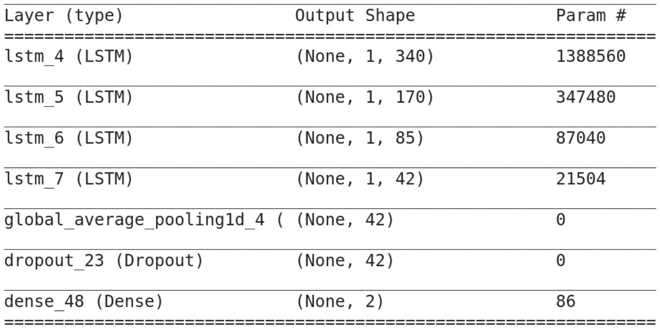
2) CNN: Figure 5 shows the architecture of the CNN. The architecture contains five CNN layers activated by the RELU and kernel size is set to seven. The next layer is the global pooling layer that come after a dense layer on the output activated by the a sigmoid activation function. The CNN model contains a total of 2,156,016 parameters.



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_10 (Conv1D)           (None, 1, 340)            1618740
_________________________________________________________________
conv1d_11 (Conv1D)           (None, 1, 170)            404770
_________________________________________________________________
conv1d_12 (Conv1D)           (None, 1, 85)             101235
_________________________________________________________________
conv1d_13 (Conv1D)           (None, 1, 42)             25032
_________________________________________________________________
conv1d_14 (Conv1D)           (None, 1, 21)             6195
_________________________________________________________________
dropout_26 (Dropout)         (None, 1, 21)             0
_________________________________________________________________
global_average_pooling1d_7 ( (None, 21)                0
_________________________________________________________________
dense_51 (Dense)             (None, 2)                 44
=================================================================
```

Fig. 5. The architecture of CNN model

3) LSTM: Figure 6 shows the architecture of the LSTM. The architecture contains four LSTM layers activated by the RELU that come after the global pooling layer followed by a dense layer on the output activated by a sigmoid activation function. The LSTM model contains a total of 1,844,670 parameters.



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_4 (LSTM)                (None, 1, 340)            1388560
_________________________________________________________________
lstm_5 (LSTM)                (None, 1, 170)            347480
_________________________________________________________________
lstm_6 (LSTM)                (None, 1, 85)             87040
_________________________________________________________________
lstm_7 (LSTM)                (None, 1, 42)             21504
_________________________________________________________________
global_average_pooling1d_4 ( (None, 42)                0
_________________________________________________________________
dropout_23 (Dropout)         (None, 42)                0
_________________________________________________________________
dense_48 (Dense)             (None, 2)                 86
=================================================================
```

Fig. 6. The architecture of LSTM model

F. Evaluation

The models are evaluated for quality using accuracy, weighted F_1 score, and binary cross-entropy loss defined as follows.

$$Accuracy = \frac{true\_P + true\_N}{true\_P + true\_N + false\_P + false\_N} \tag{12}$$

$$Recall = \frac{true\_P}{true\_P + false\_N} \tag{13}$$

$$Precision = \frac{true\_P}{true\_P + false\_P} \tag{14}$$

[4] https://www.tensorflow.org/

$$F_1 score = 2 \times \frac{recall \times precision}{recall + precision} \quad (15)$$

$$loss = -(y\log(p) + (1-y) \times \log(1-p)) \quad (16)$$

where:

- p is the probability predicted by the model.
- $true\_P$ (true positives) are males that are correctly identified.
- $true\_N$ (true negatives) are females that are correctly identified.
- $false\_P$ (false positives) are females that are identified as males.
- $false\_N$ (false negatives) are males that are identified as females.

IV. Findings and Discussions

The machine learning models are trained using 50 epochs where each epoch has a batch size of 32. From the original data, we extracted 338 speech samples of five speakers to be used later to evaluate the model on unseen data. The data was divided into 80% for training data and 20% for testing data. The training data is used to train the model and the testing data is used to evaluate the models. We use the same test set to evaluate the final model.

Table II illustrates the evaluation findings of the trained models on test data that contains speakers included in the training data. MLP obtained the lowest accuracy and $F_1 score$ of 94% compared to LSTM and CNN that obtained 97%. We show the learning curves in Figure 7 where the training performance are shown in the dashed line and testing performance are shown in solid line. We observed that MLP during training produced unstable accuracy per epoch while CNN and LSTM were stable after 25 epoch. Hence, both LSTM and CNN produced state-of-the-art results on gender classification on Sepedi speech data.

TABLE II
Results on Seen Speakers

Model	Accuracy	Weighted F_1score	Loss
MLP	0.94	0.94	0.26
LSTM	0.97	0.97	0.08
CNN	0.97	0.97	0.09

In Figure 8, we show the learning curve of the cross-entropy loss function. This curve helps to monitor if the models were overfitted. Since the learning curve did not show a negative slope but kept increasing for LSTM and CNN, then both models were not overfitted. But for MLP, the learning curve was not stable after 50 epochs. Increasing the number of epochs may overfit the model or converge the model. Since the models took a long time to finish training, we limited training iterations to 50 epochs.

We further evaluated the models on speech samples whose speakers were removed from the training data. Table III illustrates the evaluation findings of the trained models. We observed that MLP obtained an accuracy of 99% with a loss of 0.22. CNN achieved an accuracy of 98% with a loss of 0.05. LSTM achieved an accuracy of 95% with a loss of 0.11. Hence, the best models are CNN and LSTM because of the lowest loss value. The results of MLP can not be trusted since the model was not stable or did not converge.

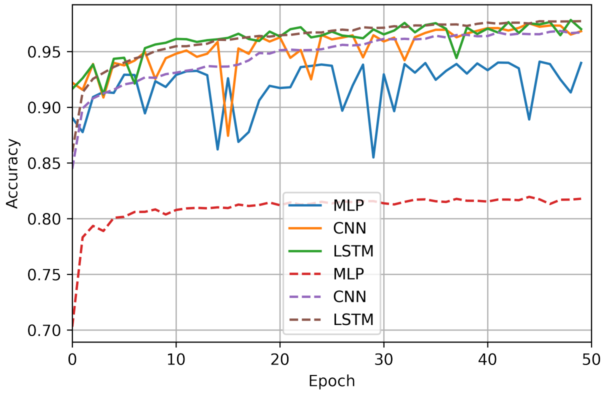


Fig. 7. The accuracy of MLP, CNN and LSTM algorithms on Sepedi data.

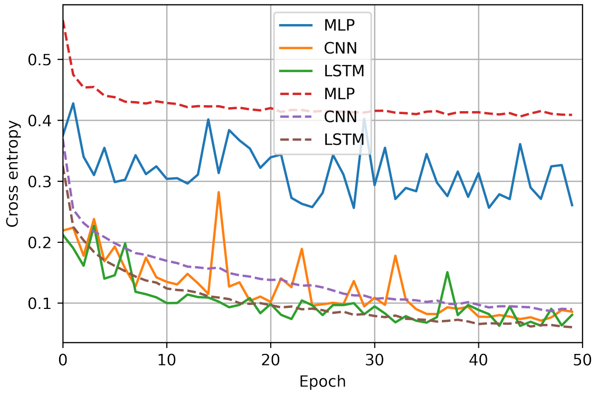


Fig. 8. The loss of MLP, CNN and LSTM algorithms on Sepedi data.

TABLE III
Results on Unseen Speakers

Model	Accuracy	Weighted F_1score	Loss
MLP	0.99	0.99	0.22
LSTM	0.95	0.95	0.11
CNN	0.98	0.98	0.05

Table IV shows the current performance of the models published in the literature. We compare our results in Table II with the results in Table IV. Our models performed better than the KNN classifier in [16] where authors used MFCC and the first four formant frequencies to build a gender recognition classifier that obtained an accuracy of 66.66%.

A CNN gender recognition model by Dat et al. [3] obtained an accuracy of 92% when authors used a subset of our features. Their model can be improved by adding more features since in our case the models obtained an additional 1% of accuracy for MLP and an additional 5% of accuracy for CNN and LSTM.

Yusnita, et al. [17] used the MLP model which obtained an accuracy of 93.3% when using LPC as features. Their accuracy is 0.7% lower than our MLP which obtained 94% and 3.7% lower than our LSTM and CNN which obtained 97%.

Liztio et al. [21] used fundamental frequency extracted from speech signal to train a gender identification model based on a backpropagation neural network. The authors obtained an accuracy of 95% that is 2% lower than our LSTM and CNN.

The difference in accuracy across different machine learning models is caused by (i) the usage of different features to train the model, (ii) parameters used to train the model, (iii) the type of algorithm used to train the model, and other factors not mentioned.

TABLE IV
COMPARISON WITH THE LITERATURE

Authors	Model	Accuracy	Features
[16]	KNN	66.66%	MFCC, first four formant frequencies
[3]	CNN	92%	Mel spectrogram, MFCC chroma, zero-crossing rate spectral contrast and tonnetz
[17]	MLP	93.3%	LPC
[21]	Backpropagation Neural Network	95%	fundamental frequency
Ours	**LSTM**	**97%**	**see Table I**
	CNN	**97%**	**see Table I**

V. CONCLUSIONS

We proposed a gender identification model trained using Sepedi NCHLT speech corpus containing a duration of 55.7 hours made of 30776 males and 28337 females. We acquired Sepedi NCHLT speech corpus from SADiLaR. Then we extracted the mid-term features. Then features were normalised using the Z-score normalisation method. We trained an XGBoost classifier to perform feature selection using its feature importance, and we noticed that all the features were significant in building the classifier. Then, we trained the gender identification model using MLP, CNN, and MLP based on all the features. The models were evaluated using accuracy and $F_1 score$ where CNN and LSTM obtained good results for gender identification in Sepedi speech corpus.

This work can be further extended by (i) increasing the dataset and using other low-resourced indigenous languages, (ii) investigating addition of other speech features not used in this work (iii) and fine-tuning the models.

REFERENCES

[1] T. B. Mokgonyane, T. J. Sefara, T. I. Modipa, and M. J. Manamela, "Automatic speaker recognition system based on optimised machine learning algorithms," in *2019 IEEE AFRICON*, 2019, pp. 1–7.

[2] V. Santarcangelo, G. M. Farinella, and S. Battiato, "Gender recognition: methods, datasets and results," in *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2015, pp. 1–6.

[3] P. T. Dat and L. The Anh, "Application of convolutional neural network for gender and age group recognition from speech," in *2019 6th NAFOSTED Conference on Information and Computer Science (NICS)*, 2019, pp. 489–493.

[4] A. Mukherjee and B. Liu, "Improving gender classification of blog authors," in *Proceedings of the 2010 conference on Empirical Methods in natural Language Processing*. Association for Computational Linguistics, 2010, pp. 207–217.

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

[6] A. Pahwa and G. Aggarwal, "Speech feature extraction for gender recognition," *International Journal of Image, Graphics and Signal Processing*, vol. 8, no. 9, p. 17, 2016.

[7] T. Jayasankar, K. Vinothkumar, and A. Vijayaselvi, "Automatic gender identification in speech recognition by genetic algorithm," *Appl. Math. Inf. Sci*, vol. 11, no. 3, pp. 907–913, 2017.

[8] T. J. Sefara, "The development of an automatic pronunciation assistant," Master's thesis, University of Limpopo, South Africa, 2019.

[9] E. Barnard, M. H. Davel, C. van Heerden, F. de Wet, and J. Badenhorst, "The NCHLT speech corpus of the South African languages," in *Proc. 4th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU)*, St Petersburg, Russia, May 2014.

[10] T. B. Mokgonyane, T. J. Sefara, M. J. Manamela, and T. I. Modipa, "Development of a text-independent speaker recognition system for biometric access control," in *Southern Africa Telecommunication Networks and Applications Conference (SATNAC)*, vol. 2018, 2018, pp. 128–133.

[11] T. B. Mokgonyane, T. J. Sefara, M. J. Manamela, T. I. Modipa, and M. S. Masekwameng, "The effects of acoustic features of speech for automatic speaker recognition," in *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*. IEEE, 2020, pp. 1–5.

[12] T. J. Sefara and T. B. Mokgonyane, "Emotional speaker recognition based on machine and deep learning," in *2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*. IEEE, 2020, pp. 1–8.

[13] K. Nugroho, E. Noersasongko, H. A. Santoso *et al.*, "Javanese gender speech recognition using deep learning and singular value decomposition," in *2019 International Seminar on Application for Technology of Information and Communication (iSemantic)*. IEEE, 2019, pp. 251–254.

[14] T. J. Sefara and A. Modupe, "Yorùbá gender recognition from speech using neural networks," in *2019 6th International Conference on Soft Computing Machine Intelligence (ISCMI)*, 2019, pp. 50–55.

[15] I. A. Modupe, T. J. Sefara, and O. Sunday, "Yorùbá gender recognition from speech using attention-based BiLSTM," in *Proceedings of The First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019 - Short Papers*. Trento, Italy: Association for Computational Linguistics, 2019, pp. 16–22.

[16] A. A. Abdulsatar, V. Davydov, V. Yushkova, A. Glinushkin, and V. Y. Rud, "Age and gender recognition from speech signals," in *Journal of Physics: Conference Series*, vol. 1410, no. 1. IOP Publishing, 2019, p. 012073.

[17] M. A. Yusnita, A. M. Hafiz, M. N. Fadzilah, A. Z. Zulhanip, and M. Idris, "Automatic gender recognition using linear prediction coefficients and artificial neural network on speech signal," in *2017 7th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, 2017, pp. 372–377.

[18] B. Zhong, Y. Liang, J. Wu, B. Quan, C. Li, W. Wang, J. Zhang, and Z. Li, "Gender recognition of speech based on decision tree model," in *Proceedings of the 3rd International Conference on Computer Engineering, Information Science & Application Technology (ICCIA 2019)*. Atlantis Press, 2019, pp. 571–577.

[19] N. A. Nazifa, C. Y. Fook, L. C. Chin, V. Vijean, and E. S. Kheng, "Gender prediction by speech analysis," *Journal of Physics: Conference Series*, vol. 1372, p. 012011, nov 2019.

[20] S. Bhukya, "Effect of gender on improving speech recognition system," *International Journal of Computer Applications*, vol. 179, pp. 22–30, 2018.

[21] L. M. Liztio, C. A. Sari, E. H. Rachmawanto *et al.*, "Gender identification based on speech recognition using backpropagation neural network," in *2020 International Seminar on Application for Technology of Information and Communication (iSemantic)*. IEEE, 2020, pp. 88–92.