The Effects of Acoustic Features of Speech for Automatic Speaker Recognition

Tumisho Billson Mokgonyane[1]
Department of Computer Science
University of Limpopo
Polokwane, South Africa
mokgonyanetb@gmail.com

Tshephisho Joseph Sefara[2]
Next Generation Enterprises and Institutions
Council for Scientific and Industrial Research
Pretoria, South Africa
tsefara@csir.co.za

Madimetja Jonas Manamela[3]
Department of Computer Science
University of Limpopo
Polokwane, South Africa
jonas.manamela@ul.ac.za

Thipe Isaiah Modipa[4]
Department of Computer Science
University of Limpopo
Polokwane, South Africa
thipe.modipa@ul.ac.za

Moses Sebaka Masekwameng[5]
Department of Computer Science
University of Limpopo
Polokwane, South Africa
sebaka.moshe@gmail.com

***Abstract*—Automatic speaker recognition is the task of automatically determining or verifying the identity of a speaker from a recording of his or her speech sample and has been studied for many decades. One of the most important steps of speaker recognition that significantly influences the speaker recognition performance is known as feature extraction. Acoustic features of speech have been researched by many researchers around the world, however, there is limited research conducted on African indigenous languages, South African official languages in particular. This paper presents the effects of acoustic features of speech towards the performance of speaker recognition systems focusing on South African low-resourced languages. This study investigates the acoustic features of speech using the National Centre for Human Language Technology (NCHLT) Sepedi speech data. Acoustic features of speech such as *Time-domain, Frequency-domain* and *Cepstral-domain* features are evaluated on four machine learning algorithms: K-Nearest Neighbours (K-NN), two kernel-based Support Vector Machines (SVM), and Multilayer Perceptrons (MLP). The results show that the performance is poor for time-domain features and good for spectral-domain features and even better for cepstral-domain features. However, the combination of these three features resulted in a higher accuracy and F_1 score of 98%.**

Keywords—speaker recognition, acoustic features of speech, time-domain, frequency-domain, cepstral-domain

I. INTRODUCTION

Automatic speaker recognition is the task of automatically determining or verifying the identity of a speaker from a recording of his or her speech sample. Studies have shown overtime that no two or more humans sound precisely identical [1]. The acoustic aspects of what differentiates among the different human voices are obscure and difficult to separate from signal aspects that reflect segment recognition. The authors in [2] report that the three sources of disparity between speakers include (1) the difference in vocal cords and vocal tract shapes, (2) the difference in accents (speaking styles), and (3) the difference in how speakers express themselves to convey a particular message. However, since it is difficult to quantify or control a speaker's tendency to use certain words (the third source), automatic speaker recognition systems exploit only the first two sources of disparity by examining low-level acoustic features of speech. The human voice consists of distinct acoustic features of speech that have the potential to uniquely distinguish between humans. The type acoustic features of speech extracted from the human voice greatly influence the performance of a speaker recognition system. Researchers around the world have researched the effects of several acoustic features of speech [3]–[12]. However, such research has focused mainly on well-resourced languages such as French, English, Chinese, Turkish and Vietnamese [13]–[16]. It has been reported that different acoustic features of speech have certain impact on a particular language [17], [18]. The effects of acoustic features of speech for speaker recognition has not been adequately investigated and explored on African indigenous languages, particularly South African low-resourced languages. Therefore, it is not known whether the acoustic features of speech do have an effect on South African under-resourced languages.

In this paper, we investigate the effects of acoustic features of speech towards the performance of speaker recognition systems focusing on South African low-resourced languages. The features are extracted from audio files recorded from the Sepedi native speakers. We chose Sepedi as it is one of the South Africa official language and reported to be low-resourced [19], [20]. This paper is structured as follows: Section II describes the speech features extracted from the human voice. Section III discusses the methodology and results are discussed in Section IV where: *(i)* we show if model performance is affected by different speech features, *(ii)* we show which speech features are compatible with each other, and *(iii)* we compare machine learning model's performance. Section V concludes the paper and highlights the future work.

II. ACOUSTIC FEATURES OF SPEECH

This section discusses the types of acoustic features of speech used for speaker recognition. There are different types of acoustic features of speech that can be extracted from speech and depending on the choice, the recognition accuracy varies. Some of the acoustic features of speech

available are Time-domain, Frequency-domain and Cepstral-domain features.

A. *Time-domain features*

Time-domain features are features that are directly extracted from the raw audio samples and include Zero Crossing Rate (ZCR), Energy and Entropy of Energy. Estimates of the spectral properties are obtained using a representation based on the short-time average ZCR. An appropriate definition of computations is given in [21]:

$$Z_n = \sum_{m=-\infty}^{\infty} | \operatorname{sgn}[x(m)] - \operatorname{sgn}[x(m-1)] | \, w(n-m) \quad (1)$$

where

$$\operatorname{sgn}[x(n)] = 1 \quad x(n) \geq 0$$
$$= -1 \quad x(n) < 0$$

and

$$w(n) = 1/(2N) \quad 0 \leq n \leq N-1$$
$$= 0 \qquad otherwise$$

An example of Time-domain features (ZCR) extracted from one audio sample is shown in Fig. 1.

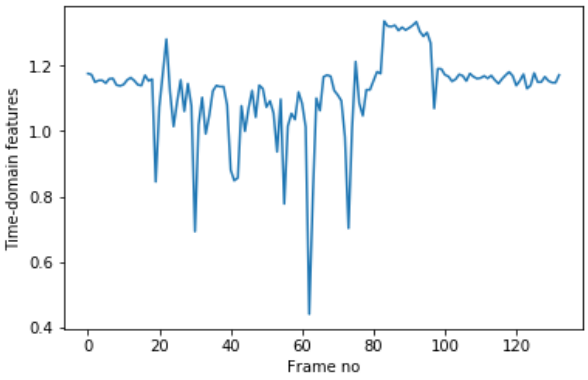

Fig. 1. Time-domain features example (ZCR).

B. *Frequency-domain features*

Frequency-domain features include Spectral Spread, Spectral Centroid, Spectral Flux, Spectral Entropy, Spectral Rolloff, Chroma Deviation and Chroma Vector, these features are based on the magnitude of the Discrete Fourier Transform (DFT) [22]. An example of Frequency-domain features (Spectral Centroid) extracted from one audio sample is shown in Fig. 2.

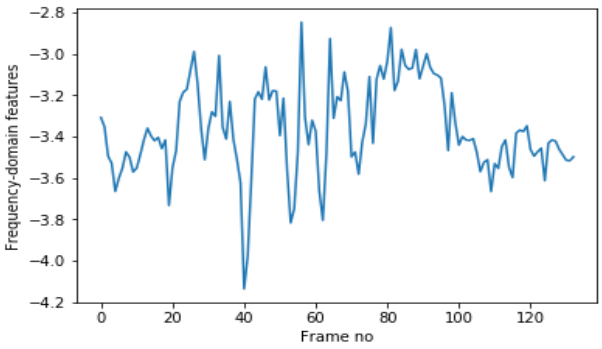

Fig. 1. Frequency-domain features example (Spectral Centroid).

C. *Cepstral-domain features*

Cepstral-domain features include Mel-Frequency Cepstral Coefficients (MFCCs) that result after the inverse DFT is applied on the logarithmic spectrum. MFCCs are determined with the help of a psychoacoustically motivated filter bank, followed by logarithmic compression and discrete cosine transform. Suppose the output of an X-channel filterbank is *Y (x); x = 1, … ,X;* the MFCCs are obtained using the following equation [2]:

$$c_n = \sum_{x=1}^{N} [\log Y(x)] \cos\left[\frac{\pi n}{X}\left(x - \frac{1}{2}\right)\right] \quad (2)$$

where n is the index of a cepstral coefficient. An example of Cepstral-domain features (MFCCs) extracted from one audio sample is shown in Fig. 3.

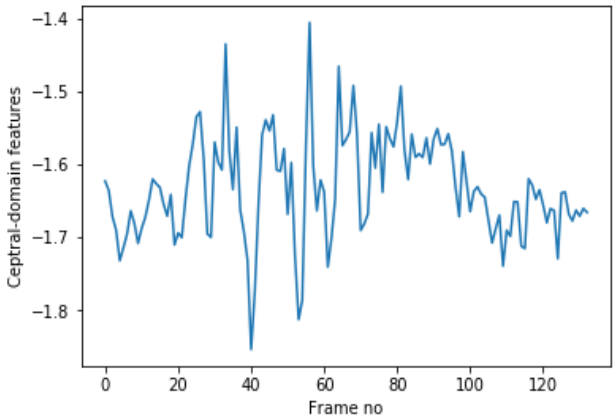

Fig. 2. Cepstral-domain features example (MFCC).

III. Methodology

This section discusses information regarding the data set, feature extraction and normalisation, classification model setup and evaluation metrics.

A. *Data*

The National Centre for Human Language Technology (NCHLT) [23] is the primary source of data used in this paper. The Sepedi NCLHT speech data, containing audio files recorded from different Sepedi native speakers is used. The data contains fifty (50) randomly selected speakers and 150 samples per speaker. The data as summarised in Table I, is partitioned into train and test partitions of 75% train data and 25% test data set.

TABLE I. Data Statistics

Unit	***Value***
No. of speakers	50
Instances per speakers	150
Total Duration (seconds)	24,681
Size (MB)	835.7

B. *Feature Extraction*

Features extraction is performed with the use pyAudioAnalysis [22] library. This library extracts a set of 34 acoustic features of speech described in Table II. The

extracted acoustic features of speech consists of three groups (Time-domain, Frequency-domain and Cepstral-domain features) discussed in Section II. These features are trained individually to compare their performances, and then combined to test their compatibility with each other. We investigate four different combinations of the features which are as follows:

- Time and Frequency (TF)
- Time and Cepstral (TC)
- Frequency and Cepstral (FC)
- Time and Frequency and Cepstral (TFC)

TABLE II. ACOUSTIC FEATURES ON SHORT-TERM WINDOWS

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.

C. Feature Normalisation

Standardisation of a data set is a common requirement for many machine learning algorithms. To achieve better results, we standardise the extracted acoustic features by removing the mean and scaling to unit variance using the following z-score normalization equation:

$$\hat{y} = \frac{x - \bar{x}}{\sigma} \tag{3}$$

where σ represents the variance and $\bar{x}$ represent the mean for each feature x.

D. Classification Model Setup and Parameter Optimization

The classification models are trained on Scikit-Learn [24] and GridSearchCV is used to find the best hyperparameters. Four machine learning algorithms are considered: K-Nearest Neighbours (K-NN), Radial Basis Function Support Vector Machines (RBF-SVM), Linear Support Vector Machines (LSVM) and Multilayer Perceptrons (MLP). The authors in [25], [26] give a detailed description of these algorithms and discuss the algorithm's application to speaker recognition.

1) ***K-Nearest Neighbours (K-NN)****:* The K-NN algorithm is a type of lazy learning or instance-based learning algorithm which only approximates the function locally and defer all computation until classification [27]. We train the K-NN classifier with 7 nearest neighbors which are weighted by the distance metric. The distance metric weights data points by the inverse of their distance, meaning closer neighbors of a data point have more value compared to neighbors which are further away.

2) ***Support Vector Machines*** *(SVM):* Support vector machine (SVM) are the advanced models with integrated learning algorithms in which classification and regression analysis is done by analyzing data and recognizing the patterns [28]. We implement the Radial Basis Function (RBF) and the Linear SVM kernels defined by the following equations:

$$Linear = \langle x, x' \rangle \tag{4}$$

$$RBF = exp(-\gamma \|x - x'\|^2) \tag{5}$$

where γ is a positive parameter.

3) ***Multilayer perceptron (MLP)****:* An MLP is a feed-forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. We train the MLP classifier with only one hidden layer and 100 neurons, the alpha parameter is set to 0.1, and 1000 epochs are performed. The alpha parameter is a parameter for regularization term (also known as the penalty term), that combats overfitting by constraining the size of the weights.

E. Evaluation

This study uses evaluation measurements such as accuracy, precision, recall, F_1 score and root mean squared error (RMSE) to evaluate the performance of each of the classification models.

IV. RESULTS AND DISCUSSION

This section discusses the results of the methodology discussed in Section III. Table III shows the accuracy scores for three different features of speech and their combinations. The features are trained with different classification models. It is observed that time-domain features on their own give the lowest accuracy for all classification models, with 25.81%, 32.69%, 32.75% and 34.40% for K-NN, LSVM, RBF-SVM, and MLP respectively. When Training the classification models with Frequency-domain features, we observe that the performance improves by 38.14%, 49.60%, 48.05% and 49.49% for K-NN, LSVM, RBF-SVM, and MLP classifiers respectively. Cepstral-domain features improves the accuracy even further to 84.32% for K-NN, 91.25% and 91.68% for LSVM and RBF-SVM and 91.89% for MLP. LSVM performs better than RBF-SVM for Frequency-domain features with a difference of 1.49%, however, the performs difference for Cepstral-domain features is only 0.43% in favour of RBF-SVM. MLP classifier outperforms all the classification models regardless of the features in use. From these results, we see that the performance of the model is indeed affected by different acoustic features of speech.

TABLE III. ACCURACY SCORES FOR DIFFERENT MODELS

Features	*Classifier Models*			
	K-NN	***LSVM***	***RBF-SVM***	***MLP***
Time	0.2581	0.3269	0.3275	**0.3440**
Frequency	0.6395	0.8229	0.8080	**0.8389**
Cepstral	0.8432	0.9125	0.9168	**0.9189**
TF	0.6715	0.8416	0.8405	**0.8576**
TC	0.8485	0.9413	0.9387	**0.9547**
FC	0.8661	0.9627	0.9621	**0.9712**
TFC	**0.8763**	**0.9712**	**0.9680**	**0.9755**

To investigate which combination of features improves the models performance, we combined Time-domain and Frequency-domain features and the accuracy score has increased by 3.25% for RBF SVM, higher than K-NN's 3.20% and 1.87% for both LSVM and MLP, resulting in a 2.53% increase on average. A combination of Time-domain and Cepstral-domain features improves the performance by an average of 2.28% and a combination of Frequency-domain and Cepstral-domain features results in a higher improvement of 4.28% on average. By this higher average, we therefore conclude that the combination of Frequency-domain and Cepstral-domain features is more compatible as compared to the combination of Time and Frequency-domain features and the combination of Time and Cepstral-domain features.

The performance of the classification models increases even further when Time, Frequency and Cepstral-domain features are combined. It is observed that MLP classifier performed best with an accuracy of 97.55% for all features combined, K-NN gives a lower accuracy of 87.63% and LSVM and RBF-SVM have an accuracy of 97.12% and 96.80% respectively. MLP outperforms LSVM by a difference of only 0.43%, similar to the difference obtained when training RBF-SVM and LSVM with Cepstral-domain features.

TABLE IV. RESULTS BASED ON BEST PERFORMING FEATURES

Classifier Models	*Performance Metrics*			
	Accuracy	***Recall***	***Precision***	***F_1 Score***
K-NN	0.8763	0.8853	0.8763	0.8729
LSVM	0.9712	0.9722	0.9712	0.9711
RBF-SVM	0.9680	0.9691	0.9680	0.9679
MLP	**0.9755**	**0.9764**	**0.9755**	**0.9755**

Table IV shows the results of the classification models trained with a combination of all features (TCF) and the accuracies are similar to that of Table III. We report on precision, recall and F_1 score. Table IV also shows that both precision and recall are similar to the classification accuracy results for their respective classification models. This results shows that the classification models did not overfit. To validate this, we calculated the F_1 score, which conveys the balance between precision and recall. A similar result for F_1 score was achieved and therefore validates that the classification models did not overfit.

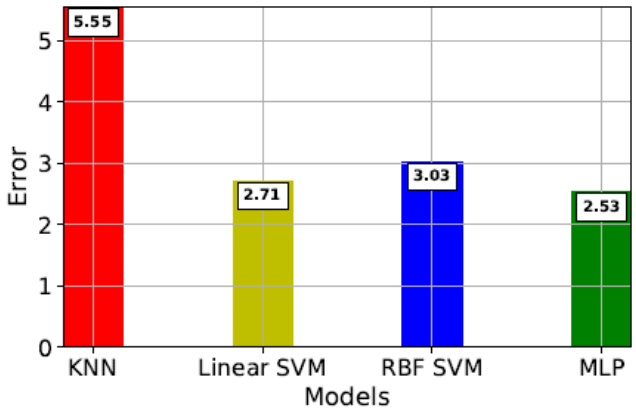

Fig. 3. The Root Mean Square Error of the classifier models.

Fig. 4 shows the RMSE which depicts that K-NN has the highest RMSE of 5.5506 followed by RBF-SVM at 3.0315. LSVM and MLP had the lowest RMSE at 2.7125 and 2.5374 respectively. This results show that K-NN misclassify most of the data whereas LSVM and MLP perform well with a difference of only less than one percent.

V. CONCLUSION AND FUTURE WORK

This paper reported on the effects of acoustic features of speech on the performance of speaker recognition systems focusing on the Sepedi language as one of the 11 South African official languages. The paper gave a general overview of speaker recognition and briefly described three acoustic features of speech (Time-domain, Frequency-domain, and Cepstral-domain features). The training and testing stages including feature extraction and normalisation, classification model setup and evaluation are clearly explained in the methodology section. The dataset of Sepedi speech data was obtained from the NCHLT project. pyAudioAnalysis tool was used for feature extraction and the extracted acoustic features of speech were then u sed to train the classification models on Scikit-Learn.

The effects of Time-domain, Frequency-domain, and Cepstral-domain features was investigated individually on different classification models (K-NN, LSVM, RBF-SVM and MLP) and it was observed that Time-domain features have the lowest accuracy score for all classification models. The accuracy of the classification models improved when trained with Frequency-domain features and improved even further when the classification models are trained with Cepstral-domain features. From this improvement, we conclude that the Cepstral-domain features are more superior and have a more positive effect towards the performance of speaker recognition systems based on African indigenous languages. It was also observed that the performance of the classification models increases even more when Time, Frequency and Cepstral-domain features are combined together. The more compatible combination was that of Frequency and Cepstral-domain features where the classification accuracy is only 0.43% less than that of a combination all the features combined together. We trained the classification models on all features and it was observed that MLP outperformed all the classification models with an accuracy of 97.55% and K-NN had the lowest accuracy of

87.63%. LSVM and RBF-SVM had an accuracy score of 97.12% and 96.80% respectively. From this classification accuracy scores, the future work on this study will focus on a combination of Time, Frequency and Cepstral-domain features and these will be trained with LSVM and MLP classification models as they have shown to give best performances. As an extension to the study, a user friendly environment (graphical user interface) will be developed for easy access to the speaker recognition system.

ACKNOWLEDGMENT

This study was conducted and facilitated at the University of Limpopo, Department of Computer Science, Telkom Centre of Excellence for Speech Technology.

REFERENCES

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," Speech communication, vol. 52, no. 1, pp. 12–40, 2010.

[2] R. P. Ramachandran, K. R. Farrell, R. Ramachandran, and R. J. Mammone, "Speaker recognition—general classifier approaches and data fusion methods," Pattern Recognition, vol. 35, no. 12, pp. 2801–2821, 2002.

[3] A. Eronen, "Comparison of features for musical instrument recognition," in Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics, 2001, pp. 19–22.

[4] T. Eriksson, S. Kim, H.-G. Kang, and C. Lee, "An information-theoretic perspective on feature selection in speaker recognition," IEEE Signal Processing Letters, vol. 12, no. 7, pp. 500–503, 2005.

[5] K. Umapathy, B. Ghoraani, and S. Krishnan, "Audio signal processing using time-frequency approaches: Coding, classification, fingerprinting, and watermarking," Hindawi Publishing Corporation EURASIP Journal on Advances in Signal Processing, vol. 28, 01 2010.

[6] A. Lawson, P. Vabishchevich, M. Huggins, P. Ardis, B. Battles, and A. Stauffer, "Survey and evaluation of acoustic features for speaker recognition," in 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2011, pp. 5444–5447.

[7] H. Baharipour, M. E. Ahmadabadi, and M. Mosleh, "A study of speaker recognition approaches based on feature selection and classification methods."

[8] S. V. Chougule and M. Chavan, "Robust spectral features for automatic speaker recognition in mismatch condition," Procedia Computer Science, vol. 58, pp. 272–279, 2015.

[9] S. S. Tirumala, S. R. Shahamiri, A. S. Garhwal, and R. Wang, "Speaker identification features extraction methods: A systematic review," Expert Systems with Applications, vol. 90, pp. 250–271, 2017.

[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) 2018, 2018, pp. 128–133.

[11] R. Font, J. M. Esp´ın, and M. J. Cano, "Experimental analysis of features for replay attack detection-results on the asvspoof 2017 challenge." in Interspeech, 2017, pp. 7–11.

[12] F. E. Abualadas, A. M. Zeki, M. S. Al-Ani, and A.-E. Messikh, "Speaker identification based on hybrid feature extraction techniques," International Journal of Advanced Computer Science and Applications, vol. 10, no. 3, 2019.

[13] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and rsr2015," Speech Communication, vol. 60, pp. 56–77, 2014.

[14] H. Celıktas and C. Hanılci "A study on turkish text — dependent speaker recognition," 2017 25th Signal Processing and Communications Applications Conference (SIU), vol. 81, pp. 1–4, 2017.

[15] W. Yanlei, Z. Heming, G. Xiaojiang, and G. Chenghui, "A study on speaker and session variability in speaker recognition of chinese whispered speech," in 2010 The 2nd International Conference on Industrial Mechatronics and Automation, vol. 2, 2010, pp. 292–295.

[16] D. T. Thu, L. T. Van, Q. N. Hong, and H. P. Ngoc, "Text-dependent speaker recognition for vietnamese," in 2013 International Conference on Soft Computing and Pattern Recognition (SoCPaR), 2013, pp. 196–200.

[17] A. Bora, J. Vajpai, and S. Gaur, "Speaker identification for biometric access control using hybrid features," International Journal of Computational Science and Engineering, vol. Vol. 9, 04 2018.

[18] M. Sugiyama, "Automatic language recognition using acoustic features," in [Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing, 1991, pp. 813–816 vol.2.

[19] F. de Wet, J. Badenhorst, and T. Modipa, "Developing speech resources from parliamentary data for south african english," Procedia Computer Science, vol. 81, pp. 45–52, 2016.

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

[21] R. Bachu, S. Kopparthi, B. Adapa, and B. D. Barkana, "Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy," in Advanced Techniques in Computing Sciences and Software Engineering. Springer, 2010, pp. 279–282.

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

[23] E. Barnard, M. H. Davel, C. v. Heerden, F. d. Wet, and J. Badenhorst, "The nchlt speech corpus of the south african languages," in Spoken Language Technologies for Under-Resourced Languages, 2014.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg et al., "Scikit-learn: Machine learning in python," Journal of machine learning research, vol. 12, no. Oct, pp. 2825–2830, 2011.

[25] 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, pp. 141–146.

[26] T. B. Mokgonyane, T. J. Sefara, M. J. Manamela, and T. I. Modipa, "The effects of data size on text-independent automatic speaker identification system," in 2019 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD). IEEE, 2019, pp. 1–6.

[27] D. Aha and D. Kibler, "Instance-based learning algorithms," Machine Learning, vol. 6, pp. 37–66, 1991.

[28] J. K. Sahoo and D. Rishi, "Speaker recognition using support vector machines," International Journal of Electrical, Electronics and Data Communication, vol. 2, no. 2, pp. 01–04, 2014.