The Effects of Data Size on Text-Independent Automatic Speaker Identification System

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

Tshephisho Joseph Sefara
Modelling and Digital Science
Council for Scientific and Industrial Research
Pretoria, South Africa
tsefara@csir.co.za

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

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

***Abstract*— Speaker recognition is a technique that automatically identifies a speaker from a recording of their speech utterance. Speaker recognition technologies are taking a new direction due to rapid progress in artificial intelligence. Research in the field of speaker recognition has shown fruitful results. There is, however, not much work done for African indigenous languages that have limited speech data resources. This paper presents how data size impacts the accuracy of an automatic speaker recognition system models, focusing on the Sepedi language as it is one of the South African under-resourced language. The speech data used is acquired from the South African Centre for Digital Language Resources. Four machine learning models, namely, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Multilayer Perceptrons (MLP) and Logistic Regression (LR) are trained under four data setting environment. LR performed better than other models with the highest accuracy of 91% while SVM obtained the highest increase of 4% in accuracy as data size increases.**

Keywords*— *Speaker recognition, text-independent, support vector machine, k-nearest neighbors, multilayer-perceptron, logistic regression

I. INTRODUCTION

The human voice is a phenomenon that is highly dependent on the speaker who produced it. Studies have shown over time that no two individuals sound exactly the same [1]. Many physical aspects of a speech signal such as the timber, tone or intensity vary a lot from one speaker to another. The range of vocabulary or expressions a speaker normally uses also varies amongst speakers. All these properties make the human voice a very powerful tool that can be used in computer-based security systems since its physical characteristics are easy to measure and compare [2], [3]. In addition, a speech signal is quite well-known and easy to use and it has been deeply studied for many years, therefore many powerful algorithms can be found to deal with it [4]–[6].

Speaker recognition is one of the important research topics in digital signal processing and has a variety of applications, especially in security systems [7]. Voice controlled systems and devices rely heavily on speaker recognition. Automatic speaker recognition is a technique used to automatically recognise the identity of a speaker from a recording of their voice. Research in the field of speaker recognition has now spanned over 50 years [6], with a considerable amount of publications available [8]–[13]. However, research in this field has been conducted mainly on readily available speaker databases such as YOHO, TIMIT/NTIMIT and ANDOSL speaker databases [14] which are built on resourced languages such as English, Chinese, Vietnamese, Turkish [15]–[17].

South African official languages are still classified as being highly under-resourced [18]. Very few speaker recognition research attempts have been made in the context of these languages. The Sepedi language is one of the South African official languages that is largely spoken in the Limpopo Province of South Africa. Census 2011 reports that Sepedi is spoken by most persons in Limpopo province with more than 2.8 million speakers [19]. This paper presents the development of an automatic speaker recognition system that incorporates the identification or classification of Sepedi speakers focusing on how data size improves the recognition accuracy. The system uses machine learning algorithms that learn features extracted from the Sepedi speech data to train the classifier models. The system can be used to automatically authenticate speaker identities using their voices to allow only the correctly identified persons an access right to information systems or to facilities that need to be protected from the intrusion of unauthorized persons.

The rest of this paper is outlined as follows: Section II gives an overview of speaker recognition. Section III details methods used to build the learning models. Section IV discusses the experimental results, and the paper is concluded in Section V.

II. OVEVIEW OF SPEAKER RECOGNITION

A speaker recognition system is composed of two different phases, a training phase and a testing phase. In the training phase (also referred to as the enrolment phase), a speaker's voice is recorded and a number of features are extracted to form a unique voice print (speaker model) that uniquely identifies the speakers. In the testing phase (the recognition phase), the speech sample provided is compared against the previously created voice print. Some applications of speaker recognition include customer verification for bank transactions, control on the use of credit cards, security control for confidential information, surveillance procedures, forensics and remote access to computers [20].

A speaker recognition system consists of two fundamental tasks, *speaker identification* and *speaker verification*. Speaker identification is the task of associating an unknown voice with one from a set of enrolled speakers. Potential speaker identification applications include automatic speaker labelling of recorded meetings for speaker-dependent audio indexing and intelligent answering machines with personalized caller greetings [21]. Speaker verification is the task of determining whether an unknown voice is from a particular enrolled speaker. Applications of speaker verification include telephone banking, computer login, cellular telephone fraud prevention and calling cards [21].

Speaker recognition systems can be further classified by the constraints placed on the text of the speech used in the system, the classification can either be text-dependent or text-independent. In the text-dependent case, the input sentence or phrase is fixed for each speaker, and in the text-independent case, there is no restriction on the sentence or phrase to be spoken [22]. Text-independent recognition is suited for application areas such as forensics and surveillance where speakers can be considered non-cooperative users, as they do not specifically wish to be recognised, whereas, the text-dependent case is suited for services such as telephone-based services and access control, where the users are considered cooperative [23]. Text-dependent recognition achieves higher recognition performance than the text-independent recognition [20]. However, due to the flexibility that the text-independent recognition provides, the increasing development trend is in the building of the text-independent recognition systems [24]. A text-independent speaker identification system is realised in this work.

III. Methodology

This section discusses the data, tools, methods, and evaluation procedure followed in this study. The overview of the proposed system is illustrated in Figure 1. Given an input speech signal, the audio feature vectors are extracted and used to train machine learning models like Logistic Regression (LR), Multilayer Perceptron (MLP), Support Vector Machines (SVM) and K-Nearest Neighbour (KNN). The best performing model is used for development.

A. Data

The dataset is obtained from the National Centre for Human Language Technology (NCHLT) project of the South African Centre for Digital Language Resources [25]. The data contain Sepedi speech audio files recorded by different speakers. We randomly sampled the data to 50 speakers, each containing 150 audio samples. The total audio files are 7500 which makes up to a total duration of 24,681 seconds. The data is partitioned into 10% (15 samples) for testing and 90% for training. We use the 15 samples per speaker to test the models, which is a total of 750 samples for all the speakers combined. To test the effectiveness of data size towards a speaker identification system, the train data is divided into different sample sizes shown in Table I.

TABLE I. Summary of the Data

Experiment	Data Size	Speakers
A	25 Samples	50
B	50 Samples	50
C	100 Samples	50
D	135 Samples	50

B. Feature Extraction

A human voice is comprised of different discriminative features that have the potential to uniquely identify human beings. Feature extraction is one of the most significant aspect of speaker recognition, this step generates feature vectors that represent each speech signal. We extract speech features using an open-source comprehensive package developed in Python called pyAudioAnalysis [26]. pyAudioAnalysis implements a total of 34 short-term features and these features are illustrated in Table II. The features are grouped into three domains, namely, Time, Frequency, and Cepstral-domain features.

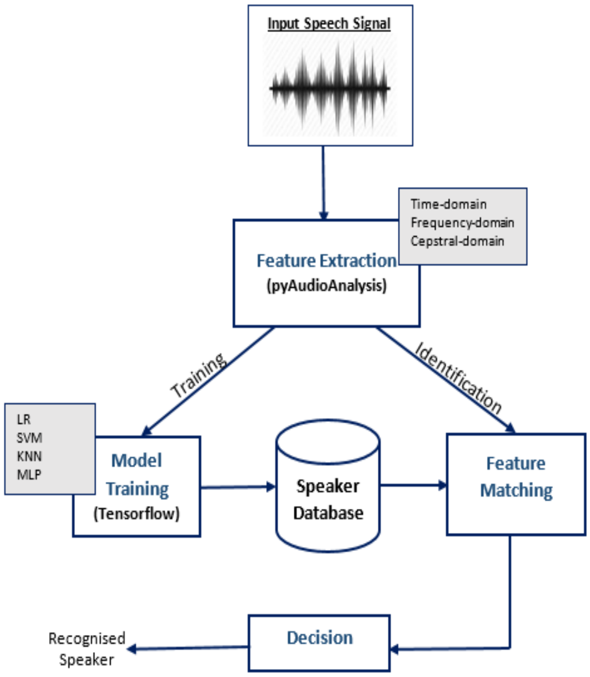

Fig. 1. Systematic overview of the proposed speaker recognition system.

TABLE II. Acoustic Features on Short Term Window

Feature ID	Feature Name	Description
1	ZCR	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	Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but 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.

*1) **Time-domain features:*** include Zero Crossing Rate (ZCR), Energy and Entropy of Energy that are extracted directly from the raw audio samples. Speech signals are broadband signals and the interpretation of average ZCR is therefore much less precise. However, rough 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 [27]. The ZCR measures the number of times in a given time interval/frame that the amplitude of the speech signals passes through a value of zero. Several features of a speech signal can be selected using energy and zero crossing. ZCR is very useful for discriminating speech from noise and for determining the start and the end of the speech segment.

*2) **Frequency-domain features**:* are based on the magnitude of the Discrete Fourier Transform (DFT), these include Spectral Spread, Spectral Centroid, Spectral Flux, Spectral Entropy, Spectral Rolloff, Chroma Deviation and Chroma Vector.

*3) **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 recognition tasks and widely used for speaker and speech recognition [28]. In the source-filter model of speech, MFCCs are understood to represent the filter (vocal tract). The 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 M-channel filterbank is $Y(m), m = 1, \dots, M$, the MFCCs are obtained using the following equation [20]:

$$c_n = \sum_{m=1}^{M} [Log Y(m)] \cos\left[\frac{\pi n}{M}\left(m - \frac{1}{2}\right)\right] \qquad (2)$$

where n is the index of a cepstral coefficient. Figure 2 shows the time domain (ZCR), frequency domain (Spectral Centroid) and cepstral domain (MFCCs) features extracted from a single speech utterance given by a single speaker.

C. Models

This section details the setup and architecture of the learning algorithms. The following machine learning algorithms are used to train the classifier models using Tensorflow [29]. To get the best results, we use 1000 iterations or epochs during training.

*1) **Logistic Regression*** is a highly accurate and robust method that uses *multinomial logistic regression* method to generalize logistic regression to multiclass problems [30]. In most cases, LR does not suffer from the overfitting problem because it has few parameters, and have a bias parameter to manage the overfitting.

*2) **Multilayer Perceptron*** has more than two hidden layers that typically uses random initialization and stochastic gradient descent to initialize and optimize the weights [31]. MLPs can handle extremely complex tasks, although they take a lot of time to train and are computationally expensive [32]. We train the MLP classifier architecture shown in Table III using Tensorflow. Dropout layers use a probability of 0.5. Tanh represents a *rectified linear unit*. The number of filters for the last layer corresponds to the number of classes of the given dataset.

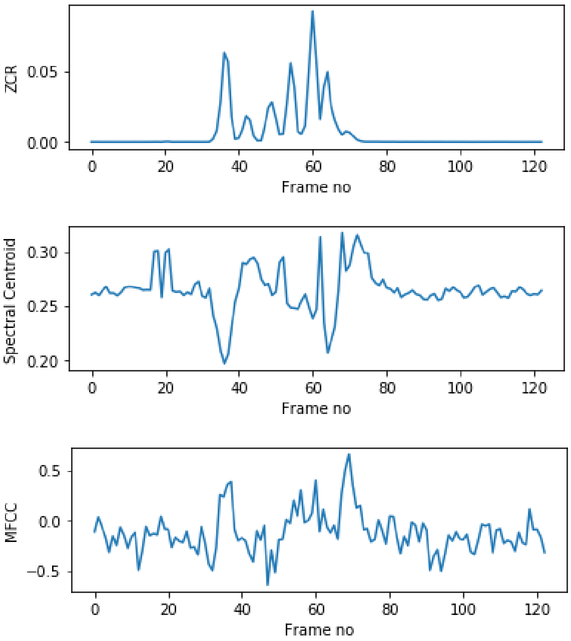

Fig. 2. Time, Frequency and Cepstral features extracted from a single audio file.

TABLE III. TABLE OF THE MLP ARCHITECTURE

Layer	Type	Filters/Neurons
1	Fully connected+tanh	68
2	Dropout	-
3	Fully connected+tanh	68
4	Dropout	-
5	Fully connected+softmax	50

*3) **Support Vector Machines*** are supervised learning models with associated learning algorithms that analyse data and recognize patterns, used for classification and regression analysis [32]. SVM is a popular discriminative classifier which models the boundary between a speaker and a set of impostors and has proven to be a powerful technique for pattern classification. Currently, the SVM classifier is one of

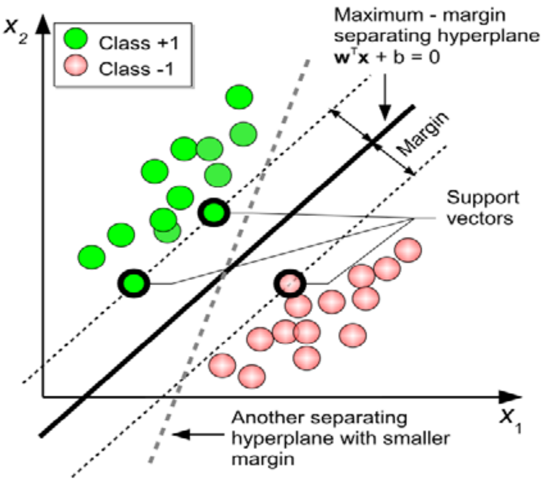
x_2
Class +1
Class -1
Maximum - margin separating hyperplane
$w^T x + b = 0$
Margin
Support vectors
Another separating hyperplane with smaller margin
x_1

Fig. 3. A maximum-margin hyperplane that separates the positive (+1) and negative (-1) training examples is found by an optimization process.

the most robust classifiers in speaker verification. One good reason for the SVM to be popular is its good generalization performance to classify unseen or new data [33]. As shown in Figure 3, the SVM is a classifier which models the decision boundary between two or more classes as a separating hyperplane. We implement radial basis function (RBF) and linear SVM kernels defined by the following equations:

- Linear SVM = $\langle x, x' \rangle$ (3)
- RBF SVM = $exp(-\gamma\|x - x'\|^2)$ (4)

where γ is a positive integer.

*4) **K-Nearest Neighbor*** belongs to a family of instance-based and lazy learning algorithms [34]. Whenever there is a need to classify unknown data sample from a set of testing data, the KNN's task is to search through the training dataset for the k most similar samples [35].

D. Evaluation

The behaviour of each model is evaluated based on certain criteria to assess its performance. The performance is affected by the size of the training data, the quality of the speech signal, and most importantly the type of learning algorithm employed. The following evaluation measurements are used to evaluate the performance of the models:

*1) **Accuracy*** is the percentage of the examples which are correctly classified from all the examples given. We report on validation accuracy to avoid overfitting and testing accuracy to avoid underfitting.

*2) **Categorical cross entropy*** loss function since the data is categorical.

E. System Specification

We use Ubuntu server to conduct the experiments, the environment is configured with one NVIDIA Tesla K80 GPU, 11 Gig Memory and Intel CPU at 2.3GHz shown in Table 4.

TABLE IV. SYSTEM SPECIFICATIONS.

Hardware	Specification
System	Ubuntu x86_64
GPU Memory	11441MiB
CPU	4 x Intel(R) Xeon(R) CPU E5-2686 v4 @ 2.30GHz
GPU	NVIDIA Tesla K80 with CUDA v10

IV. EXPERIMENTS AND RESULTS

This section discusses the experimental results. All the experiments tested on the same test data set that is not used during model training.

A. Results on overfitting

To avoid overfitting, we illustrate the learning curves in Figure 4 for LR and Figure 5 for MLP. Both models are trained for 1000 epochs and the curves did not decrease. Hence, this shows models are not overfitted. Furthermore, Figure 6 shows the loss function curve of LR and MLP for all data sizes. These shows the curves did not increase but continued to decrease, hence these results validate overfitting did not occur.

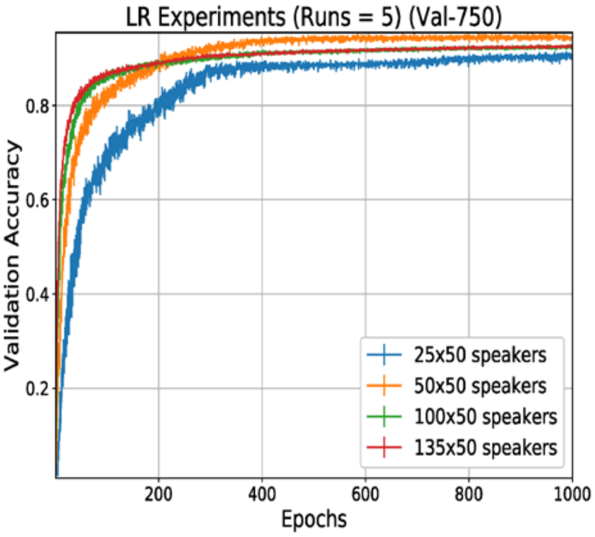

Fig. 4. Effect of data size on speaker recognition using LR.

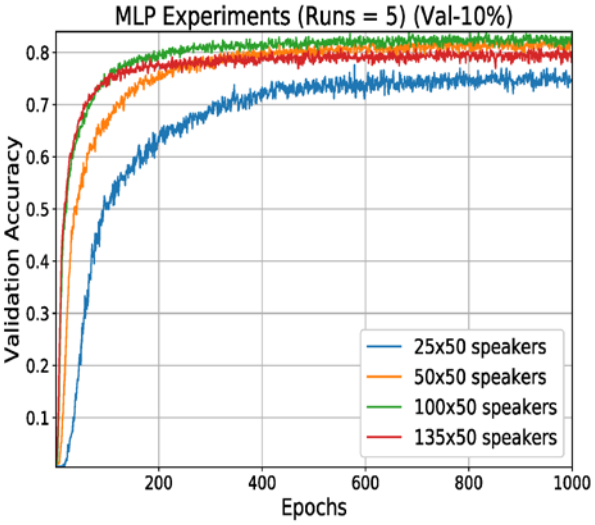

Fig. 5. Effect of data size on speaker recognition using MLP.

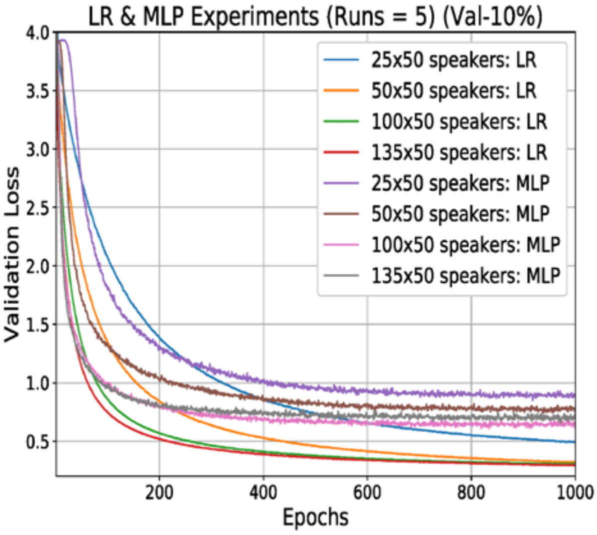

Fig. 6. Result of loss function learning curves of LR and MLP.

B. Results on testing accuracy

We observe lower testing accuracy results for 25x50 samples data set (Experiment A). This shows that all the learning models did not converge under limited data. Hence, larger training data is required to have robust models. As such, we increased the amount of data to 50x50 samples (Experiment B). We observe LR performing well compared to other models with a testing accuracy of 0.81. While Linear

SVM scored a lower accuracy of 0.36. The RBF SVM performed better than linear SVM, KNN and MLP. In Experiment C and D (100x50 and 135x50 samples), LR outperformed all of the models, this happens since LR does not contain complex parameters and the model learns easily by achieving higher accuracy of 0.83 and 0.91 respectively. The test results are shown in Table V. All the models increase accuracy when the data size increases. We observe linear SVM increasing by 4.1% when data is increased. This may infer that under big data settings this model may obtain much better results. From these results, we implement the speaker recognition system based on the LR model.

TABLE V. SUMMARY OF TEST RESULTS.

	Accuracy per data size				
Models	**25**	**50**	**100**	**135**	**Increase**
LR	**0.75**	**0.81**	**0.83**	**0.91**	1.21%
RBF SVM	0.69	0.74	0.79	0.82	1.18%
MLP	0.66	0.71	0.75	0.76	1.15%
KNN	0.66	0.63	0.69	0.70	1.06%
Linear SVM	0.15	0.36	0.56	0.62	**4.1%**

V. CONCLUSION

This paper reported how data size impact accuracy when training speaker recognition models, focusing on the Sepedi speech dataset. All the stages of training and testing, covering feature extraction, model training, and evaluation are described. The dataset of Sepedi speech data was obtained from the South African Centre for Digital Language Resources. Features were extracted using a Python library (pyAudioAnalysis) for audio analysis. We implemented well-known MLPs and machine-learning algorithms. We observed LR performing well under both limited and larger data settings and that all learning algorithms increased accuracy when data is increased. As an extension to the study, the future work will focus on:

- Testing the effects of different types of **features**.
- Increasing number of speakers to test the **complexity** of our approach.
- Adding noise to speech samples to test the **scalability** and **robustness** of the models.

ACKNOWLEDGMENT

This study was 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, pp. 12-40, 2010.

[2] E. D. Casserly and D. B. Pisoni, "Speech perception and production," *Wiley Interdisciplinary Reviews: Cognitive Science,* vol. 1, pp. 629-647, 2010.

[3] D. S. Rodríguez, "Text-Independent Speaker Identification," Kraków, Poland, 2008.

[4] K. Hashimoto, J. Yamagishi and I. Echizen, "Privacy-preserving sound to degrade automatic speaker verification performance," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on,* 2016.

[5] T. Marciniak, R. Weychan, A. Stankiewicz and A. Dabrowski, "Biometric speech signal processing in a system with digital signal processor," *Bulletin of the Polish Academy of Sciences Technical Sciences,* vol. 62, pp. 589-594, 2014.

[6] S. Furui, "50 years of progress in speech and speaker recognition research," *ECTI Transactions on Computer and Information Technology (ECTI-CIT),* vol. 1, pp. 64-74, 2005.

[7] N. Singh, R. A. Khan and R. Shree, "Applications of speaker recognition," *Procedia engineering,* vol. 38, pp. 3122-3126, 2012.

[8] H. Chenchen, G. Wei, F. Wenlong and F. Dongyu, "Research of speaker recognition based on the weighted fisher ratio of MFCC," in *Proceedings 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC)*, 2013.

[9] B. Ruiz, P. Domingo and L. Hernandez, "A dual speech/speaker recognition using GMM in speaker identification and a HMM in keyword speech recognition," in *Proceedings IEEE 33rd Annual 1999 International Carnahan Conference on Security Technology (Cat. No.99CH36303),* 1999.

[10] N. Wang and L. Wang, "Robust speaker recognition based on multi-stream features," in *2016 IEEE International Conference on Consumer Electronics-China (ICCE-China),* 2016.

[11] Z. N. Karam, W. M. Campbell and N. Dehak, "Graph relational features for speaker recognition and mining," in *2011 IEEE Statistical Signal Processing Workshop (SSP),* 2011.

[12] M. R. Leonard and J. H. L. Hansen, "In-set/out-of-set speaker recognition: leverging the speaker and noise balance," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008.

[13] L. Chen and Y. Yang, "Emotional speaker recognition based on i-vector through Atom Aligned Sparse Representation," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing,* 2013.

[14] B. R. Wildermoth and K. K. Paliwal, "GMM based speaker recognition on readily available databases," in *Microelectronic Engineering Research Conference, Brisbane, Australia*, 2003.

[15] H. Çelıktaş and C. Hanılçı, "A study on Turkish text — Dependent speaker recognition," *2017 25th Signal Processing and Communications Applications Conference (SIU),* vol. 81, pp. 1-4, 2017.

[16] 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*, 2010.

[17] D. 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.

[18] F. 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.

[19]

[20] 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, pp. 2801-2821, 2002.

[21] D. A. Reynolds, "Automatic speaker recognition using Gaussian mixture speaker models," in *The Lincoln Laboratory Journal*, 1995.

[22] Y. Liu, Y. Qian, N. Chen, T. Fu, Y. Zhang and K. Yu, "Deep feature for text-dependent speaker verification," *Speech Communication,* vol. 73, pp. 1-13, 2015.

[23] L. Gbadamosi, "Text independent biometric speaker recognition system," *International Journal of Research in Computer Science,* vol. 3, p. 9, 2013.

[24] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Advances in Signal Processing,* vol. 2004, p. 101962, 2004.

[25] N. J. De Vries, M. H. Davel, J. Badenhorst, W. D. Basson, F. Wet, E. Barnard and A. De Waal, "A smartphone-based ASR data collection tool for under-resourced languages," *Speech communication,* vol. 56, pp. 119-131, 2014.

[26] T. Giannakopoulos, "pyaudioanalysis: An open-source python library for audio signal analysis," *PloS one,* vol. 10, 2015.

[27] R. G. 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.

[28] V. Tiwari, "MFCC and its applications in speaker recognition," *International journal on emerging technologies,* vol. 1, pp. 19-22, 2010.

[29] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard and others, "Tensorflow: a system for large-scale machine learning.," in *OSDI*, 2016.

[30] F. E. Harrell, "Ordinal logistic regression," in *Regression modeling strategies,* Springer, 2015, pp. 311-325.

[31] F. Richardson, D. Reynolds and N. Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE Signal Processing Letters,* vol. 22, pp. 1671-1675, 2015.

[32] J. K. Sahoo and D. Rishi, "Speaker Recognition using Support Vector Machines," *International Journal of Electrical, Electronics and Data Communication,* vol. 2, pp. 1-4, 2014.

[33] 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 African Telecommunication and Networks and Application Conference (SATNAC) 2018*, Arabella, Western Cape, South Africa, 2018.

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

[35] 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 *International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD 2018)*, Durban, South Africa, 2018.