Support Vector Machine based automatic speaker recognition system

M.Subba Rao[#1], K.Umamaheswari[*2], P.Venkata jagadeesh[#3]
[1]*Professor in Department of Information Technology, AITS Rajampet, AP, India*
[23]*Students in Deparment of Computer Science and Engineering, AITS Rajampet, AP, India*
[1]msraoswap@gmail.com
[2]Umamaheswarianusha77@gmail.com
[3]pendlimarrijagadeesh@outlook.com

Abstract--Speaker recognition is a technique of identifying the person talking to a machine using the voice features and acoustics. It has multiple applications ranging in the fields of Human Computer Interaction (HCI), biometrics, security, and Internet of Things (IoT). At present, multiple biometric techniques co-exist with each other, for instance, iris, fingerprint, voice, facial, and more. Voice is one metric which apart from being natural to the users, provides comparable and sometimes even higher levels of security when compared to some traditional biometric approaches. This paper aims to evaluate different pre-processing, feature extraction, and machine learning techniques on audios recorded in unconstrained and natural environments to determine which combination of these works well for speaker recognition and classification. Thus, we present several methods of audio preprocessing like trimming, split and merge, noise reduction, and vocal enhancements to enhance the audios obtained from real-world situations. Mel Frequency Cepstral Coefficients (MFCC) are extracted for each audio, along with their differentials and accelerations to evaluate machine learning classification techniques such as k-Nearest Neighbor, Multilayer Perceptron, Support Vector Machines, and Random Forest Classifiers. In this paper, we used Support Vector Machine Algorithm to get best classification rate with its hyper-parameter

Index Terms—Speaker recognition, Audio pre-processing, Mel Frequency Cepstral Coefficients (MFCC), Support Vector Machine, Performance measures

I. INTRODUCTION

The human voice is a phenomenon which depends heavily on the speaker who produces it. Studies show that no two individuals sound the same [1], the acoustic aspects of what determines the discrepancies between the voices of the speakers are unclear and difficult to differentiate from signals that represent the identification of segments. The causes of variability between speakers are threefold, (1) the difference in speech styles (accent included), (2) the difference in vocal tract shapes and vocal cords, and (3) the way speakers articulate themselves to communicate a specific message (words or phrases used). Nonetheless, because a speaker's propensity to use certain phrases, sentences, and syntactic structures (referring to the third source) in an experiment is not easy to quantify or control, automatic speaker recognition systems use the first two sources only by exploring a speech signal's low-level acoustic characteristics.

Recognition of speakers is an important topic in signal processing and has a variety of applications, especially in security systems [3]. The voice controlled systems and apps rely heavily on the recognition of speakers. Many speaker recognition security control systems for confidential information, customer verification of bank transactions, forensics and remote computer access [2]. Researchers have published many publications in the field of speaker recognition [4]- [10] but very few (if not none) attempts to build speaker recognition systems developed using under-resourced languages have been made public. The official languages are still listed as strongly under-resourced according to [11]. Census 2011[12] estimates that with more than 2.8 million speakers. This paper presents the development of an automated voice recognition system that combines language speaker classification and recognition. To train the classifier model, the system uses machine learning algorithms which learn features extracted from the speech data. The program can be used to authenticate speaker identities automatically using their voices, so that only the verified people have access to information systems or facilities which must be protected against intrusion by unauthorized persons. The outline of this paper is as follows: Section II offers a theoretical context for the recognition of voices, including roles, styles, phases and areas of voice recognition systems. Section III explains how to execute the proposed system. The experimental results & Conclusion and future work are discussed in Sections IV & V Respectively.

II. BACKGROUND

A. Fundamental Tasks of Speaker Recognition

The recognition of speakers has two basic tasks, namely the authentication of speakers and the identification of speakers as shown in Fig. 1. Speaker verification is the process of determining if an unknown voice originates from a given enrolled speaker, the speaker must assert an identity and the program validates the claimed identity. Speaker authentication applications include telephone banking, machine registration, prevention of fraud by mobile phones, and calling cards [10]. Identification of speakers is the job of associating an unknown voice to one from a group of enrolled speakers. In this case, the speaker offers a voice sample (without claiming an identity) and the program decides which of the identified set of enrolled speakers does belong to the voice sample. Applications for speaker recognition include automated speaker labeling of recorded meetings for speaker-dependent audio indexing and smart response machines with customized caller greetings.

B. Classification of Speaker Recognition Systems

Speaker recognition systems are further categorized by the restrictions imposed on the speech text used in the system; either language-dependent or language-dependent classification may be applied. In the text-dependent case, for each speaker the spoken text or sentence used to train and check the program is fixed [13]. Text-dependent voice recognition systems are mainly used in services such as access control and telephone-based services, where users are considered cooperative [1]. The spoken phrase or text used for the training and testing is not set in the text-independent scenario. Text-independent speaker recognition systems are the most versatile and commonly used in situations where speakers may be considered non-cooperative participants, as they do not explicitly need to be identified such as forensic examination and surveillance procedures. Text-dependent recognition achieves higher recognition performance than text-independent recognition [2], but despite the versatility offered by text-independent recognition, there is an increasing trend in the development of text in-dependent recognition systems [9].

C. Phases of Speaker Recognition

A voice recognition program consists of two separate stages, one training phase and one test phase, as shown in Fig. A speaker voice is captured in the training process, and many vectors of audio features are extracted to create a specific model (speech-print) that uniquely identifies the speaker. In the phase of research, (also known as process of recognition)

D. Applications of Speaker Recognition Systems

The aim of speaker recognition is to identify who is speaking automatically based on individual details used in speech or voice statements. Speaker recognition work has now spanned more than five decades, and has yielded promising results [8], [14], Per Say is one of the leading providers of advanced voice biometric recognition products used to safely, easily and efficiently verify speaker identities 15], [16] Speaker recognition technology is applicable to a broad range of applications, but Authentication and forensics are the two main fields of research. Authentication speaker recognition permits uses or automated systems to recognize a person using their voices. This form of method of authentication is known as authentication by biometric person. An authentication of a biometric person can be used to complement the knowledge-based (usernames and passwords) or token-based (use of physical tokens such as identification cards) methods to allow users to access information or knowledge systems securely and safely [17], [18]. Forensic is an essential relevant service, which speaker recognition technology will support [3]. During telephone conversations, a great deal of information is shared between parties (law-abiding) [19]. When there is a recorded sample of speech during the crime investigation, the words of the offenders can be matched with the recorded voice to help.

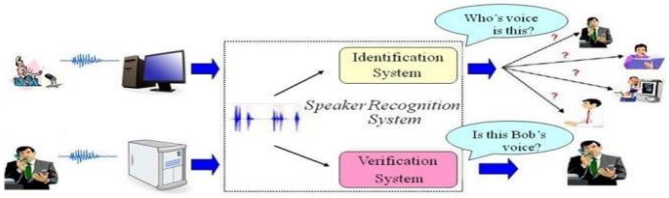
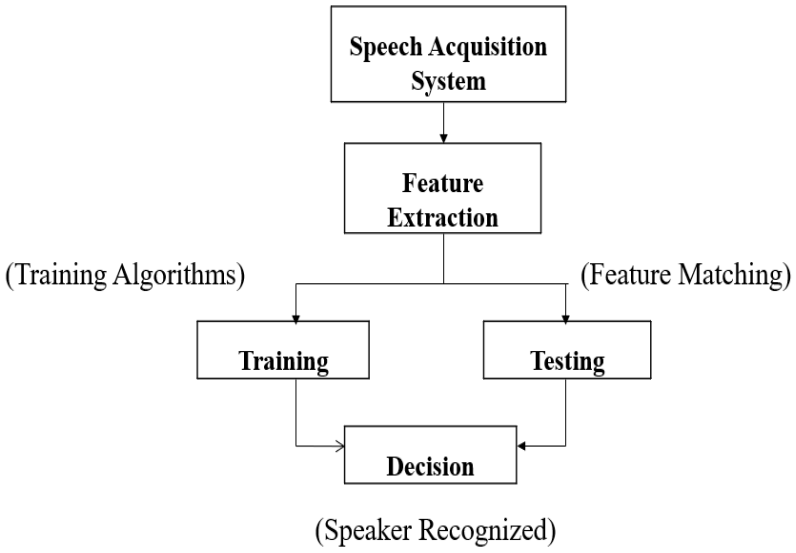


Fig. 1. Speaker Recognition fundamental tasks (Verification and Identification)

III. METHODOLOGY

A systematic workflow of the proposed voice recognition system is given below. The voice activity detection is performed with an input speech signal in the given speech signal to determine the presence of speech or absence of speech [20]. The audio function vectors are then extracted and used to train four standard machine learning algorithms.



MODULES:

1.Dataset

VoxCeleb is a recently published publicly available dataset which serves this purpose appropriately [21]. The two most common audio formats that exist presently are WAV and MP3. For most researches, WAV files are considered more useful as they cover the entire range of frequencies that are audible to the human ear. On the other hand, MP3 files are compressed and hence do not contain all the information that is usually encompassed in a WAV file of the corresponding audio. Additionally, feature extraction from these WAV files is extremely crucial. This phase essentially forms the basis of the machine learning algorithms to be used for classification. Thus, WAV files are largely preferred audio studies, and VoxCeleb does a great job of providing WAV files at constant sampling rates. Consistency in sampling rate is very important in an audio study to ensure that the extracted coefficients represent the same underlying calculations. Hence, this is another reason why VoxCeleb dataset is apt for this work as it has sampled all audios to a constant sampling rate of 16000 Hz.

2.Feature Extraction

In order to determine the sound features relevant to speech processing for speaker recognition, we first need to understand the different type of audio features. Audio features can be broadly classified into three types:

- Rhythmic features
- Temporal features
- Spectral features

Rhythmic features primarily deal with features related to musical notes and are heavily used in applications related to music information retrieval (MIR). Temporal features describe an audio signal over a sampled period. Few examples of temporal

features are zero-crossing rate, minimum amplitude, or maximum frequency. These features are generally used in applications which deal with understanding the continuity of the audio signal – for instance, detecting a sudden change in oceanic wave sounds. Temporal features are sometimes also used in MIR while performing genre classification. For speech processing, spectral features are most effective. Spectral features are based on the frequencies of the audio waves and are used for converting temporal features to equivalents domain of frequency. This is very like how the human ear treats audio signals. The human ear receives the temporal signals, which are then converted to their frequency domains resulting in corresponding vibrations in the human ear giving us the ability to hear and comprehend an audio signal. Precisely for this reason of being very similar to how humans perceive audio, spectral features are the most widely used features in speech and speaker recognition systems. There are several methods used for generating frequency domains for spectral features. Few of the notable ones used in speaker recognition are Linear Predictive Coding (LPC), Rasta filter, and Mel Frequency Cepstral Coefficients (MFCC). Studies suggest that MFCCs represent the closest relation to the human earing model and are becoming increasingly popular in speech recognition.

Mel Frequency Cepstral Coefficients (MFCC):

Mel Frequency Cepstral Coefficients are a set of spectral audio features which are effective in speaker recognition systems. MFCCs are a list of coefficients that in totality represent a Mel Frequency Cepstrum (MFC) [22]. The steps to identify the MFCCs are as follows:

1. Frame or window the signal into blocks that usually range between 20 - 40ms. As discussed earlier, framing is essential before extracting features from the audios. However, unlike most application where a frame length of 20-40ms is considered fine, for speaker recognition systems, scientists have studied the frame length of up to 250ms can yield good results. Taking wider frames makes the frames less specific but also greatly reduces the total number of frames, thereby speeding up the overall process of feature extraction.
2. Perform a Discrete Fourier Transform (DFT) on this framed signal and calculate the powers of the spectrum. Fourier transform is used to split the time-signal function into the multiple frequencies it is composed of. This step is motivated by the fact that the cochlea in human ear vibrates in different patterns and in multiple spots depending on the incoming frequency [23]. Assuming *Si (k)* denotes the DFT for the framed signal *i* and *k* denotes the DFT length, then the formula is:

$$Si(k) = \sum_{n=1}^{N} si(n)h(n)e^{-j2\pi kn/N}$$

Where *h(n)* is an analysis window for the Nth sample.

Further, the spectrum power is calculated as $Pi(k) = (1/N) \mid Si(k) \mid 2$

3. Map the above calculated power spectrums to the mel-scale. The mel-scale represents the multiple pitch scales that an audio receiver or listener interprets. The formula to convert the hertz to mels is given as:

$$m = 2595 \log_{10}(1 + f/700) = 1127 \ln(1 + f/700)$$

Additionally, triangular overlapping filters are also used for the mapping. These are a set of 26 vectors and each is 257 in length. Every value of this filter bank is multiplied with the spectrum power calculated in the previous step followed by adding up the coefficients.

4. Compute the logarithm of energies This step gives us the values of the filter banks. It is obtained by taking the logarithms of the energies obtained in the previous step.
5. Finally, compute the discrete cosine transform (DCT) of the filter bank energies to get the MFCCs.

MFCC Delta: Differentials

Apart from the standard set of coefficients known as the MFCCs, there are variations of it
that exist too. The most prominent ones are those which are obtained by calculating the deltas of the coefficients. They represent the path that the MFCCs encounter, which is known to increase the accuracy of speech recognition researches in general. The first order MFCC deltas are sometimes also referred to as differentials. The formula for calculating the MFCC delta coefficients is:

$$d_t = \sum_{n=1}^{N} n(c_{t+n} - c_{t-n}) / 2 \sum_{n=1}^{N} n2$$

where, *dt* is the delta value coefficient for the window *t*, and *c* represents the MFCC coefficients.

MFCC Delta-Delta: Accelerations

Similarly, the MFCC delta-delta coefficients are calculated when the formula is applied on the MFCC delta coefficients. These are called as MFCC Delta-Delta coefficients, sometimes also referred to as accelerations. The formula for calculating the MFCC delta-delta coefficients is the same as MFCC delta coefficients with the only change being that here we use the delta coefficients instead of the original MFCC coefficients. Thus, the formula is

$$dd_t = \sum_{n=1}^{N} n(d_{t+n} - d_{t-n}) / 2 \sum_{n=1}^{N} n2$$

where *ddt* is the delta-delta coefficient for window *t*, and *dt* represents the MFCC delta coefficients. For simplicity, the MFCC delta-delta calculation can also be represented as follows:

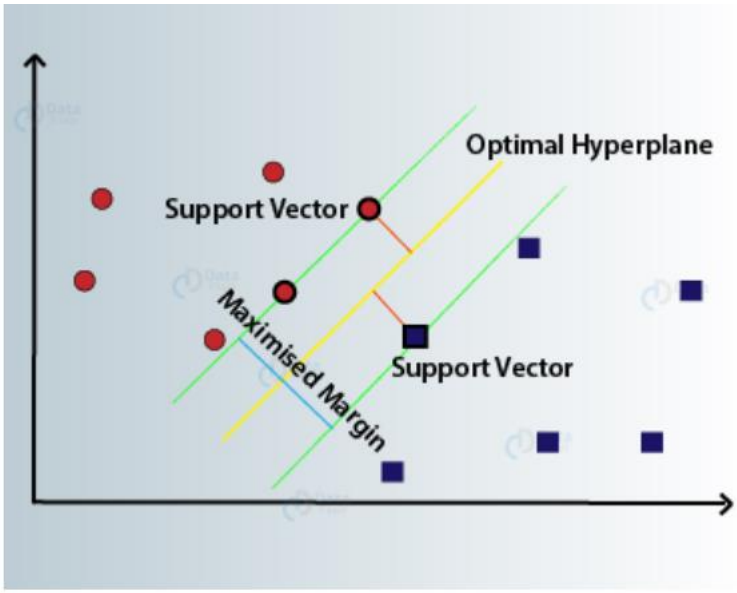
MFCC delta-delta = Δ (MFCC delta) = Δ (Δ (MFCC))) where Δ represents the delta function.

4. Machine Learning Classifier:

Like the feature extraction techniques, the machine learning classifiers also play a vital role in determining the overall effectiveness of speaker recognition model. As we have an intention to classify audios and determine the speaker in them, this is a classification problem and hence we shall discuss about an effective supervised classification machine learning algorithm, Support Vector Machine.

4.1. Support Vector Machine (SVM)

Support Vector Machine, sometimes abbreviated as SVM is a well-known supervised machine learning classification technique. The objective of a SVM algorithm is to construct an n-dimensional hyperplane which can be used for classification or regression.



We can understand the working of support vector machine algorithm with the help of following steps –

1. Consider a data set containing various samples of different types.
2. We have to create a hyperplane that separates the dataset into classes.
3. Our goal is to create a line that classifies the data into two classes, creating a distinction between red triangles and blue circles as shown in above diagram.
4. Let us visualize some of the lines that can differentiate between the two classes in a x-y plane.
5. If you choose one of the line, then it is the ideal line that partitions the two classes properly. However, we still have not concretized the fact that it is the universal line that would classify our data most efficiently.
6. According to SVM, we must find the points that lie closest to both the classes. These points are known as support vectors.

7. We find the proximity between our dividing plane and the support vectors. The distance between the points and the dividing line is known as margin. The aim of an SVM algorithm is to maximize this very margin. When the margin reaches its maximum, the hyperplane becomes the optimal one.
8. The SVM model tries to enlarge the distance between the two classes by creating a well-defined decision boundary.
9. In the above case, our hyperplane divided the data. While our data was in 2 dimensions, the hyperplane was of 1 dimension. For higher dimensions, say, an n-dimensional Euclidean Space, we have an n-1-dimensional subset that divides the space into two disconnected components.

(E) Evaluation

Every model's behavior is assessed based on certain parameters for evaluating its efficiency. The size of the training data, the quality of audio files and, most importantly, the type of machine-learning algorithm used affect the results. The following measures are used to determine the efficiency of the models:

Accuracy: Percentage of examples correctly categorized from all given examples. It is calculated as:

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

Precision: The percentage of true x-class instances for all those listed as class x. It is calculated as:

$$\textbf{Precision}=\frac{tp}{tp+fp}$$

Recall: The percentage of examples listed as class x among all examples of class x. It is calculated as:

$$\textbf{Recall}=\frac{tp}{tp+fn}$$

F_1- measure: is the harmonic mean of precision and recall. It is calculated as:

$$F_1= 2*\frac{precision*recall}{precision+recall}$$

Where

tp = true positives: number of examples predicted positive that are actually positive

fp = false positives: number of examples predicted positive that are actually negative

tn = true negatives: number of examples predicted negative that are actually negative

fn= false negatives: number of examples predicted negative that are actually positive

IV.EXPERIMENT RESULTS:

Experiment 1: SVM + MFCC Coefficients

In this experiment, we test our dataset with the MFCC coefficients (13 attributes) using the SVM classifier. we first perform the experiment using the Polynomial kernel. Based on our implementation, we invoke this method as *svc_poly = SVC(gamma = 'scale', kernel = 'poly')*. The total number of correctly classified instances using the Polynomial kernel are 3350 out of 4304 and the total number of incorrectly classified instances are 954 out of 4304.

Overall model summary: SVC - Poly Kernel
Total number of speakers tested: 4304
Correctly classified speakers: 0.7783 3350/4304
Incorrectly classified speakers: 0.2217 954/4304
Precision: 0.7796
Recall : 0.7773
F1 Score: 0.7771

Experiment 2: SVM + MFCC Delta Coefficients

In this experiment, we test our dataset with the MFCC Delta coefficients (26 attributes) using the SVM classifier. we first perform the experiment using the Polynomial kernel. We invoke the method as *svc_poly = SVC(gamma = 'scale', kernel = 'poly')*. The total number of correctly classified instances using the Polynomial kernel are 3324 out of 4304 and the total number of incorrectly classified instances are 980 out of 4304.

Overall model summary: SVC - Poly Kernel
Total number of speakers tested: 4304
Correctly classified speakers : 0.7723 3324/4304
Incorrectly classified speakers: 0.2277 980/4304
Precision: 0.7742
Recall: 0.7717
F1score: 0.7716

Experiment 3: SVM + MFCC Delta Delta Coefficients

In this experiment, we test our dataset with the MFCC Delta Delta coefficients (39 attributes) using the SVM classifier. we first perform the experiment using the Polynomial kernel. We call the method as *svc_poly = SVC(gamma = 'scale', kernel = 'poly')*. The total number of correctly classified instances using the Polynomial kernel are 3324 out of 4304 and the total number of incorrectly classified instances are 980 out of 4304.

Overall model summary: SVC - Poly Kernel
Total number of speakers tested: 4304
Correctly classified speakers: 0.7711 3319/4304
Incorrectly classified speakers: 0.2289 985/4304
Precision: 0.7726
Recall: 0.7706
F1 Score: 0.7703

Thus, a total of 3 experiments were performed using three datasets having different number of MFCC coefficients - with and without delta coefficients. These datasets were then loaded into SVM machine learning classifiers. As this is effectively a multi-class classification research, the metrics recall, F1 score, and the classification accuracy are of most significance to the study. Below is a summary of how each combination of feature extraction and machine learning technique perform when compared with every other technique.

ML Classifier	Data set	Classification accuracy	Precision	Recall	F1Score
SVM	MFCC	0.7783 (3350/4304)	0.7796	0.7773	0.7771
SVM	MFCC Delta	0.7723 (3324/4304)	0.7742	0.7717	0.7716
SVM	MFCC Delta-Delta	0.7711 (3319/4304)	0.7726	0.7706	0.7703

V. CONCLUSION & FUTURE WORK

This work is broadly divided into three parts – audio preprocessing, feature extraction, and machine learning classification. As the audios used in our study were not recorded in constrained environments, audio pre-processing was an extremely crucial part of the research. The two most important things we focused on for pre-processing were reducing the ambient noise and highlighting the human vocals. we believed that just using the MFCC coefficients could be

effective in the study. Our results proved it correct as our F1 scores increased by 1.37%, using on models trained with just the MFCC coefficients instead of the MFCC delta-delta coefficients.

Apart from the existing audio processing techniques, it will be interesting to see how other approaches can impact the feature extraction and classification accuracy of the machine learning models. The current approaches generalize the high-shelf and low-shelf frequencies to limit their gain for suppressing the ambient noise and enhancing the human vocals. Although this approach performs well, noise continues to exist in audios which limits the accuracy. More research in this area to achieve complete elimination or almost 'close to complete' elimination of noise can greatly improve the subsequent feature extraction and classification tasks for speaker recognition.

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] N. Singh, R. Khan, and R. Shree, "Applications of speaker recognition," Pmcedia engineering, vol. 38, pp. 3122-3126, 2012.

[3] 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.

[4] E. Aliyu, O. Adewale, and A. Adetunmbi, "Development of a textdependent speaker recognition system," International Journal of Computer Applications, vol. 69, no. 16, 2013.

[5] 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.

[6] 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.

[7] A. K. Jain, K. Nandakumar, and A. Ross, "50 years of biometric research: Accomplishments, challenges, and opportunities," Pattern Recognition Letters, vol. 79, pp. 80-105, 2016.

[8] Statistics South Africa (STATS SA), Census 2011, [Online]. Available: http://www.statssa.gov.za/publications/Report-03-01-78/Report-0301-782011.pdf. [Accessed 25 August 2018].

[9] 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

[10] F. Bimbot, J.-ille, G. Gravier, I. MagrinChagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcfa, D. PetrovskaDelacr£taz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," EURASIP Journal on Advances in Signal Processing, vol. 2004, no. 4, p. 101962, 2004.

[11] D. Ferbrache, "Passwords are broken-the future shape of biometrics," Biometric Technology Today, vol. 2016, no. 3, pp. 5-7, 2016.

[12] L. Hamid, "Biometric technology: not a password replacement, but a complement," Biometric Technology Today, vol. 2015, no. 6, pp. 7-10, 2015.

[13] L. Gbadamosi, "Text independent biometric speaker recognition system," International Journal of Research in Computer Science, vol. 3, no. 6, p. 9, 2013.

[14] N. J. De Vries, M. H. Davel, J. Badenhorst, W. D. Basson, F. de 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.

[15] J. Ramirez, J. C. Segura, C. Benitez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," Speech communication, vol. 42, no. 3-4, pp. 271-287, 2004.

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

[17] 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.

[18] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5-32, 2001.

[19] 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.

[20] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5-32, 2001.

[21] A. Nagrani, J.S. Chung, A. Zisserman, VoxCeleb: a large-scale speaker identification
dataset, Interspeech, Department of Engineering Science, University of Oxford, 2017. Available at: http://www.robots.ox.ac.uk/~vgg/data/voxceleb/

[22] Mel-frequency Cepstrum, Wikipedia, Available at: https://en.wikipedia.org/wiki/Melfrequency_ cepstrum [online]

[23] J. Lyons, Mel Frequency Cepstral Tutorial, Practical Cryptography. Available at:
http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstralcoefficients- mfccs/ [online]

[24] Tumisho Billson Mokgonyane . Tshephisho Joseph Sefara, Thipe Isaiah Modipa etl “Automatic Speaker Recognition System Based on Machine learning Algotithms” , SAUPEC/RobMech/PRASA Conference Bloemfontein, South Africa, January 28-30, 2019, 978-1-7281-0369-3/19/$31.00 c2019 IEEE