International Journal of Intelligent Information Systems
2024, Vol. 13, No. 5, pp. 94-108
https://doi.org/10.11648/j.ijiis.20241305.11

Research Article

Speaker Recognition System Using Hybrid of MFCC and RCNN with HCO Algorithm Optimization

Stephen Nyakuti Otenyi*, **Livingstone Ngoo**, **Henry Kiragu**

Electrical and Telecommunication Engineering Department, Multimedia University of Kenya, Nairobi, Kenya

Abstract

Though there are advancements in speaker recognition technology, available systems often fail to correctly recognize speakers especially in noisy environments. The use of Mel-frequency cepstral coefficients (MFCC) has been improved using Convolutional Neural Networks (CNN) yet difficulties in achieving high accuracies still exists. Hybrid algorithms combining MFCC and Region-based Convolutional Neural Networks (RCNN) have been found to be promising. In this research features from speech signals were extracted for speaker recognition, to denoise the signals, design and develop a DFT-based denoising system using spectrum subtraction and to develop a speaker recognition method for the Verbatim Transcription using MFCC. The DFT was used to transform the sampled audio signal waveform into a frequency-domain signal. RCNN was used to model the characteristics of speakers based on their voice samples, and to classify them into different categories or identities. The novelty of the research was that it used MFCC integrated with RCNN and optimized with Host-Cuckoo Optimization (HCO) algorithm. HCO algorithm is capable of further weight optimization through the process of generating fit cuckoos for best weights. It also captured the temporal dependencies and long-term information. The system was tested and validated on audio recordings from different personalities from the National Assembly of Kenya. The results were compared with the actual identity of the speakers to confirm accuracy. The performance of the proposed approach was compared with two other existing speaker recognition the traditional approaches being MFCC-CNN and Linear Predictive Coefficients (LPC)-CNN. The comparison was based the Equal Error Rate (EER), False Rejection Rate (FRR), False Match Rate (FMR), and True Match Rate (TMR). Results show that the proposed algorithm outperformed the others in maintaining a lowest EER, FMR, FRR and highest TMR.

Keywords

Speaker Recognition, Verbatim Transcription, Spectrum Subtraction, Mel-Frequency Cepstral Coefficients (MFCC), Linear Predictive Cepstral Coefficients, HCO Algorithm

1. Introduction

Speaker recognition is the process of identifying a person based on their voice characteristics. It has various applications in security, forensics, biometrics, and speech transcription [1]. However, speaker recognition is a challenging task due to the variability and complexity of speech signals, as well as the presence of noise and interference that are common in parliaments [1, 2]. The national assemblies of African countries often have to deal with problems including multilingualism, dialectal variation, and low-resource languages [3, 4]. These can make speaker recognition methods less effective or even

*Corresponding author: nyakuti10@gmail.com (Stephen Nyakuti Otenyi)

Received: 3 July 2024; **Accepted:** 5 August 2024; **Published:** 10 October 2024

unusable.

In Kenya, the National Assembly has four official languages: English, Swahili, Kenyan Sign Language, and Braille [5]. Therefore, there is a need for a better and more robust speaker recognition method that can handle the diversity and complexity of the speech signals in the national assemblies of African countries, especially Kenya. Such a system should be able to extract features from different languages and dialects, denoise speech signals from various sources of interference, and perform better than the available approaches that rely on simple methods or limited data [1, 6]. Variability and complexity of speeches in the National Assembly of Kenya makes it challenging for the plenary proceedings to be accurately recorded [7].

Despite significant research in speaker recognition systems and technologies, current systems still struggle with correctly recognizing speaker especially in noisy environments [1, 6]. There is still a challenge and a need of recognizing speakers in noisy environments. Noting that the Parliament of Kenya and the 47 County Assemblies in Kenya and expected to produce verbatim record of the plenary proceedings as well as committees. There is need to achieve a robustness to address challenges such as variability of the speech signal due to different microphones, distances, orientations, and acoustic conditions. The interference of the background noise, such as applause, laughter, coughing, and other speakers and the overlap of multiple speakers, especially in debates and discussions need to also be addressed. Factors such as the diversity of the speakers, such as gender, age, accent, language, and speaking style are challenges that can be handled by a robust system.

The researchers aimed to achieve robustness using Mel-frequency cepstral coefficients (MFCC) as the acoustic features and region-based convolutional neural networks (R-CNN) reinforced with hosted cuckoo optimization (HCO) algorithm to classify the features. The novel optimized R-CNN combines the advantages of CNNs and recursive neural networks (RNNs) to address problem of variable lengths of speech inputs by dividing the input into regions and applying CNNs to each region. The algorithm then aggregates the outputs using RNNs to extract both local and global features. It has been found applicable in real-time speaker identification relevant to parliamentary settings.

In this research, speaker recognition method for the Verbatim Transcription of the Kenyan National Assembly was developed using advanced techniques for feature extraction, denoising, and deep learning. Discrete Fourier Transform (DFT) is used to perform spectrum subtraction for noise reduction, and Mel-frequency cepstral coefficients (MFCC) are extracted from speech signals. A region-based convolutional neural network (R-CNN) optimized with hybrid cuckoo optimization (HCO) algorithm is used to model and classify the speakers based on their voice samples. Performance of the proposed system on audio recordings from different personalities from the National Assembly of Kenya is evaluated and compared with two other existing speaker recognition approaches: MFCC-CNN and LPC-CNN.

2. Speaker Recognition

2.1. Components of Speaker Recognition Method

There are many components involved in the process of speaker recognition. They include feature extraction, normalization, selection, transformation, classification and decision [8, 9]. There are also other aspects that need to be considered, such as database design, evaluation metrics, security issues, ethical issues, etc. [10]. Feature extraction component of speaker recognition method extracted relevant information from the speech signal. The features are used to represent the characteristics of the speaker's voice. The component transforms the raw speech signal into a set of features that capture the characteristics such as pitch, timbre, accent, pronunciation, energy, spectral shape, etc. [11]. These features were then used as input to a classifier. There were different types of features that were to be extracted from speech signals, such as spectral features, cepstral features, prosodic features, and phonetic features [8, 12]. Spectral features are based on the frequency spectrum of the speech signal, which reflects the shape of the vocal tract and the resonance of the vocal cords. Cepstral features are derived from the spectral features by applying a logarithmic transformation and a discrete cosine transform to reduce the correlation between adjacent features and enhances the speaker-specific information [13]. Prosodic features are related to the variations in pitch, intensity, and duration of speech segments, which reflect the speaker's emotion, attitude, and intonation [14]. Phonetic features are based on the linguistic content of the speech signal, such as the pronunciation of vowels, consonants, and phonemes, which can vary depending on the speaker's native language, dialect, and accent [15].

Convolutional Neural Networks (CNNs) can learn to extract hierarchical and nonlinear features from the input signal. Two most widely used method for feature extraction are linear predictive coding (LPC) coefficients and Mel-frequency cepstral coefficients (MFCCs) [16, 17]. LPC coefficients can be derived from the autocorrelation method, the covariance method, or the Burg method, among others [18]. LPC coefficients represent the spectral envelope of the speech signal, which is related to the vocal tract shape and hence to the identity of the speaker. MFCCs has a lower dimensionality than LPC coefficients, which means that they require less computation and memory [19]. However, a lower dimensionality may also result in a loss of information and a degradation of performance. MFCCs are generally more robust than LPC coefficients, especially in noisy conditions, because they reduce the effects of the high-frequency components that are more susceptible to noise [20]. LPC coefficients are more sensitive to noise and channel distortion, because it relies on

the accurate estimation of the linear prediction coefficients.

Feature normalization component reduces the variability of the features due to factors such as noise, channel distortion, or different recording conditions [21]. This can improve the performance of the system by making the features more robust and discriminative. Feature normalization techniques can be divided into two categories: short-term and long-term normalization. Short-term normalization operates on a frame-by-frame basis, while long-term normalization operates on a segment or utterance level. Examples of short-term normalization include cepstral mean subtraction (CMS), cepstral mean and variance normalization (CMVN), and relative spectral transform analysis (RASTA) filtering [22]. Some examples of long-term normalization are feature warping, histogram equalization, and test normalization. CMS is a simple method that subtracts the mean of each cepstral coefficient from the corresponding coefficient in each frame. CMS can reduce the effect of channel distortion, but it cannot cope with non-stationary noise or speaker variability [23]. CMVN is an extension of CMS that also divides each cepstral coefficient by its standard deviation. CMVN can reduce the effect of both channel distortion and non-stationary noise, but it still cannot handle speaker variability [24]. RASTA is a more complex method that applies a band-pass filter to each cepstral coefficient along the time axis. According to some studies, RASTA is generally superior to CMS and CMVN in terms of speaker recognition accuracy, especially in noisy or mismatched conditions [25]. However, RASTA may not be suitable for real-time applications due to its high computational cost. Therefore, the best recommended normalization process for speaker recognition depends on the specific scenario and the trade-off between accuracy and efficiency [25]. A possible solution is to combine RASTA with other techniques, such as feature selection or dimensionality reduction, to reduce the complexity and redundancy of the normalized features.

Feature selection component chooses the most relevant features for speaker recognition, based on some criteria such as information content, redundancy, or computational complexity [26]. The main goals of feature selection are to reduce the dimensionality of the feature space, to improve the performance of the recognition system, and to reduce the computational complexity and storage requirements. There are different methods for feature selection, such as filter methods, wrapper methods, and embedded methods. Filter methods evaluate the features independently of the classifier and rank them according to some criteria, such as information gain, mutual information, or Fisher's ratio. Wrapper methods use the classifier as a black box and search for the optimal subset of features that maximizes the classification accuracy. Embedded methods integrate the feature selection process into the classifier training and select the features that are most relevant for the classifier complexity [26].

Feature transformation component transforms the features into a new space that can better capture the speaker-specific information. Feature transformation techniques are often applied to extract more robust and discriminative features that can capture the characteristics of the speaker. Feature transformation using linear discriminant analysis (LDA) or principal component analysis (PCA) can be used to project the features onto a lower-dimensional subspace that maximizes the inter-speaker variability and minimizes the intra-speaker variability [27]. Fast Fourier Transform (FFT) can be used to convert a time-domain signal into a frequency-domain representation, which can reveal the spectral properties of the signal [28]. It can also be used to compute features such as Mel-frequency cepstral coefficients (MFCCs). CNNs can be used to directly process raw waveform data or spectrogram images, and can achieve state-of-the-art performance in speaker recognition tasks.

Classification component classifier compares features with a database of known speakers and assigns a label to the unknown speaker. There are many methods for classification, such as Gaussian mixture models (GMMs), hidden Markov models (HMMs), CNNs, SVMs, FFT etc. The choice of the classifier depends on factors such as accuracy, speed, scalability, and adaptability. CNNs have been shown to achieve the highest accuracy in speaker recognition, especially when combined with other techniques such as i-vectors and attention mechanisms [1]. CNNs can learn complex and high-level features from the speech signals that are discriminative for speaker identification. GMMs and HMMs are also widely used methods that have good accuracy, but they rely on hand-crafted features such as MFCCs and require careful tuning of parameters such as the number of mixtures and states [29]. SVMs are another popular method that can achieve high accuracy with a suitable kernel function and regularization parameter. However, SVMs suffer from the curse of dimensionality and may not perform well on high-dimensional feature spaces. FFT is the fastest method, as it can perform a linear transformation on a speech signal in O (N log N) time, where N is the length of the signal [30, 31]. CNNs are the slowest methods, as they involve multiple layers of nonlinear transformations and require large amounts of training data and computational resources. However, CNNs can benefit from parallelization and optimization techniques such as GPU acceleration and batch normalization to improve their speed.

Decision component of speaker recognition method makes a final decision based on the output of the classifier. There are different ways to make decision, such as thresholding, voting, fusion, or rejection. The decision can also be influenced by prior knowledge, such as speaker models, enrollment data, or background information.

2.2. Denoising by Spectral Subtraction

Speech denoising is the process of removing noise from speech signals. It can improve the quality and intelligibility of speaker recognition methods. Spectrum subtraction assumes that the noise spectrum is additive and can be estimated from the noisy speech spectrum. It can be performed in different domains, such as time, frequency, or cepstral, but the most

common one is the frequency domain, where the Fast Fourier Transform (FFT) is used to convert the speech signal into a sequence of short-time spectra [32].

The basic idea of spectrum subtraction is to subtract an estimate of the noise spectrum from the noisy speech spectrum, and then apply an inverse FFT to obtain the enhanced speech signal [33]. However, this simple method can introduce some artifacts, such as musical noise and speech distortion, due to inaccurate noise estimation or over-subtraction. Spectral floor is a method that sets a lower limit for the subtracted spectrum, so that any negative value is replaced by a small positive value. This prevents the generation of musical noise, which is a common artifact of spectral subtraction. However, spectral floor may introduce some bias in the enhanced speech spectrum, and reduce the signal-to-noise ratio (SNR) improvement. Spectral gain function is a method that applies a nonlinear function to the subtracted spectrum, so that the negative values are mapped to zero or near-zero values [34]. This also reduces the musical noise, but preserves more of the original speech spectrum than spectral floor. However, spectral gain function may introduce some distortion in the enhanced speech spectrum, and affect the speech quality. Spectral smoothing function is a method that applies a low-pass filter to the subtracted spectrum, so that the high-frequency variations caused by noise are smoothed out [35].

The most suitable method among these three depends on the application and the type of noise. For example, spectral floor may be more suitable for low SNR conditions, where noise reduction is more important than speech quality. Spectral gain function may be more suitable for medium SNR conditions, where both noise reduction and speech quality are important. Spectral smoothing function may be more suitable for high SNR conditions, where speech quality and intelligibility are more important than noise reduction [36].

2.3. Mel-frequency Cepstral Coefficient Subtraction

Mel-frequency cepstral coefficient (MFCC) subtraction is a technique for enhancing the speaker recognition performance of speech systems in noisy environments [37]. MFCC subtraction assumes that the noise spectrum is relatively stationary and can be estimated from the silent segments of the speech signal. By subtracting the noise spectrum from the speech spectrum, the MFCC features of the clean speech can be recovered.

MFCC subtraction can improve the speaker recognition accuracy by reducing the mismatch between the training and testing conditions. MFCC subtraction can also be combined with other noise reduction methods, such as spectral subtraction, Wiener filtering, or cepstral mean normalization, to further enhance the speech quality and speaker recognition performance [38, 39]. Pre-emphasis stage entails filtering the signal through a high-pass filter to emphasize the higher frequencies. This is followed by framing that breaks down the audio samples into small frames of between 20 and 40 milliseconds. Windowing is then carried out through consideration of the next block in the feature extraction process and integrating all the adjacent frequency lines. FFT is then performed to convert time to frequency domain followed by Mel filter bank processing. DFT transforms the pitch energies to time domain to enable frame analysis [40].

2.4. Region-based Convolutional Neural Networks (R-CNN) Modeling

Region-based convolutional neural networks (R-CNNs) are a type of deep learning model that combines the merits of convolutional neural networks (CNN) with recursive neural networks (RNN) [41, 42]. CNN can capture complex patterns and features from high-dimensional speech signals. In speaker recognition methods, CNNs can be used to model the characteristics of speakers based on their voice samples, and to classify them into different categories or identities [42]. One of the challenges of speaker recognition is to deal with the variability and noise in speech signals, which can affect the performance of the system. To address this issue, CNNs can use multiple layers of filters and pooling operations to extract robust and discriminative features from the raw speech data [43]. Moreover, CNNs can use techniques such as batch normalization, dropout, and regularization to prevent overfitting and improve generalization.

RNNs can process sequential data, such as speech, by maintaining a hidden state that encodes the previous inputs. They can learn long-term dependencies and capture the temporal dynamics of speech signals and can be used to extract features from speech frames or segments, and then feed them to a classifier, such as a softmax layer or a support vector machine (SVM) [44]. R-CNNs can perform object detection and segmentation by applying CNNs to region proposals generated by a selective search algorithm. They can learn features that are invariant to scale, rotation, and translation, and can handle complex backgrounds and occlusions [45].

One way to use R-CNNs for speaker recognition is to integrate them with the i-vector framework. It is a popular method for extracting low-dimensional speaker embeddings from high-dimensional acoustic features. The i-vector framework consists of two steps: first, a universal background model (UBM) is trained on a large set of speakers to capture the general variability of speech; second, a total variability matrix is estimated to model the speaker- and channel-dependent variability of speech [46]. Hourri et al. proposed ConvVector framework extracts speaker characteristics by constructing CNN filters linked to the speaker. It achieves an equal error rate (EER) of 1.05% on a gender-dependent corpus under different noise conditions [47]. The framework consists of two main components: a ConvVector generator and a ConvVector classifier. The ConvVector generator takes a speaker's utterance as input and produces a set of filters that capture the speaker's acoustic features. The ConvVector clas-

sifier then applies these filters to another utterance and computes a similarity score between the two utterances based on the filter responses. The similarity score reflects how likely the two utterances belong to the same speaker [47]. Conv2D can also handle noise and channel variability well, but it requires more computational resources than ConvVector [2].

3. Research Methodology

Figure 1 illustrates the proposed method for speaker recognition using MFCC and region-based convolutional neural networks optimized with HCO algorithm. Speech signals were input into the system and were preprocessed by adding known white Gaussian noise [23, 32]. The signal-to-noise ratio was adjusted between 1 and 20dB, followed by denoising through spectrum subtraction. The denoised signal underwent MFCC feature extraction. These features were then utilized to create speaker models, which were preserved in a database. Subsequently, speech signal features were compared with the database entries to ascertain speaker identity. The system then determined the speaker's identity as either known or unknown, which constituted the system's output.

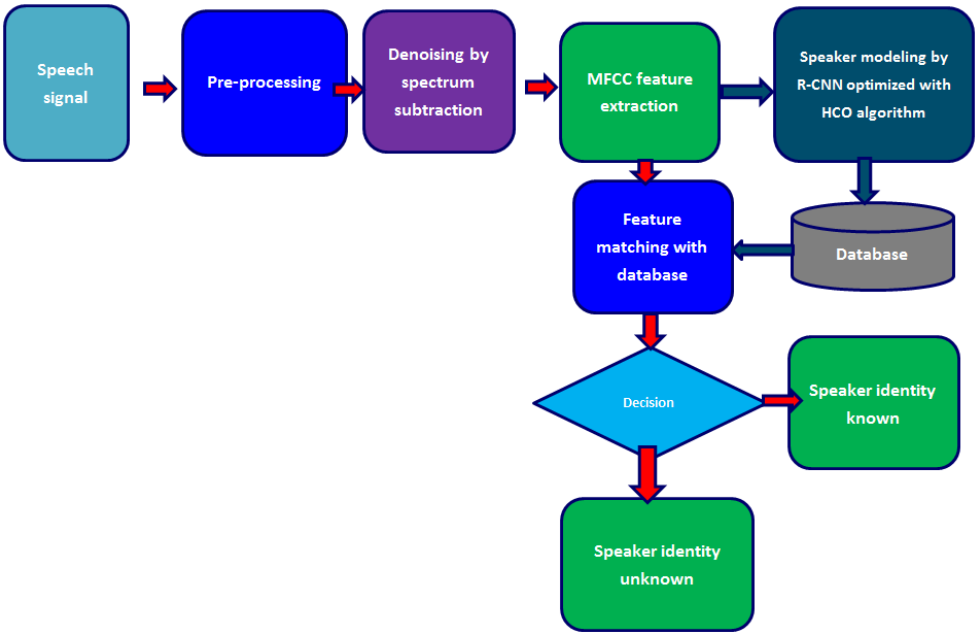

Figure 1. *Proposed Speaker Recognition method.*

3.1. Denoising by Spectrum Subtraction

A time smoothing window was utilized on a noisy signal to mitigate high-frequency noise and utilize the correlation between consecutive samples. Subsequently, a noise reduction filter was applied to the smoothed signal to approximate the desired speech signal. Additionally, a Wiener filter was employed to reduce the mean square error between the estimated signal and the actual one. [38]. Next, Kalman filter was used on the same noisy signals to estimate the state of a dynamic system from a series of noisy measurements [48].

The SNR of the estimated speech signal was calculated as shown in Equation 1.

$$SNR = 10log_{10}\left(\frac{\sum_{n=1}^{N} s^2(n)}{\sum_{n=1}^{N}(s\{n\}-\hat{s}(n))^2}\right) \quad (1)$$

Where;

s was the true speech signal;
$\hat{s}$ was the estimated speech signal; and
N was the number of samples.

3.2. Feature Extraction from Speech Signals

Feature extraction was carried out using MFCC. Pre-emphasis filter was applied to boost the high frequencies and any DC offset was removed from the signal. Equation 2 illustrates how the filter was applied to improve the signal power.

$$S_{out}(n) = S_{in}(n) - \alpha S_{in}(n-1) \quad (2)$$

Where,
$S_{out}(n)$ was the output speech signal;
$S_{in}(n)$ was the input speech signal;
α was the filtering constant ranging between 0.9 and 1;
Speech was framed into 25 ms overlapping segments using

Q samples, with P=100 and Q=256 creating common overlaps; Hamming window, $W(n)$, applied.

$$W(n) = W_0\left(n - \frac{N-1}{2}\right), for\ 0 \le n \le N-1 \quad (3)$$

Where,

N was number of samples per bin;

W_0 was the window coefficient applied to samples;

The Output speech signal, $S_{out}(n)$, was defined in terms of input speech signal, $S_{in}(n)$, and hamming window, $W(n)$, as shown in Equation 4.

$$S_{out}(n) = S_{in}(n)W(n) \quad (4)$$

Hamming window was used to extract pitch coefficients to help reduce the signal value towards zero at the window boundary and to avoid interruption. Equation 5 illustrates the impulse response of the Hamming window.

$$W(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \quad (5)$$

The sample in the time domain, $h(t)x(t)$, was converted into the frequency domain $Y(f)$ for each frame using Equation 6.

$$Y(f) = FFT\big(h(t)x(t)\big) = H(f)X(f) \quad (6)$$

Where,

$H(f)X(f)$ is the FFT of $h(t)x(t)$

The Discrete Fourier Translate (DFT) of each frame was computed to obtain the magnitude spectrum [49]. A filter bank of triangular filters spaced according to the Mel-scale was applied. The Mel-scale was defined in Equation 7.

$$mf = 2595\, log_{10}\left(1 + \frac{f}{700}\right) \quad (7)$$

Where,

mf was the Mel-frequency and

f was the linear frequency in Hz.

Logarithmic compression and DCT applied to filter bank outputs yielded decorrelated MFCC vectors, reducing redundancy and mimicking auditory perception as shown in Equation 8.

$$C_n = \sum_{k=1}^{k}(\log S_k)\cos\left(\frac{n\pi}{k}\left(k - \frac{1}{2}\right)\right) \quad (8)$$

Where,

$n = 1,2,\cdots,k$; and

$S_k$is the output of the last step

3.3. MFCC-R-CNN Based Speaker Recognition

The system identified speakers by modeling their voices using MFCC-based speech features and an R-CNN optimized by the HCO algorithm. The R-CNN processed spectrograms of speech signals, pinpointing phonetic elements to extract features. This MFCC-R-CNN fusion aimed to create a robust speaker model, capturing spectral and temporal aspects of speech. However, R-CNN training was resource-intensive and may overfit. The HCO algorithm optimized parameters of R-CNN to mitigate this. R-CNN also used max-pooling to learn scale, rotation, and translation invariant features, reducing overfitting and computational demands. A pooling layer handled complex scenarios, focusing on pertinent image areas. Speaker embedding layer mapped features to vectors, encapsulating speaker identity. The speaker embedding layer, trained via triplet loss, optimized feature distances for accurate speaker recognition. It was enhanced with an attention mechanism that prioritized relevant features. To combat overfitting and bolster generalization, techniques like batch normalization, dropout, and regularization were employed. [50]. The dropout randomly dropped out some units in each layer during training to reduce the co-adaptation of features and increase the diversity of the network [51]. Batch normalization ensured consistent input distribution, dropout prevented feature co-dependence, and regularization controlled weight complexity, aiding in model generalization. The model adeptly handled sequential data, encoding past inputs into a hidden state. It leveraged RNN, CNN, and LSTM networks for robust feature extraction from speech, culminating in a softmax layer for final classification or regression.

For R-CNN, the output vector, y, was expressed in Equation 9.

$$y = f(W^T X(n) + b) \quad (9)$$

Where,

f was the activation function;

$W^T$was the transform of the weight matrix;

$X(n)$ was an input vector, and

b was a bias vector.

HCO algorithm was used optimize the R-CNN. HCO consists of three operators: levy flight, egg laying, and host bird selection [42]. Levy flight is a random walk process that follows a power-law distribution. It was used to simulate the movement of cuckoos in search of host nests. The levy flight was expressed in Equation 10.

$$X_{i,t+1} = X_{i,t} + \alpha L(\lambda) \quad (10)$$

Where,

$X_{i,t}$was the position of the i^{th} cuckoo at iteration t,

α was a scaling factor, and

$L(\lambda)$ was a levy distribution with exponent lambda.

Egg laying is a method for creating novel solutions by altering existing ones, akin to cuckoos secretly nesting in host nests. [52]. The egg laying was expressed as shown in Equation 11.

$$y_i = x_i + \beta(x_j - x_k) \quad (11)$$

Where,

y_i was the new solution (egg) generated by the i[th] cuckoo;

x_j and x_k were randomly selected solutions from the current population, and

β was a mutation factor.

Host bird selection involved choosing superior solutions from existing and new populations, predicated on host birds identifying and removing certain cuckoo eggs. The host bird selection was expressed as illustrated in Equation 12.

$$if\ f(y_i) < f(x_i) then\ x_i = y_i \quad (12)$$

Where,

$f(y_i)$ and $f(x_i)$ were the objective functions to be minimized, and

x_i and y_i were the old and new solutions, respectively.

With HCO algorithm, the new solution, $X_{i,t+1}$, for the i[th] R-CNN at iteration t+1 was expressed as Equation 13.

$$X_{i,t+1} = X_{i,t} + L(\lambda)(X_{i,t} - X_{j,t}) \quad (13)$$

Where,

$X_{i,t}$ is the current solution for the i[th] R-CNN at iteration t;

$L(\lambda)$ is a levy flight with scale parameter; and

$X_{j,t}$ was a randomly chosen solution from the population, j, at iteration t.

4. Results and Discussions

Results in Figure 2 illustrate original, noisy and denoised audios. Periods of silence were less with original signal with background noise compared to original signal without background noise. When Gaussian noise of SNR=10 was added, the noisy signal is shown in top left graph of Figure 2. The denoised signal had a lower amplitude compared to noisy signal.

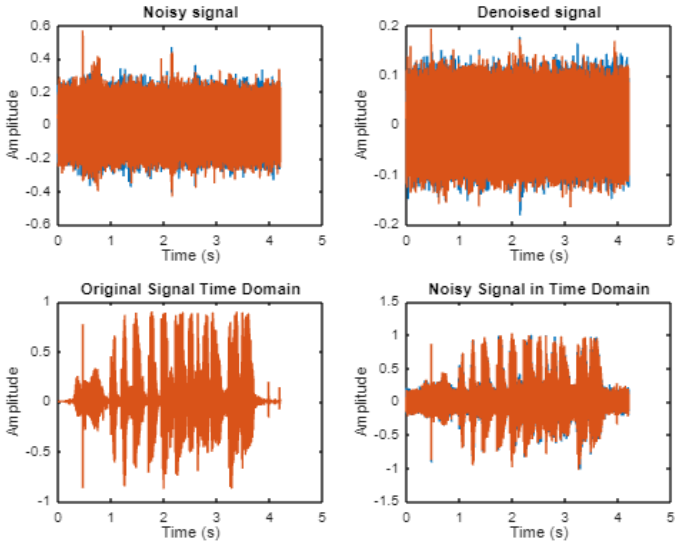

Figure 2. *Original, noisy and denoised audio signals.*

Figure 3 illustrates magnitude and phase response versus normalized audio frequency. The filter phase rises exponentially from -0.1 at normalized frequency of 0 to maximum of 1.25 at normalized frequency of 1. The results suggest that the filter introduces more delay as the frequency increases, which could be indicative of a higher-order all-pass filter characteristic.

The magnitude response of the filter revealed a pronounced peak. The corresponding graph depicted a pattern where the magnitude lessened with rising frequency, aligning with a typical behavior of low pass. This suggested a reduction in the audio intensity at elevated frequencies. Conversely, the phase response graph illustrated the phase alteration in relation to frequency. An increasing phase shift with frequency implied a progressive delay in the signal correlating with frequency. The data indicated that the audio signal underwent processing that impacted both its amplitude and phase. Starting at -27dB at a normalized frequency of 0, the signal abruptly climbed to a peak of 5dB at a normalized frequency of 0.1, then plummeted back to -27dB at a normalized frequency of 1. This behavior typifies a band-pass filter, particularly one with a narrow passband centered at the normalized frequency of 0.1. The sharp ascent to the apex implied a significant resonance at that frequency, while the decline thereafter reflected the filter's rate of attenuation for frequencies falling outside the passband.

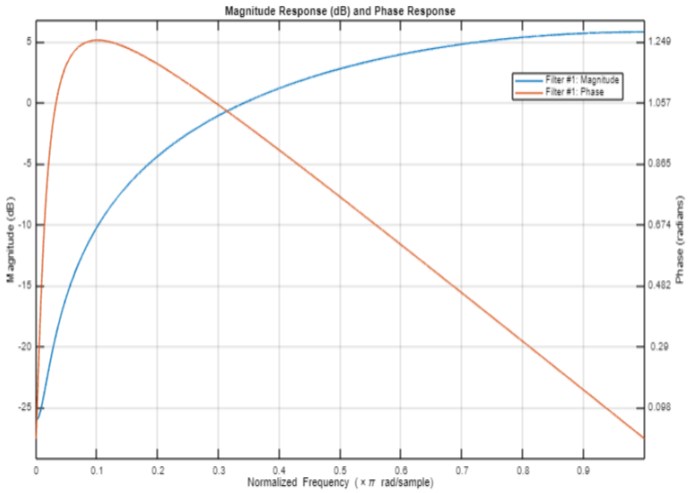

Figure 3. *Magnitude and Phase response against normalized frequency.*

Figure 4 illustrates power spectrum of original and noisy signals. The power for original signal ranged between -30dB and -140dB. For noisy signal the power ranged between -30dB and -70dB. The original signal has a power range that extends from -30dB to -140dB, which suggests a wide dynamic range and possibly a high degree of variation within the signal itself.

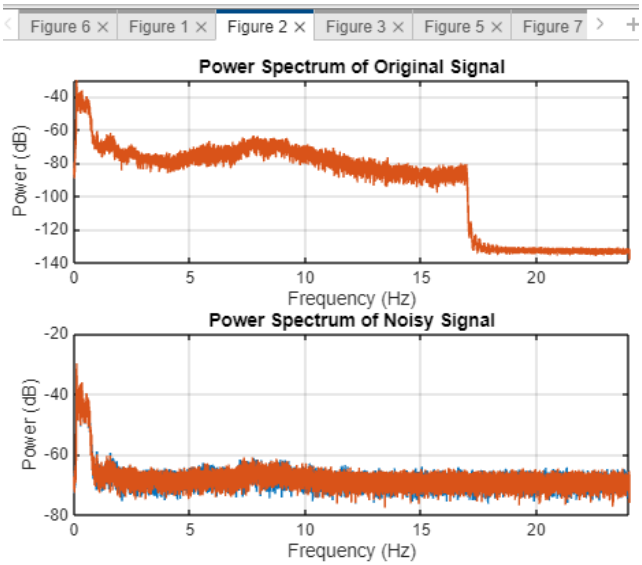

Figure 4. *Power spectra of original and noisy signals.*

This indicated a complex signal characterized by multiple frequency components or a signal with a high signal-to-noise ratio. Conversely, the noisy signal presented a much narrower power spectrum, ranging from -30dB to -70dB. This suggested that the noise within the signal was quite substantial, possibly dominating the original signal at certain frequencies. The power levels' convergence at -30dB for both signals might have implied that the noise floor was at this level, with any signal components beneath this threshold likely obscured by noise.

The denoising scheme used in the proposed algorithm was effective in that despite the Gaussian noise SNR increasing from 1 to 20dB, the denoised signals had greater SNR compared to input signals. Generally, SNR of output signal increased with increase of SNR for input signal. Figure 5 illustrates variations of SNR for output (denoised) signals compared to that of input signals.

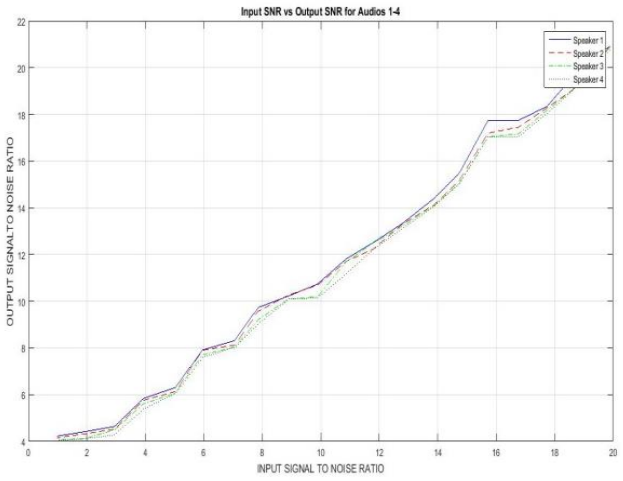

Figure 5. *Input against Output Signal to Noise Ratio.*

Figure 6 illustrates in amplitude versus time as well as power versus frequency of pre-emphasized signal (the description is poorly done. What is the essence of Figure 6 in this paper?

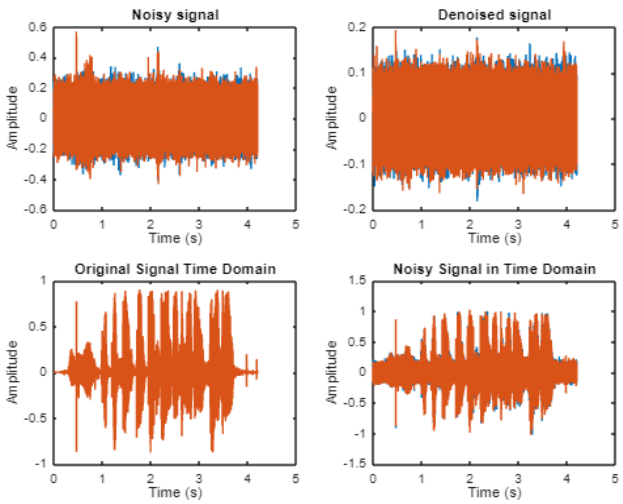

Figure 6. *Pre-emphasized signal in frequency and time domains.*

The windowed frames in time and frequency domain are shown Figure 7. The highest amplitude was 0.2 and highest power was -40dB and lowest -200dB. Conversely, the lowest power level was significantly weaker at -200dB, suggesting that there were parts of the signal that had very little energy. This wide range between the highest and lowest power levels implied that the signal had a high dynamic range or that there were periods of low activity interspersed with peaks of higher intensity.

The dynamic range was useful to understand the efficiency of processing audio signal and to detect any potential issues with signal integrity. The measurements could help in assessing the loudness and clarity of the recorded sound.

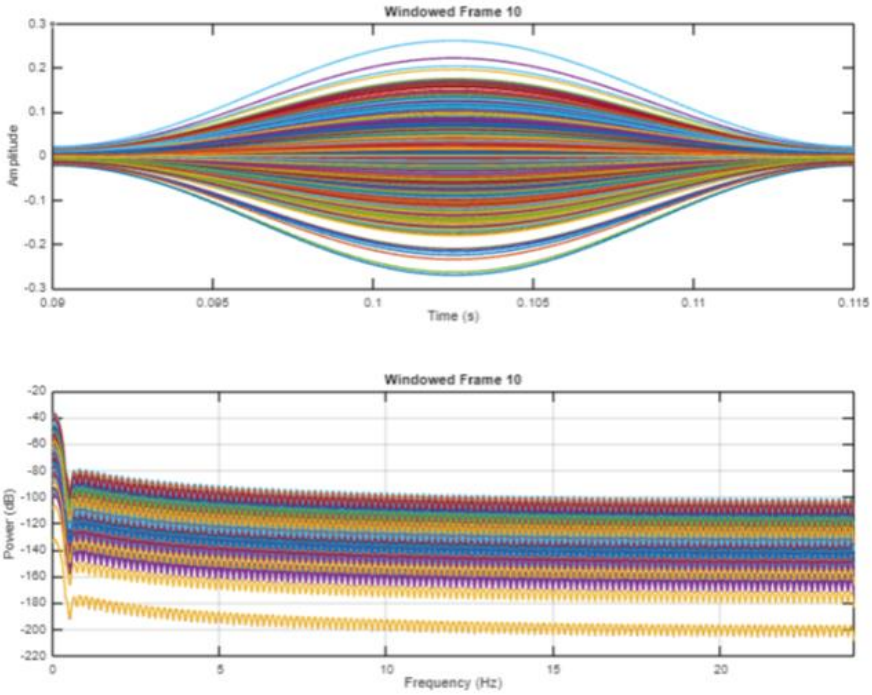

Figure 7. *Windowed frames in time and frequency domains.*

Figure 8 illustrates Mel frequency filter bank response for the four audio signals recorded in the National Assembly of Kenya. At low frequencies, the response intensity was high. Increase in frequency of the response reduced intensity gradually. The high response intensity at low frequencies aligns with the greater sensitivity of human ear to lower frequencies. As the frequency increased, the intensity diminished, which is typical due to the logarithmic nature of the Mel scale that mimics response of the human auditory system. The spacing of the lines in the filter bank response reflected the property Mel scale of having a higher resolution at lower frequencies and a lower resolution at higher frequencies. The observed pattern in the filter bank response was indicative of the role of filter bank in emphasizing the formant structure of speech important for distinguishing phonetic elements and understanding spoken language.

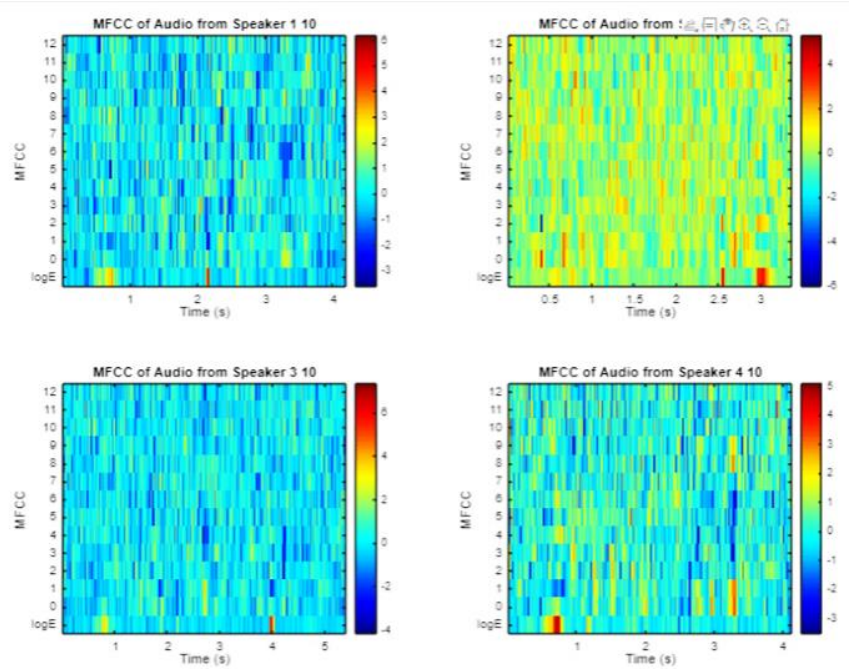

Figure 8. *Mel filter bank response.*

The Mel filter bank illustrated in Figure 9 shows while the filter index increased from 1 to 20 with increase in frequency index, the spectrum reduced from 1 to 0. The gradual decline of the spectrum from 1 to 0 across the filters reflects the logarithmic nature of the Mel scale, which is more aligned with human auditory perception than the linear frequency scales.

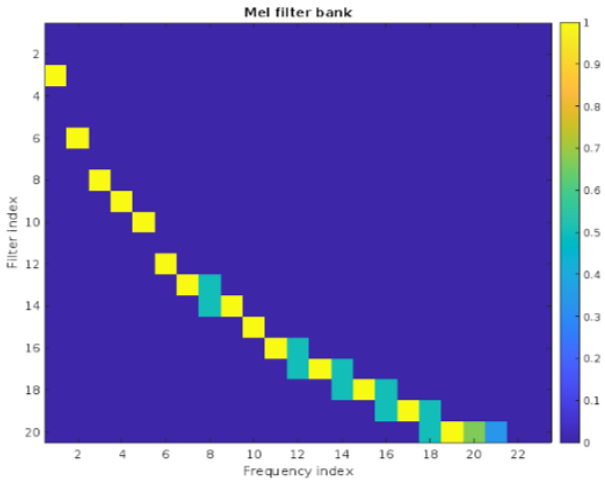

Figure 9. *Mel filter bank.*

Results in Figure 10 illustrates MFCCs for four audio signals with respect to time. MFCC spectral range of Audio 1 was mainly between -3 and 1, for Audio 2 between -1 and 2, Audio 3 were mainly between -4 and 1 and Audio 4 were mainly between -1 and 3. Audio 1 and Audio 3 had a lower range, potentially indicating a deeper voice quality, while Audio 2 and Audio 4 exhibited a higher ranges corresponding to a lighter or more variable voice.

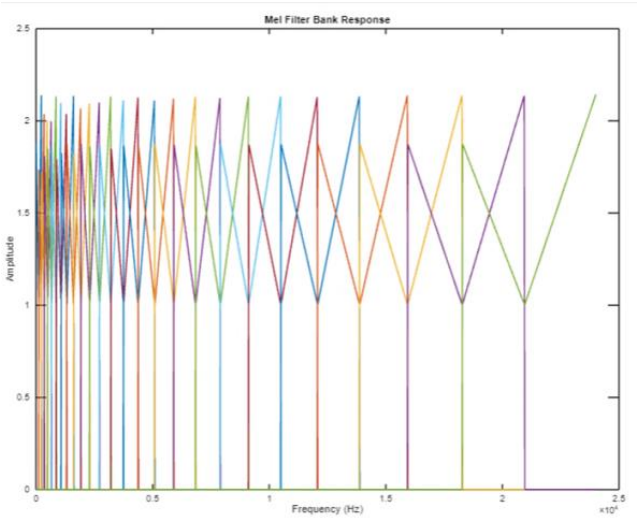

Figure 10. *MFCC of four audio signals.*

The histograms in Figure 11 illustrate the MFCCs for four audio signals with respect to coefficient values. All the four audio signals had mean of probability density below zero coefficient value. This means on average, the energy in the frequency bands represented by the coefficients was less than the logarithm of the overall energy. This also means the voices had less energy in the frequency bands that were crucial for distinguishing between different speakers. The histograms show that audio signal from Speaker 2 had mean coefficient value of -0.364421148 and a mean probability density value

of 1.052896054 which was the highest probability density among the four speakers. This suggests that the audio characteristics of Speaker 2 are more consistently represented across the sample, potentially making the speech pattern the most distinct and recognizable. On the other hand, audio signal of Speaker 3 had the highest mean coefficient value of -0.259147726 but the lowest mean probability density value of 0.81563408. This was the least consistent speech pattern, which could pose challenges for recognition. Speakers 1 and 4 have almost similar mean coefficient values, but Speaker 4 has a slightly higher probability density, indicating a slightly more consistent pattern than Speaker 1. These results suggest that while all the four speakers could be differentiated based on their MFCCs, Speaker 2 stood out as having the most distinct and consistent speech pattern, which was advantageous for the proposed speaker recognition system.

Results in Figure 12 illustrate variations of performance indicators for 50 audio samples tested with the three algorithms MFCC-RCNN-HCO, LPC-CNN and MFCC-CNN. The performance indicators were Equal Error Rate (EER), False Match Rate (FMR), False Rejection Rate (FRR) and True Match Rate (TMR). The proposed algorithm demonstrates a robust performance across indicators. The results show that for the graph of variation between FRR against FMR, the proposed algorithm had the lowest FRR, followed by LPC-CNN and lastly, MFCC-CNN. As FRR decreased, there was increase in FMR for each of the algorithms. The proposed algorithm maintains the lowest FRR which is essential identifying speakers. This is particularly important in the context of the National Assembly, where incorrect rejection of a speaker could lead to incorrect Hansard transcription of speech. Conversely, the increase in FMR as FRR decreased was an expected trade-off in the speaker recognition system, reflecting a balance between correct identification and convenience in transcription. Relationship between TMR and EER show that as EER varied, the proposed algorithm had the highest TMR of all the three followed by LPC-CNN and then MFCC-CNN. TMR of all the algorithms increased with increase in EER. While the proposed algorithm achieves the lowest EER at lower FMR levels, its performance dips as FMR exceeds 40%. The proposed algorithm's ability to keep FMR increases at bay, up to a 40% threshold, was indicative of its sophisticated design, which incorporated advanced features extraction and classification techniques inherent to MFCC-RCNN-HCO. This could imply that the algorithm is highly accurate under typical conditions, though its reliability is low but reasonable in scenarios with higher noise levels or poor audio quality, which are plausible in lively parliamentary settings. The relationship between TMR and EER further underscores the algorithm's initial accuracy, but also highlights a potential area for improvement in maintaining performance consistency across varying error rates. The increase in TMR with EER suggests that the algorithm is more lenient at higher error rates, which could be beneficial in some scenarios but might also introduce vulnerabilities. Graph of the relationship between TMR and FMR illustrates that TMR for the proposed algorithm was the highest across all values of FMR. There was increase in TMR as FMR increased. Superior TMR of the proposed algorithm suggests that it had the highest probability of correctly identifying speakers from audio signals.

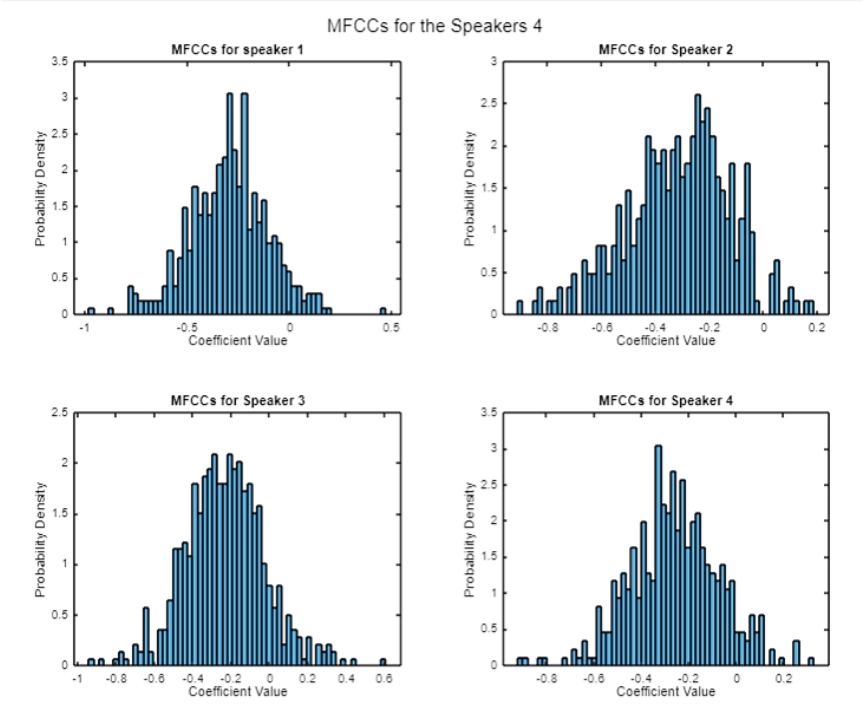

Figure 11. *Histograms of the MFCCs for the four audios.*

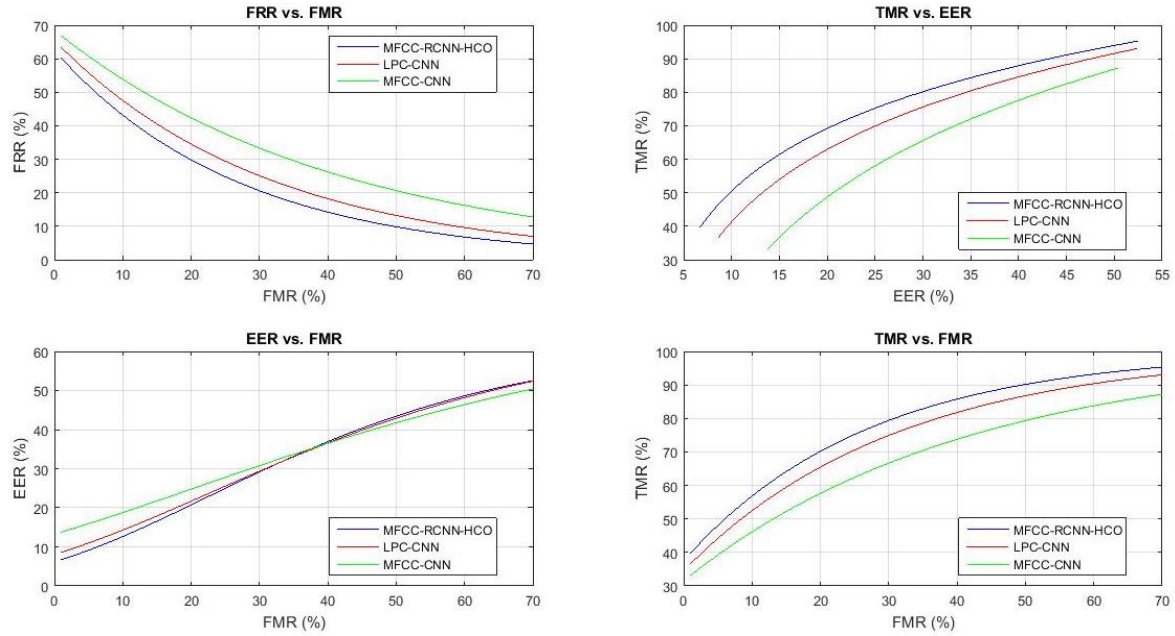

Figure 12. Performance indicators of the algorithms.

5. Conclusions

The proposed algorithm outperformed the others in maintaining a lowest EER, FMR, FRR and highest TMR. The algorithm consistently had the highest TMR, even as FMR rose, indicating a robust ability to accurately identify speakers from audio signals. A low FRR ensures reliable Hansard transcriptions, while the trade-off with increasing FMR is a common aspect of such systems. The algorithm excels in achieving a low EER, indicating high accuracy. The proposed algorithm was also robust under typical parliamentary conditions that have high noise levels.

Abbreviations

SNR	Signal-to-Noise Ratio
FFT	Fast Fourier Transform
MFCC	Mel-Frequency Cepstral Coefficient
DC	Direct Current
DFT	Discrete Fourier Transform
R-CNN/RCNN	Region-based Convolutional Neural Network
HCO	Host-Cuckoo Optimization
CNN	Convolutional Neural Network
RNN	Recursive Neural Network
SVM	Support Vector Machine
EER	Equal Error Rate
FMR	False Match Rate
FRR	False Rejection Rate
TMR	True Match Rate
LDA	Linear Discriminant Analysis
PCA	Principal Component Analysis
GMM	Gaussian Mixture Models
HMM	Hidden Markov Models
LPC	Linear Predictive Coefficients
CMS	Cepstral Mean Subtraction
CMVN	Cepstral Mean and Variance Normalization
RASTA	Relative Spectral Transform Analysis

Author Contributions

Stephen Nyakuti Otenyi: Conceptualization, Data Curation, Software, Writing

Livingstone Ngoo: Supervision, Project Administration, Validation

Henry Kiragu: Supervision, Validation, Formal Analysis, Software, Data Curation

Funding

This work is not supported by any external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] R. M. Hanifa, K. Isa, and S. Mohamad, "A review on speaker recognition: Technology and challenges," *Comput. Electr. Eng.*, vol. 90, p. 107005, 2021. https://doi.org/10.1016/j.compeleceng.2021.107005

[2] M. Jakubec, E. Lieskovska, and R. Jarina, "An Overview of Automatic Speaker Recognition in Adverse Acoustic Environment," presented at the 2020 18th International Conference on Emerging eLearning Technologies and Applications (ICETA), IEEE, 2020, pp. 211–218. https://doi.org/10.1109/iceta51985.2020.9379245

[3] D. Hershcovich *et al.*, "Challenges and strategies in cross-cultural NLP," *ArXiv Prepr. ArXiv220310020*, 2022. https://doi.org/10.18653/v1/2022.acl-long.482

[4] V. N. Ngoni, "English–Bukusu Automatic Machine Translation for Digital Services Inclusion in E-governance," 2022.

[5] N. O. Ogechi, "On language rights in Kenya," *Nord. J. Afr. Stud.*, vol. 12, no. 3, pp. 19–19, 2003.

[6] S. S. Tirumala, S. R. Shahamiri, A. S. Garhwal, and R. Wang, "Speaker identification features extraction methods: A systematic review," *Expert Syst. Appl.*, vol. 90, pp. 250–271, 2017. https://doi.org/10.1016/j.eswa.2017.08.015

[7] National Assembly of Kenya, "Standing Orders." National Assembly of Kenya, 2013. [Online]. Available: http://www.parliament.go.ke/sites/default/files/2022-08/National%20Assembly%20Standing%20Orders%20-%206th%20Edition,%202022_0.pdf

[8] R. Jahangir *et al.*, "Text-independent speaker identification through feature fusion and deep neural network," *IEEE Access*, vol. 8, pp. 32187–32202, 2020. https://doi.org/10.1109/access.2020.2973541

[9] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," *Appl. Acoust.*, vol. 158, p. 107020, 2020. https://doi.org/10.1016/j.apacoust.2019.107020

[10] J. V. E. López, "Adaptation of Speaker and Speech Recognition Methods for theAutomatic Screening of Speech Disorders Using Machine Learning," 2023. https://doi.org/10.14232/phd.11491

[11] M. M. Kabir, M. F. Mridha, J. Shin, I. Jahan, and A. Q. Ohi, "A survey of speaker recognition: Fundamental theories, recognition methods and opportunities," *IEEE Access*, vol. 9, pp. 79236–79263, 2021. https://doi.org/10.1109/access.2021.3084299

[12] S. J. Jainar, P. L. Sale, and B. Nagaraja, "VAD, feature extraction and modelling techniques for speaker recognition: a review," *Int. J. Signal Imaging Syst. Eng.*, vol. 12, no. 1–2, pp. 1–18, 2020. https://doi.org/10.1504/ijsise.2020.10036128

[13] S. K. Sarangi and G. Saha, "Improved speech-signal based frequency warping scale for cepstral feature in robust speaker verification system," *J. Signal Process. Syst.*, vol. 92, pp. 679–692, 2020. https://doi.org/10.1007/s11265-020-01517-2

[14] S. Bharadwaj and P. B. Acharjee, "Analysis of Prosodic features for the degree of emotions of an Assamese Emotional Speech," presented at the 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), IEEE, 2020, pp. 1441–1452. https://doi.org/10.1109/iceca49313.2020.9297453

[15] C. Quam and S. C. Creel, "Impacts of acoustic ‐ phonetic variability on perceptual development for spoken language: A review," *Wiley Interdiscip. Rev. Cogn. Sci.*, vol. 12, no. 5, p. e1558, 2021. https://doi.org/10.1002/wcs.1558

[16] S. Ali, S. Tanweer, S. S. Khalid, and N. Rao, "Mel frequency cepstral coefficient: a review," *ICIDSSD*, 2020. https://doi.org/10.4108/eai.27-2-2020.2303173

[17] S. Pangaonkar and A. Panat, "A Review of Various Techniques Related to Feature Extraction and Classification for Speech Signal Analysis," presented at the ICDSMLA 2019: Proceedings of the 1st International Conference on Data Science, Machine Learning and Applications, Springer, 2020, pp. 534–549. https://doi.org/10.1007/978-981-15-1420-3_57

[18] B. A. Aicha and F. Kacem, "Conventional Machine Learning and Feature Engineering for Vocal Fold Precancerous Lesions Detection Using Acoustic Features," *Circuits Syst. Signal Process.*, pp. 1–33, 2023. https://doi.org/10.1007/s00034-023-02551-8

[19] K. Jagadeeshwar, T. Sreenivasarao, P. Pulicherla, K. Satyanarayana, K. M. Lakshmi, and P. M. Kumar, "ASERNet: Automatic speech emotion recognition system using MFCC-based LPC approach with deep learning CNN," *Int. J. Model. Simul. Sci. Comput.*, vol. 14, no. 04, p. 2341029, 2023. https://doi.org/10.1142/s1793962323410295

[20] M. Ramashini, P. E. Abas, K. Mohanchandra, and L. C. De Silva, "Robust cepstral feature for bird sound classification," *Int. J. Electr. Comput. Eng.*, vol. 12, no. 2, p. 1477, 2022. https://doi.org/10.11591/ijece.v12i2.pp1477-1487

[21] G. Vanderreydt and K. Demuynck, "A Novel Channel estimate for noise robust speech recognition," *Comput. Speech Lang.*, p. 101598, 2023. https://doi.org/10.1016/j.csl.2023.101598

[22] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, 1994. https://doi.org/10.1109/89.326616

[23] N. N. Alrouqi, "Additive Noise Subtraction for Environmental Noise in Speech Recognition," 2021.

[24] S. Alharbi *et al.*, "Automatic speech recognition: Systematic literature review," *IEEE Access*, vol. 9, pp. 131858–131876, 2021. https://doi.org/10.1109/access.2021.3112535

[25] Wen-Jie Song, Chen Chen, Tian-Yang Sun, and Wei Wang, "A Robust Equalization Feature for Language Recognition," *J. Inf. Sci. Eng.*, vol. 36, no. 3, pp. 561–576, May 2020, https://doi.org/10.6688/JISE.202005_36(3).0006

[26] G. Manikandan and S. Abirami, "Feature Selection Is Important: State-of-the-Art Methods and Application Domains of Feature Selection on High-Dimensional Data," in *Applications in Ubiquitous Computing*, R. Kumar and S. Paiva, Eds., in EAI/Springer Innovations in Communication and Computing. Cham: Springer International Publishing, 2021, pp. 177–196. https://doi.org/10.1007/978-3-030-35280-6_9

[27] O. Ghahabi, P. Safari, and J. Hernando, "Deep Learning in Speaker Recognition," in *Development and Analysis of Deep Learning Architectures*, W. Pedrycz and S.-M. Chen, Eds., in Studies in Computational Intelligence, Cham: Springer International Publishing, 2020, pp. 145–169. https://doi.org/10.1007/978-3-030-31764-5_6

[28] Y. Wang, L. Zheng, Y. Gao, and S. Li, "Vibration Signal Extraction Based on FFT and Least Square Method," *IEEE Access*, vol. 8, pp. 224092–224107, 2020, https://doi.org/10.1109/ACCESS.2020.3044149

[29] J. Agrawal, M. Gupta, and H. Garg, "A review on speech separation in cocktail party environment: challenges and approaches," *Multimed. Tools Appl.*, vol. 82, no. 20, pp. 31035–31067, Aug. 2023, https://doi.org/10.1007/s11042-023-14649-x

[30] P. Golik, "Data-driven deep modeling and training for automatic speech recognition," 2020.

[31] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix, and T. Nakatani, "Far-Field Automatic Speech Recognition," *Proc. IEEE*, vol. 109, no. 2, pp. 124–148, Feb. 2021, https://doi.org/10.1109/JPROC.2020.3018668

[32] A. Lauraitis, R. Maskeliūnas, R. Damaševičius, and T. Krilavičius, "Detection of Speech Impairments Using Cepstrum, Auditory Spectrogram and Wavelet Time Scattering Domain Features," *IEEE Access*, vol. 8, pp. 96162–96172, 2020, https://doi.org/10.1109/ACCESS.2020.2995737

[33] S. A. El-Moneim *et al.*, "Speaker recognition based on pre-processing approaches," *Int. J. Speech Technol.*, vol. 23, no. 2, pp. 435–442, Jun. 2020, https://doi.org/10.1007/s10772-019-09659-w

[34] N. Chen and S. Fu, "Uncertainty quantification of nonlinear Lagrangian data assimilation using linear stochastic forecast models," *Phys. Nonlinear Phenom.*, vol. 452, p. 133784, Oct. 2023, https://doi.org/10.1016/j.physd.2023.133784

[35] A. P. Fellows, M. T. L. Casford, and P. B. Davies, "Spectral Analysis and Deconvolution of the Amide I Band of Proteins Presenting with High-Frequency Noise and Baseline Shifts," *Appl. Spectrosc.*, vol. 74, no. 5, pp. 597–615, May 2020, https://doi.org/10.1177/0003702819898536

[36] N. Saleem, J. Gao, M. I. Khattak, H. T. Rauf, S. Kadry, and M. Shafi, "DeepResGRU: Residual gated recurrent neural network-augmented Kalman filtering for speech enhancement and recognition," *Knowl.-Based Syst.*, vol. 238, p. 107914, Feb. 2022, https://doi.org/10.1016/j.knosys.2021.107914

[37] P. Bansal, S. A. Imam, and R. Bharti, "Speaker recognition using MFCC, shifted MFCC with vector quantization and fuzzy," presented at the 2015 International Conference on Soft Computing Techniques and Implementations (ICSCTI), IEEE, 2015, pp. 41–44. https://doi.org/10.1109/icscti.2015.7489535

[38] L.-M. Dogariu, J. Benesty, C. Paleologu, and S. Ciochină, "An Insightful Overview of the Wiener Filter for System Identification," *Appl. Sci.*, vol. 11, no. 17, Art. no. 17, Jan. 2021, https://doi.org/10.3390/app11177774

[39] Y. Zouhir, M. Zarka, and K. Ouni, "Power Normalized Gammachirp Cepstral (PNGC) coefficients-based approach for robust speaker recognition," *Appl. Acoust.*, vol. 205, p. 109272, Mar. 2023, https://doi.org/10.1016/j.apacoust.2023.109272

[40] A. Ahmed, Y. Serrestou, K. Raoof, and J.-F. Diouris, "Empirical Mode Decomposition-Based Feature Extraction for Environmental Sound Classification," *Sensors*, vol. 22, no. 20, Art. no. 20, Jan. 2022, https://doi.org/10.3390/s22207717

[41] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Netw.*, vol. 140, pp. 65–99, 2021. https://doi.org/10.1016/j.neunet.2021.03.004

[42] Gaurav, S. Bhardwaj, and R. Agarwal, "An efficient speaker identification framework based on Mask R-CNN classifier parameter optimized using hosted cuckoo optimization (HCO)," *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 10, pp. 13613–13625, Oct. 2023, https://doi.org/10.1007/s12652-022-03828-7

[43] Z. Touati-Hamad and M. R. Laouar, "Enhancing Education Decision-Making with Deep Learning for Arabic Spoken Digit Recognition," p. 4321405 Bytes, 2023, https://doi.org/10.6084/M9.FIGSHARE.24902382.V3

[44] W.-C. Lin and C. Busso, "Chunk-Level Speech Emotion Recognition: A General Framework of Sequence-to-One Dynamic Temporal Modeling," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1215–1227, Apr. 2023, https://doi.org/10.1109/TAFFC.2021.3083821

[45] S. Agarwal, J. O. D. Terrail, and F. Jurie, "Recent Advances in Object Detection in the Age of Deep Convolutional Neural Networks." arXiv, Aug. 20, 2019. https://doi.org/10.48550/arXiv.1809.03193

[46] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal Intelligence: Representation Learning, Information Fusion, and Applications," *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 3, pp. 478–493, Mar. 2020, https://doi.org/10.1109/JSTSP.2020.2987728

[47] S. Hourri, N. S. Nikolov, and J. Kharroubi, "Convolutional neural network vectors for speaker recognition," *Int. J. Speech Technol.*, vol. 24, no. 2, pp. 389–400, Jun. 2021, https://doi.org/10.1007/s10772-021-09795-2

[48] G. Hu, Z. Zhang, A. Armaou, and Z. Yan, "Robust extended Kalman filter based state estimation for nonlinear dynamic processes with measurements corrupted by gross errors," *J. Taiwan Inst. Chem. Eng.*, vol. 106, pp. 20–33, Jan. 2020, https://doi.org/10.1016/j.jtice.2019.10.015

[49] O. Deshpande, K. Solanki, S. P. Suribhatla, S. Zaveri, and L. Ghodasara, "Simulating the DFT Algorithm for Audio Processing," *ArXiv Prepr. ArXiv210502820*, 2021.

[50] M. Awais, Md. T. Bin Iqbal, and S.-H. Bae, "Revisiting Internal Covariate Shift for Batch Normalization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 5082–5092, Nov. 2021, https://doi.org/10.1109/TNNLS.2020.3026784

[51] H. Moayed and E. G. Mansoori, "Improving Regularization in Deep Neural Networks by Co-adaptation Trace Detection," *Neural Process. Lett.*, vol. 55, no. 6, pp. 7985–7997, Dec. 2023, https://doi.org/10.1007/s11063-023-11293-2

[52] R. Gadagkar, "How to Design Experiments in Animal Behaviour," *Resonance*, vol. 25, no. 10, pp. 1419–1455, Oct. 2020, https://doi.org/10.1007/s12045-020-1061-4

Biography

Stephen Nyakuti is a masters Student at the Multimedia University of Kenya undertaking Master Degree Course in Multimedia and Communication Engineering. He has degree in Telecommunication Engineering and Information Technology and Post Graduate Diploma in Mass Communication. He has Ten years of working experience in the Engineering and ICT fields.

Livingstone Ngoo: Prof. Dr-Eng. Livingstone M. H. Ngoo is a multifaceted professional with a distinguished career in electrical engineering, university administration, research, and education. Currently serving as the Acting Deputy Vice Chancellor for Academic Affairs, Research, and Innovation at Multimedia University of Kenya (MMU), Dr. Ngoo brings a wealth of experience and expertise to his role. Dr. Ngoo holds a Ph.D. in Electrical Power Systems, specializing in automation using fuzzy logic techniques and Master of Science Degree in (Control Engineering). His extensive experience encompasses designing, supervising, and commissioning electrical works and generators for diverse institutions.

Henry Kiragu has over twenty-five (25) years of teaching experience in the fields of Electronics and Telecommunication Engineering at undergraduate and graduate levels. He has supervised many research projects and theses for undergraduate as well as postgraduate level students in addition to publishing numerous papers in peer-reviewed journals and conference proceedings. Kiragu holds a Doctor of Philosophy (PhD) in Electrical and Electronics Engineering (2020) and a Master of Science (MSc) in Electrical and Electronics Engineering (2013) degrees from the University of Nairobi, Kenya. He is also a holder of a Bachelor of Technology (BTech) in Electrical and Communications Engineering degree (1994) from Moi University in Kenya. Currently, he works as a Senior Lecturer of Electronics and Telecommunication Engineering at the Multimedia University of Kenya. He is a member of the Engineers Board of Kenya (EBK) as well as the Institute Of Electrical and Electronics Engineers (IEEE).

Research Field

Stephen Nyakuti Otenyi: Telecommunications, Communication, Information Technology

Livingstone Ngoo: Telecommunications, Control Engineering, Power Systems, Electrical and Electronics,

Henry Kiragu: Telecommunications, Electricals and Electronics