






## Real-Time Lie-Speech Determination Using Voice-Stress Technology

Fadi K. Al-Dhafer <sup>1</sup>, Duraid Y. Mohammed <sup>2</sup>, Mohammed Khalaf <sup>3</sup>,  
Khamis A. Al-Karawi <sup>4</sup>, Mohammad Sarfraz <sup>5</sup>, Muhammad Mazin Al  
Maathidi <sup>6</sup>\*

<sup>1,2</sup> Computer Engineering Department, College of Engineering, Al-Iraqia University, Baghdad, Iraq

<sup>3</sup> General Directorate of Education Anbar, Ramadi 31001, Iraq

<sup>3</sup> Computer Science Department, Al-Maarif University College, Ramadi 31001, Iraq

<sup>4</sup> University of Diyala, Iraq

<sup>5</sup> Electrical Engineering Department, Aligarh Muslim University, Aligarh, India

<sup>6</sup> Computer Science Department, University of Bahrain, Bahrain

\*Corresponding Author: Fadi K. Al-Dhafer

DOI: <https://doi.org/10.52866/ijcsm.2024.05.02.008>

Received September 2023; Accepted January 2024; Available online March 2024

**ABSTRACT:** Lie detection has gained importance and is now extremely significant in a variety of fields. It plays an important role in several domains, including law enforcement, criminal investigations, national security, workplace ethics, and personal relationships. As advances in lie detection continue to develop, real-time approaches such as voice stress technology have emerged as a feasible alternative to traditional methods such as polygraph testing. Polygraph testing, a historical and generally established approach, may be enhanced or replaced by these revolutionary real-time techniques. Traditional lie detection procedures, such as polygraph testing, have been challenged for their lack of reliability and validity. Newer techniques, such as brain imaging and machine learning, might offer better outcomes, although they are still in their early phases and require additional testing. This project intends to explore a deception-detection module based on sophisticated speech-stress analysis techniques that might be applied in a real-time deception system. The purpose is to study stress and other articulation cues in voice patterns, to establish their precision and reliability in detecting deceit, by building upon previous knowledge and applying state-of-the-art architecture. The performance and accuracy of the system and its audio aspects will be thoroughly analyzed. The ultimate purpose is to contribute to the advancement of more accurate and reliable lie-detection systems, by addressing the limitations of old techniques and proposing practical solutions for varied applications. This paper proposes an efficient feature-selection strategy, which uses random forest (RF) to select only the significant features for training when a real-life trial dataset consisting of audio files is employed. Next, utilizing the RF as a classifier, an accuracy of 88% is reached through comprehensive evaluation, thereby confirming its reliability and precision for lie-detection in real-time scenarios.

**Keywords:** lie speech, voice stress, real time, random forest, machine learning, deception detection.

## 1. INTRODUCTION

### 1.1 Literature Review

Lie detection is a growing field of study, which is emerging as increasingly important across the fields of law, security, work-ethics, and relationships. Detection of deception is crucial to solving crimes, ensuring public safety, and maintaining ethical standards. In navigating this constantly changing field, it is crucial to clearly outline the basic idea of lie detection. Detecting lies involves identifying deceptive behavior, a crucial process for making informed decisions in various situations. In the past, lie detection relied chiefly on methods such as polygraph tests. However, there exist concerns about the reliability and validity of these tests, leading researchers to look for better methods [1].

Given the recent advances in technology and continued research efforts, new methods such as brain imaging [2], machine learning [3], and voice stress analysis [4] have come to the forefront. These advancements represent a significant change in how we approach the comprehension of and dealing with deception, thereby presenting exciting opportunities for more successful lie detection strategies. Some of the most commonly used machine learning techniques in lie detection research include support vector machines, decision trees, neural networks, and random forests (RFs). These techniques are used to analyze features of speech or physiological responses, such as heart rate or skin conductance, to identify deception patterns. These new methods offer the potential for more accurate and non-invasive lie detection. However, while these advanced technologies provide novel insights into lie detection, it is crucial to recognize their inherent ethical considerations. The potential intrusion into individual privacy, especially

since these methods explore cognitive processes or facial expressions, raises valid concerns. In environments such as banking, wherein the expectation of privacy is paramount, the use of certain technologies, such as electroencephalogram for lie detection, may be impractical or even unacceptable due to perceived intrusiveness. Despite these challenges, lie-detection research continues to advance, with researchers exploring innovative approaches to identifying deception in real-time. By leveraging recent advances in technology and drawing upon insights from fields such as psychology, neuroscience, and computer science, lie-detection researchers are working to develop more accurate and reliable methods that can be applied in a range of contexts. The potential benefits of such methods are significant, not only in law enforcement and criminal investigation, but also in other areas where honesty and trust are essential components of successful relationships and outcomes [5].

Fathima and Shajee proposed an efficient lie-detection technique based on the non-invasive method, which involved recording of the subject's speech utterances alone. Discriminative and meaningful features were extracted from the speech; subsequently, classifiers were built using support vector machine (SVM) to discriminate between truth and lie. However, this study uses Mel-frequency cepstral coefficients (MFCC) and mean MFCC only to distinguish between truth and lie speech [5]. Several studies have explored the use of machine learning techniques for lie detection. One study used a recurrent neural network in long-short term memory (LSTM) architecture and generated a database of audio recordings for neural network training. The study found that the best model achieved a precision of 72.5% for voice stress analysis [4].

Another study proposed a deep neural network (DNN) system for speech detection using various DNN types, with the best performance being obtained for convolutional neural networks (CNNs). The proposed system achieved up to 99.13% accuracy for the CENSREC-1-C datasets and 97.60% for the TIMIT dataset [6]. Again, another study trained machine learning models and a sequential neural network using acoustic features in speech for lie detection. A majority-voting ensemble learning classifier constructed through the combined use of a gradient boosting classifier (GBC), an SVM, and a stochastic gradient descent (SGD) proved to be the most effective one. The maximum accuracy achieved by this model was 55.8%. Finally, one study presented an automatic deception-detection framework achieved by integrating prior domain knowledge in deceptive-behavior understanding. The proposed model reached state-of-the-art performance on the Daily Deceptive Dialogues corpus of Mandarin (DDDM) database, with an 80.61% unweighted accuracy recall in deception recognition [7].

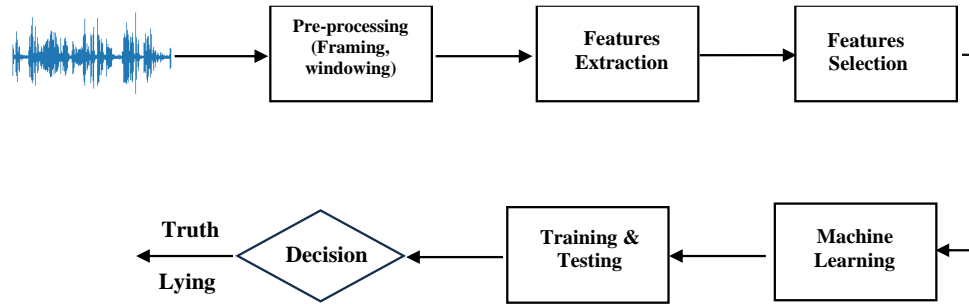
Sinead and Muhammad conducted a study to examine the differences in speech patterns between truthful and deceptive speakers, focusing on human speech perception. The study recorded a speaker under stress during a police interrogation, capturing two truthful and two deceitful responses at different times of the day. The study then extracted and analyzed four nonlinear speech features, namely the MFCC, delta cepstrum, time-difference cepstrum, Bark band energy, delta energy, and time-difference energy features from audio recordings of all three sessions. The Levenberg-Marquardt and LSTM classification methods were applied to nine training and testing combinations of the cepstrum and spectral energy features extracted from three sessions, to assess the accuracy of deception detection. The researchers also employed principal component analysis to reduce the dimensionality of the extracted features for improved performance. The results showed that the projected principal components of the four types of features significantly improved the accuracy of distinguishing between truthful and deceptive speech patterns. Moreover, with regard to time-difference spectral energy features, the LSTM classification method had the highest recognition rate, when compared to the Levenberg-Marquardt algorithm using other cepstral and spectral features [3].

Hongliang Fu and colleagues proposed an innovative, semi-supervised approach for speech-deception detection, by combining acoustic statistical features with two-dimensional time-frequency features. The method involved establishing a hybrid, semi-supervised neural network, incorporating a semi-supervised autoencoder network (AE) and a mean-teacher network. In the process, static artificial statistical features were fed into the semi-supervised AE to extract robust advanced features. Simultaneously, three-dimensional (3D) Mel-spectrum features were input into the mean-teacher network to capture features rich in time-frequency two-dimensional information. The integration of these features was further enhanced by a consistency regularization method, effectively mitigating overfitting and enhancing the model's generalization ability. Experimental evaluations of a self-built corpus for deception detection revealed that the proposed algorithm achieved a recognition accuracy of 68.62%, thereby surpassing the baseline system by 1.2% and demonstrating notable improvements in detection accuracy [8].

## 1.2 Contribution of the Present Study

Following a comprehensive review of existing literature and the identification of pertinent research gaps, our work introduces a contribution to the field of lie-speech detection, which uses voice stress technology. Our research study proposes an innovative feature-selection strategy that leverages the importance of features through the application of an RF algorithm. This methodology enables the identification and prioritization of significant features crucial for training. By employing RF for feature-selection, our approach ensures that only the most informative features contribute to the training process, thereby optimizing the model's efficiency and reducing learning time. Thus, the focus on selecting only the most important features not only enhances the accuracy of lie detection but also proves to be a highly effective strategy for reducing learning time in real-time applications. This efficiency is crucial for ensuring swift and reliable lie detection in dynamic, real-world settings.

## 2. THE PROPOSED METHOD



**FIGURE 1. Block Diagram of the Proposed Method for the Deception Detection**

### 2.1 Dataset (Real-life Trial Dataset)

The Real-life Trial dataset used in this research has been obtained from [9]; it consists of trial-hearing recordings obtained from public sources. The videos have been carefully selected to be of reasonably good audio-visual quality and portray a single subject with his/her face visible during most of the clip duration. Videos have been collected from trials with varied outcomes: guilty verdict, non-guilty verdict, and exoneration. For guilty verdicts, deceptive clips have been collected from a defendant in trial and truthful videos have been collected from witnesses in the same trial. In some cases, deceptive videos have been collected from a suspect denying a crime committed and truthful clips have been taken from the same suspect while answering questions concerning some facts verified by the police as truthful. In case of the witnesses, testimonies verified by police investigations have been labeled as truthful, whereas testimonies in favor of a guilty suspect have been considered as lying. The dataset consists of 121 videos including 61 deceptive and 60 truthful trial clips. The average length of the videos in the dataset is 28 seconds. Whereas the average length of deceptive and truthful clips is 27.7 seconds and 28.3 seconds, respectively, the data consists of 21 unique female and 35 unique male speakers, with their ages approximately ranging between 16 and 60 years [10]. In this study, only the audio of the above-described dataset has been included, whereas the image frames have not been determined during decision making. The sampling rate (SR) of the audio signal of this dataset is found to be 22050 Hz.

## 2.2 Signal Processing

### 2.2.1 Framing

In audio analysis and processing, dividing the audio signal into short-term frames or windows is common practice. This segmentation allows for the computation of features on each individual frame. This step is particularly important during the feature extraction stage. Typically, the duration of these short-term frames ranges from 10 to 50 milliseconds [11]. In this research study, a frame length of 40 milliseconds has been utilized. This duration was selected because the frame length directly impacts the frequency resolution of the spectrum, considering the sampling frequency. In simpler terms, longer frames provide better frequency resolution but sacrifice time resolution. Conversely, shorter windows offer more detailed representation in the time domain but often result in poorer frequency resolution.

Additionally, the hop size, which determines the overlap between adjacent frames, was set at 50% of the frame size, as illustrated in Figure 2. The choice of hop length has an effect on the degree of overlap between consecutive frames. A smaller hop length increases the overlap, meaning that successive frames share more data. Conversely, a larger hop length decreases the overlap between frames [11].

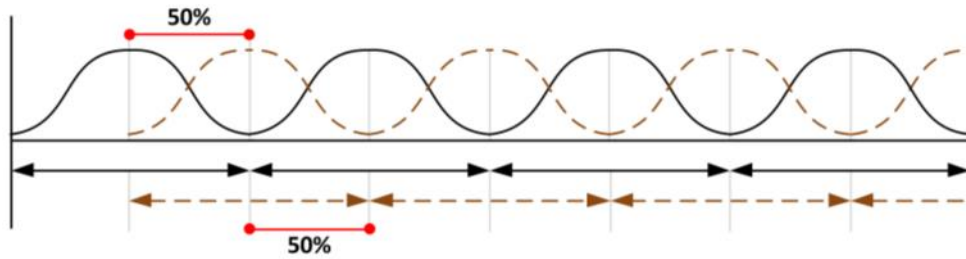
To calculate the frame length and number of frames,  $x(n)$  has been considered as the sampled audio signal,  $S_r$  is the sampling rate,  $n$  represents the time index of the audio signal samples,  $f(n)$  is the sampled audio frame,  $i$  ( $1 \leq i \leq NF$ ) is the index of the frame position in the audio signal,  $NF$  is the total number of frames in the audio signal  $x(n)$ ,  $L$  denotes the length of the frame (total number of samples in each frame), and the hop is the number of time samples between two successive frames [13].

The  $x(n)$  is split into a number of frames ( $f$ ), wherein all frames have the same length. The frame length has been obtained as shown in equation (1) [13].

$$L = \text{duration time in millisecond} \times S_r \quad (1)$$

Also, the number of frames ( $NF$ ) can be calculated according to equation (2):

$$NF = \lfloor (n - L)/hop \rfloor + 1 \tag{2}$$



**FIGURE 2. Overlapped Frame with Hop Size 50% [12]**

### 2.2.2 Windowing

To mitigate the edge effect that arises during framing, a window function is employed. There are several types of window functions available, with the Hamming Window being widely used [13]. The Hamming Window is favored due to its ability to effectively isolate the signal, as it exhibits a rapid decay in its spectrum. While it covers the entire spectrum, it also produces side lobes (higher harmonics) [14]. The signal truncate can be stated mathematically as shown in equation 3:

$$x_w(n) = x(n)w(n) \tag{3}$$

Also, the hamming window is defined using equation 4:

$$\begin{cases} 0.54 - 0.46 \cos\left(\frac{2\pi n}{L-1}\right) & 0 \leq n \leq L-1 \\ 0 & otherwise \end{cases} \tag{4}$$

### 2.3 Feature Extraction

Feature extraction plays a critical role in pattern recognition and machine learning tasks; it aims at deriving a set of informative features from the relevant dataset. These features should capture essential characteristics of the original data. Further, feature extraction can be viewed as a data-rate reduction technique when algorithms rely on a large number of features [11]. In this study, the extracted features have been divided into two categories: those obtained from the time domain and those derived from the frequency domain. Frequency domain features include cepstral features such as MFCC, delta cepstrum, and time-difference cepstrum. Time domain features encompass zero crossings, jitter, centroids, root mean square (RMS), and Energy. These features have been employed to analyze the distinctive patterns differentiating truthful and deceitful speech.

For each frame, audio features were extracted, resulting in a matrix ( $m \times n$ ); here,  $m$  represents the number of frames and  $n$  represents the number of features. The size of this matrix is  $169476 \times 37$  for the combined lie and truth audio features. Averaging of the frames has been performed to enhance the fidelity of the audio signal, by mitigating the influence of noise and other distortions. This approach can enhance the reliability and accuracy of audio analysis techniques, including speech recognition, speaker identification, and deception detection. However, careful consideration should be given to the appropriate method of averaging, as various techniques can impact the resulting feature values and their interpretation.

In this study, the first step involves averaging of all audio samples of the same class to generate one average audio sample. Subsequently, the moving average method is employed. This method entails calculating the mean of the feature values over a sliding time window. In this process, fluctuations in the feature values are smoothed out, providing a more stable representation of the feature trends over time. This calculation method has been applied to all the features described in the following section. Figure 3 illustrates the spectral features extraction.

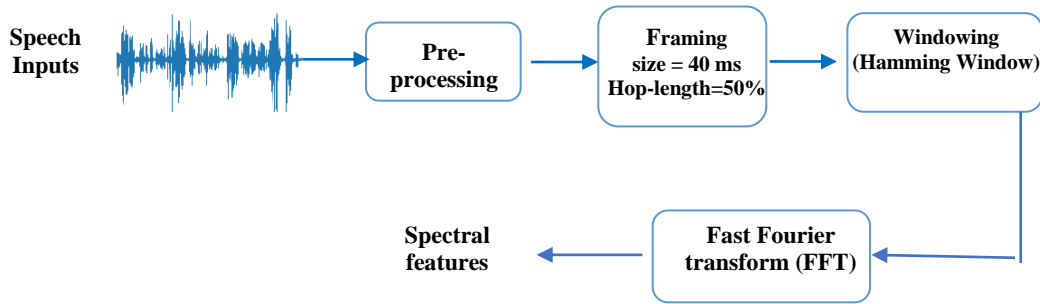


FIGURE 3. Steps in Spectral Features Extraction

### 2.2.2 Time-Domain Audio Features

In general, the time-domain properties of the audio stream are extracted directly from the audio samples. Common examples of such properties include the zero-crossing rate (ZCR), energy, jitter, and RMS.

#### 1- ZCR

ZCR refers to the rate at which the sign of a signal changes within an audio frame. In other words, it represents the number of times the signal transitions from positive to negative and vice versa, divided by the duration of the frame. Higher ZCR values are typically observed in the presence of noisy segments within the signal. The ZCR is mathematically defined as follows [11]:

$$z(i) = \frac{1}{2L} \sum_{n=1}^N |sgn[x_i(n)] - sgn[x_i(n-1)]| \tag{5}$$

where  $n$  = sequence of the audio sample,  $i$  = index of frames,  $N$  = frame length, and  $sgn$  = sign function,

$$sgn[x_i(n)] = \begin{cases} 1, & x_i(n) \geq 0 \\ -1, & x_i(n) < 0 \end{cases} \tag{6}$$

ZCR has been applied in this research study, because all speech-processing tasks, including voice synthesis, speech improvement, and speech recognition employ ZCR. ZCR plays a crucial role in identifying brief and powerful noises; it effectively recognizes small variations in signal amplitude [15].

Figure (4) illustrates the average ZCR calculated as the average of total frames: (a) truth speech and (b) lying speech. It is worth noting that these frames represent all speakers.

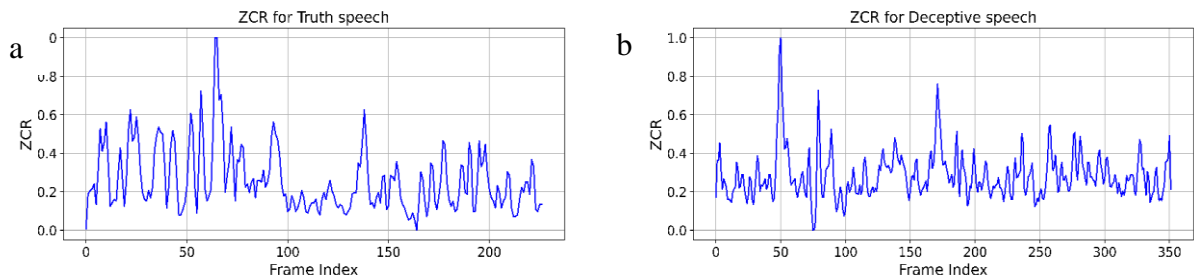


FIGURE 4. ZCR for All Speakers: (a) for Truth Speech and (b) for Lie Speech

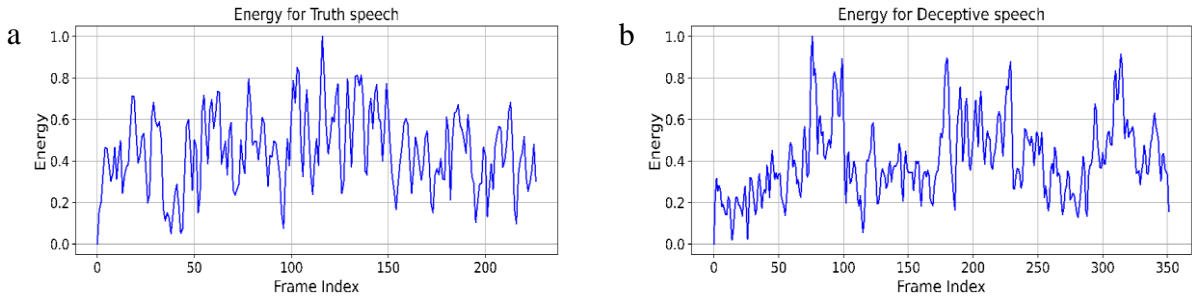
#### 2- Energy

Across subsequent speech frames, short-term energy is predicted to vary significantly, meaning that the energy envelope is predicted to quickly switch between high and low energy states. The reason for this can be attributed to the weak phonemes and brief silences present in speech signals [11]. The energy factor has been applied in this research study, because the energy of the signal is influenced by the amplitude of the wave. It is a loud signal, if the amplitude of the signal is high [14]. This mean can measure changes in volume or intensity over time, and these changes in energy might be a sign of vocal distinctions between truthful and deceptive speech, as shown in Figure (5).

Assuming that  $x_i(n)$ ,  $n = 1 \dots, WL$  is the sequence of audio samples for the  $i$ th frame, and  $WL$  is the frame length. The following formula is used to calculate the short-term energy [11]:

$$E(i) = \sum_{n=1}^L |x_i(n)|^2 \tag{7}$$

Figure (5) illustrates the average Energy for all speakers, calculated as average of total frames: (a) truth speech and (b) lying speech.



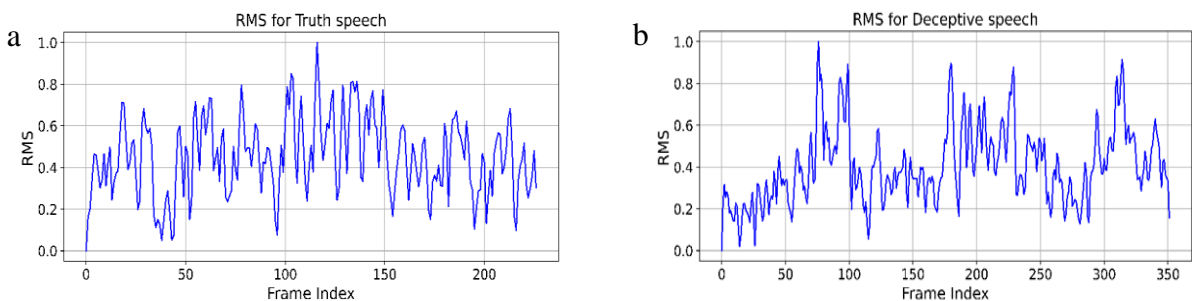
**FIGURE 5.** Energy Feature for All Speakers: (a) for Truth Speech (b) for Lie Speech

### 3- RMS

Typically, the determined RMS of the audio signal is given in decibels. Kenny and Keeping defined it initially (1962). It is a frequently employed feature. According to Tzanetakis et al., frames with silence had a lower RMS compared to frames without silence. RMS is used in this research study, because it is the measure of the loudness of an audio signal [16], which is indicative of changes in stress levels or emotional states. RMS is defined according to Equation 4 [13]:

$$RMS(i) = \sqrt{\frac{1}{L} \sum_{n=1}^L f_i(n)^2} \tag{8}$$

where  $L$  is the length of the frame,  $f_i$  is an audio frame, and  $n$  indicates the time of index of the audio signal sample. Figure (6) illustrates the average RMS calculated on the basis of the following: average of total frames (a) truth speech and (b) lying speech.

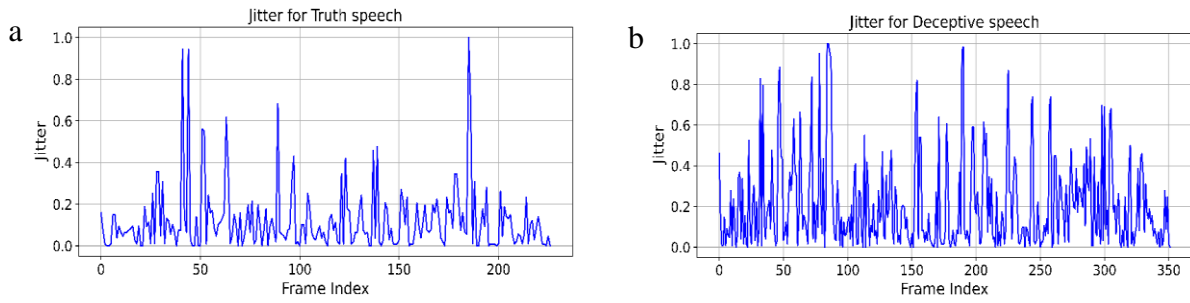


**FIGURE 6.** RMS Feature for All Speakers: (a) for Truth Speech (b) for Lie Speech

### 4- Jitter

Involuntary fundamental frequency variations over a brief period cause jitter, also known as micro-tremors. To put it another way, jitter is an irregularity or oscillation in the voice pitch [4]. It is used in this research study, because it relates to voice quality and can be used to assess the stability of the vocal folds. Jitter changes might be a sign of stress or anxiety, which could be related to lying, as shown in figure (7).

Figure (7) illustrates the average Jitter, calculated as an average of total frames: (a) truth speech and (b) lying speech. It is worth noting that these frames represent all speakers.



**FIGURE 7.** Jitter Feature for All Speakers: (a) for Truth Speech and (b) for Lie Speech

### 2.2.3 Frequency-Domain Audio Features

The discrete Fourier transform (DFT) is frequently employed in audio signal analysis, because it gives a convenient representation of the distribution of the frequency content of sounds, that is, of the sound spectrum. We will now examine a few frequently-utilized audio features, which are based on the DFT of the audio stream. Features of this type are also called frequency-domain (or spectral) audio features, and these include the following: spectral centroids ( $C_i$ ), spectral entropy (SE), MFCCs, chroma vector, pitch, and spectral roll-off [11].

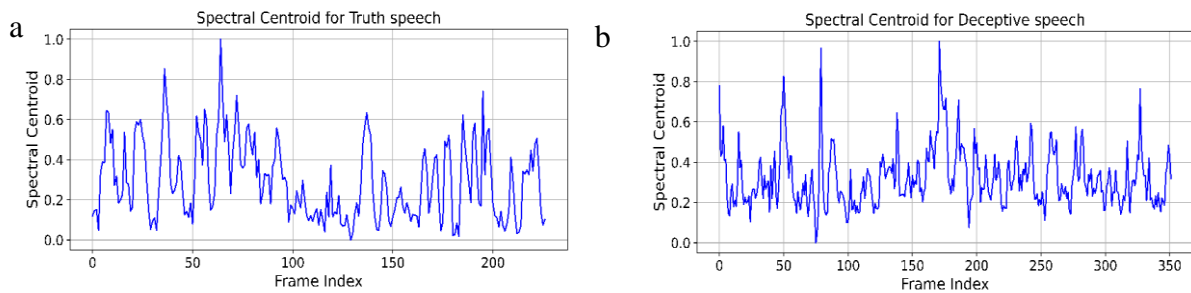
#### 1- $C_i$

$C_i$  is the center of ‘gravity’ of the spectrum: higher values are seen to relate to louder sounds in the  $C_i$ .  $C_i$  provides information about the distribution of energy across different frequencies in the voice signal [11]. The fact that the centroid can detect changes in pitch or spectral variation can indicate emotional or physiological changes. The value of  $C_i$ , for each frame, is defined as follows:

$$C_i = \frac{\sum_{k=1}^{wf_L} k X_i(k)}{\sum_{k=1}^{wf_L} X_i(k)} \quad (9)$$

where  $wf_L$  indicates the number of coefficients used in the computation, and  $X_i(k)$ ,  $k = 1, \dots$  is the magnitude of the DFT coefficients of the  $i^{th}$  audio frame [11].

Figure (8) illustrates the average  $C_i$ , calculated as the average of total frames: (a) truth speech and (b) lying speech. It is worth noting that these average frames represent all speakers.



**FIGURE 8.**  $C_i$  Feature for All Speakers: (a) for Truth Speech (b) for Lie Speech

#### 2- Spectral Entropy

SE has been used in this research study, because its most popular feature in automated speech recognition is speech detection: to distinguish between clean and noisy speech, where clean speech has a lower degree of spectral entropy [13]. Also, it can provide information about the randomness or complexity of the spectral content of a signal, as shown in figure (9).

Similar to energy entropy, SE is calculated by first dividing the short-term frame's spectrum into  $L$  sub-bands (bins). The energy of the  $f$ th sub-band ( $E_f$ ) is first normalized by the total spectral energy with  $f = 0, \dots, L - 1$ , which is then normalized by the total spectral energy ( $nf$ ) and is calculated according to the formula in equation 10 [11] [13]:

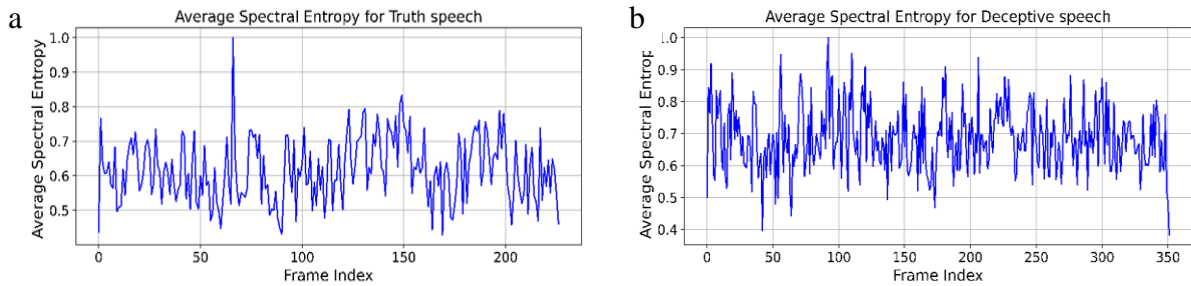
$$nf = \frac{E_f}{\sum_{f=0}^{L-1} E_f} \tag{10}$$

The entropy of the normalized spectral energy  $nf$  is finally computed according to the equation 11:

$$H = - \sum_{f=0}^{L-1} nf \cdot \log_2(nf) \tag{11}$$

where  $L$  is sub-bands (bins),  $E_f$  is the energy, and  $nf$  is normalized spectral energy.

Figure (9) illustrates the average SE, calculated as the average of total frames: (a) truth speech and (b) lying speech. It is worth noting that these average frames represent all speakers.



**FIGURE 9.** Average SE Feature for All Speakers: (a) for Truth Speech (b) for Lie Speech

### 3- Spectral Rolloff

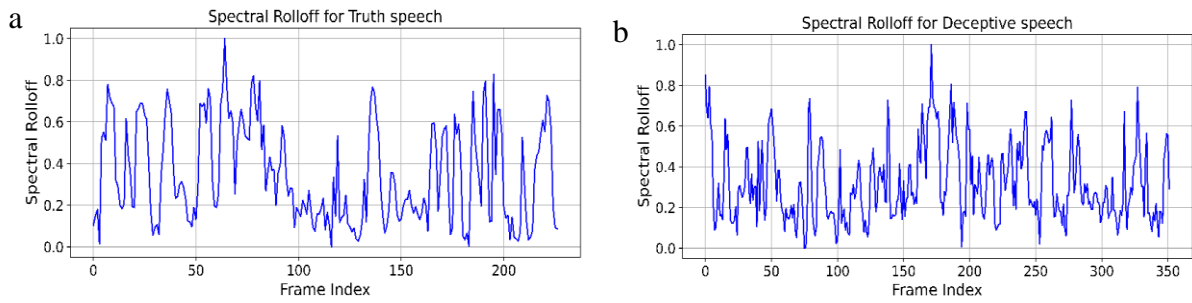
Spectral rolloff is described as the frequency below which a specific portion of the spectrum's magnitude distribution is concentrated (often about 90%). This feature has been used in this research study, because it has been discovered that the energy of a voice signal falls within a certain frequency band when it has emotional content. The frequency contents below which particular percentages of the total quantity of energy remain are shown by spectral roll-off. Depending on the signal's qualities, a value between 0.95 and 0.85 is typically chosen for this purpose, in terms of convenience. In addition, spectral roll-off offers details of the spectral shape, which establishes the quantity of high-frequency component present in a voice signal [17].

When used to distinguish between voiced and unvoiced audio signals, spectral roll-off can also be thought of as a spectral form descriptor of an audio signal, and it satisfies equation 12 below [11]:

$$\sum_{k=1}^m X_i(k) = C \sum_{k=1}^{wf_L} X_i(k) \tag{12}$$

where  $C$  is the adopted percentage (user parameter). Figure (10) illustrates the average RollOff, calculated as the average of total frames: (a) truth speech and (b) lying speech. It is worth noting that these average frames represent all speakers.





**FIGURE 10.** RollOff Feature for All Speakers: (a) for Truth Speech (b) for Lie Speech

**4- MFCCs**

One of the most popular audio features, MFCC has been employed across a variety of audio applications (speech recognition, speaker recognition, sound classification). It has been demonstrated that the MFCC parameters developed by Davis and Mermelstein (1980) are effective in gathering important auditory information. These parameters are also shown to be successful, across a wide variety of classification systems. The process through which the MFCC is calculated can be described as follows [13], [18]:

- for each frame, the fast Fourier transform (FFT) is computed to convert the time-domain data to the frequency domain
- to produce the magnitude spectrum, the absolute value is taken
- Mel frequency scale, a pitch measurement used to convert frequencies ( $f$ ) in hertz into its equivalent value in Mel, is calculated according equation 13:

$$Mel = 1127.0 \log \left( 1 + \frac{f(HZ)}{700} \right) \tag{13}$$

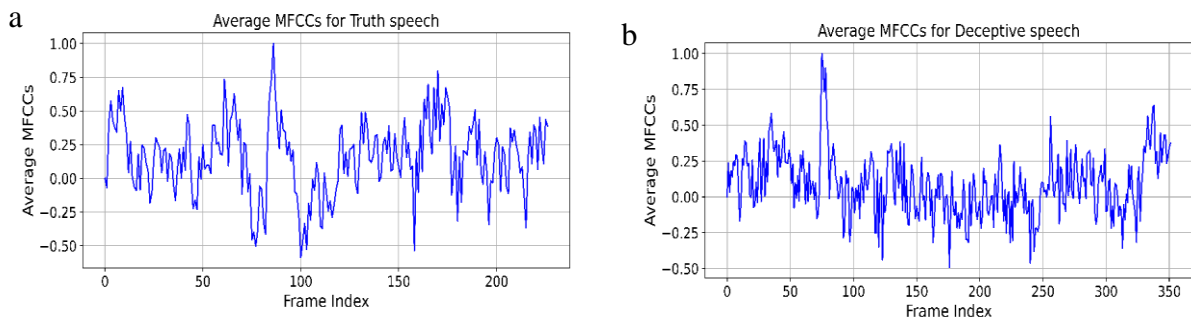
- the reduced spectrum is measured by handling the spectrum via a triangle Mel filter bank
- finally, MFCC is computed by calculating the discrete cosine transform (DCT) of the reduced log energy spectrum:

$$C_{co} = \sum_{j=1}^{N_f} \log(E_j) \cos \left[ \left( j - \frac{1}{2} \right) \frac{i\pi}{N_f} \right] \tag{14}$$

where  $E_j$  refers to the spectral energy calculated in the band of the  $j$ th Mel filter;  $N_f$  represents the total number of Mel triangular filters in the bank.

$N_c$  refers to the total number of applied cepstral coefficients ( $C_{co}$ ) extracted from each window frame. The default value for  $N_c$  is 12 [13]. In this research study, the number of coefficients is 23.

Figure (11) illustrates the average MFCCs calculated as the average of total frames: (a) truth speech and (b) lying speech. It is worth noting that these average frames represent all speakers



**FIGURE 11.** Average MFCCs Feature for All Speakers: (a) for Truth Speech (b) for Lie Speech

### 5- Pitch

Pitch is the fundamental frequency (F0), which is thought to be the main indicator of harmonics. The analysis and synthesis of speech and music, as well as their segmentation process, all require pitch. Pitch is typically only well established in spoken speech and harmonic music [13]. It may also convey the loudness of the sound and establish whether it is loud or weak. Pitch is an auditory perception that lets people know whether a sound is low or high by giving them a sense of its pitch [4]. Figure (12) illustrates the average Pitch, calculated as the average of total frames: (a) truth speech and (b) lying speech. It is worth noting that these average frames represent all speakers.

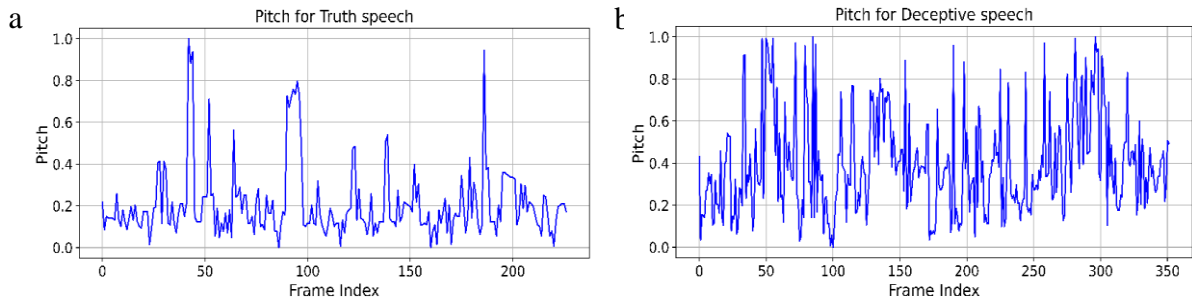


FIGURE 12. Pitch Feature for All Speakers: (a) for Truth Speech (b) for Lie Speech

## 2.3 Features Selection

One of the most-used scenarios to minimize the number of features in our model involves identifying the most important features list and retraining the model using just the most significant features. For instance, we might seek to minimize the variance of the model or increase interpretability by just adding the most crucial features [19]. In this paper, RF method has been chosen to identify the most significant features for lie detection.

### 2.3.1 RF

The basic idea of RF involves ensembles of slightly different trees created through training on random training subsets. To achieve a better overall goal with a lower error rate, more reliable results, and better noise insensitivity than could be obtained from a solitary module, an ensemble methodology was initially used to combine a set of existing modules, with each module working on the same classification problem (Breiman, 2001a; Maimon & Rokach, 2008). First, RF is used for selection of important features, wherein the model trains overall features to get the important feature, as shown in Figure (13). Subsequently, the model is trained only on the most important feature [13].

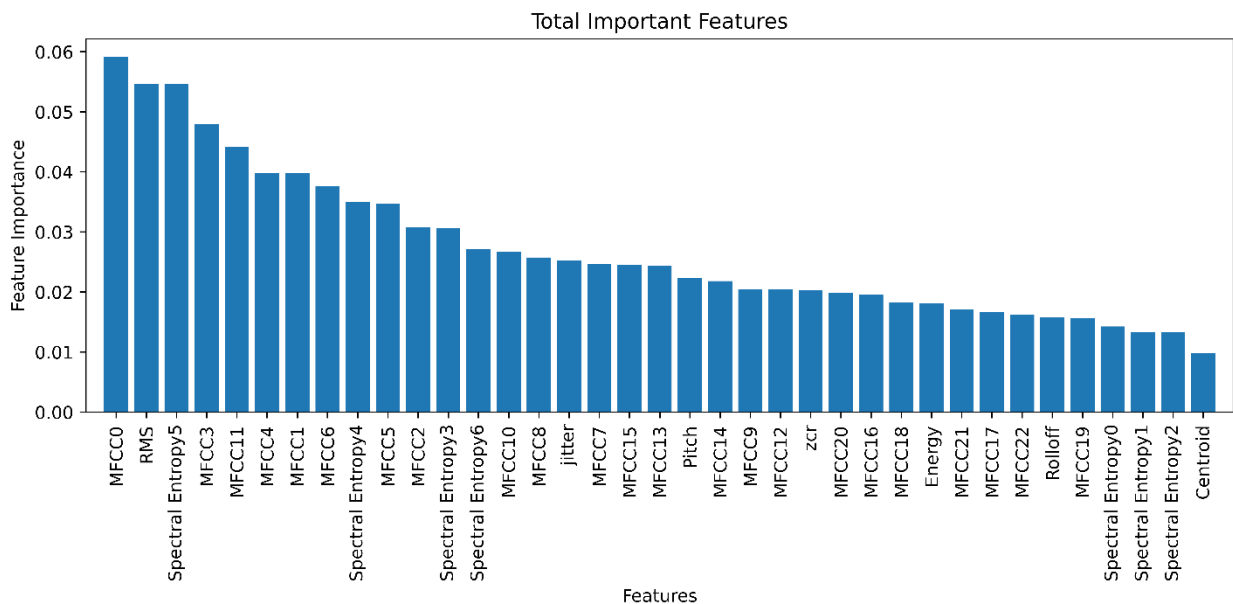


FIGURE 13. All Important Features Obtained Using RF

## 2.4 Training and Testing

In this section, the training and testing of the model will be described. Post execution of the feature-extraction process and after applying feature-selection techniques, the data has been found ready for training and testing the model. The data has been labeled, with a value of 1 assigned to lie speech and 0 assigned to true speech: the final matrix had dimensions of 169,476 rows and 37 columns, representing the extracted features and corresponding labels. RF has been chosen as the classification model for this task. RF is an ensemble learning method that utilizes multiple decision trees to make predictions. RF has been selected due to its ability to handle large feature sets and provide robust performance. The data has been split into a training set and a testing set, with a ratio of 0.8 for training and 0.2 for testing. The training set, which consists of 80% of the data, has been used to train the RF model. During training, RF has created 100 estimators, which are individual decision trees, and has combined their predictions to make the final classification decision. The remaining 20% of the data, the testing set, has been used to evaluate the performance of the trained RF model. The model has been applied to the testing set, and its predictions compared to the ground truth labels. The model has learned patterns from the training data, so as to make accurate predictions on unseen data. To address the issue of overfitting, which arises when a model performs well on the training data but fails to generalize to unseen data, cross-validation has been employed. Cross-validation involves assessing the model’s performance on multiple subsets of the data, known as folds. In this particular study, a five-fold cross-validation approach has been adopted. This technique aids in evaluating the model’s ability to generalize, by testing its performance on unseen data, thus providing a more robust assessment of its effectiveness.

## 3. DISCUSSION OF RESULTS

**Table 1.** Model Accuracy Using RF

Number of Important Features	Accuracy
10	0.834606874
15	0.862959139
20	0.874966809
25	0.88071987
30	0.882342528
35	0.879775778

The results indicate a trend of increase in accuracy along with a rise in the number of selected features . The model achieved the highest and purest levels of accuracy, when 25 features were utilized. This suggests that the additional information captured by those features contributes positively to the performance of classification.

It is also important to provide insights into the performance of the model using the confusion matrix. The confusion matrix provides a detailed breakdown of the model. The performance of the trained RF classifier has been assessed using varying numbers of selected features, and the corresponding accuracy values have been presented in Table 1. To gain further insights into the model’s performance, the confusion matrix has been calculated for the number of features chosen (25 features). When 25 features are selected, the model achieves an accuracy of 0.8807. The corresponding confusion matrix is shown in Table 2.

**Table 2.** Confusion Matrix for Lie and Truth Speech

<i>Predicted</i>	<i>True Speech</i>	<i>Lie Speech</i>
<i>Actual</i>		
<b>True Speech</b>	44.64670305	5.64685057
<b>Lie Speech</b>	6.28116241	43.42528397
<b>Performance Measurements at 25 features</b>		
<b>Recall</b>	0.89	0.87
<b>Precision</b>	0.88	0.88
<b>F1 Score</b>	0.88	0.88

## 4. CONCLUSION

In conclusion, this study introduces a deception-detection technique specifically designed for isolated speech utterances and utilizing voice-stress analysis. It addresses the limitations of existing approaches that rely on psychological and behavioral measures such as polygraph testing, which have faced criticism due to their lack of reliability and validity. While newer techniques such as brain imaging and machine learning hold promise, they are still in the early stages of development and require further testing to effectively detect deception.

The proposed technique has been validated using the “real-life trial data” dataset, comprising 61 lie samples and 60 truth samples. The analysis has involved extracting time domain and spectral features from audio signals, employing overlapped frames, and utilizing an RF-based classification algorithm. The experimental results demonstrate an overall classification accuracy of 88% for distinguishing between lie and truth classes. Additionally, the findings reveal that lying speech exhibits higher levels of randomness compared to true speech, as depicted in the feature figures presented above. These outcomes indicate the potential effectiveness of the proposed technique in detecting deception in isolated speech utterances.

However, it is crucial to acknowledge the study’s limitations. Generating a valid baseline for an individual’s regular voice patterns poses challenges. This corresponds with other research identifying valid baseline generation as a common limitation of the field. External factors such as exhaustion, illness, or emotional state can introduce variations in stress markers, thereby influencing the reliability of the deception-detection process.

While the aforementioned outcomes underscore the potential effectiveness of the proposed technique in detecting deception during isolated speech utterances, further research and testing are imperative. Validation across different datasets and real-world scenarios, coupled with an exploration of its robustness in diverse contexts and with varied populations, will contribute to a more comprehensive understanding of the technique’s applicability and reliability.

## Funding

None

## ACKNOWLEDGEMENT

None

## CONFLICT OF INTEREST

The author declares that there is no conflict of interest.

## REFERENCES

- [1] A. Xue, H. Rohde, and A. Finkelstein, ‘An Acoustic Automated Lie Detector’, 2019.
- [2] I. J. Mohammed and L. E. George, ‘Lie Detection and Truth Identification form EEG signals by using Frequency and Time Features’, *J Algebr Stat*, vol. 13, no. 3, pp. 4102–4121, 2022, [Online]. Available: <https://publishoa.com><https://publishoa.com>
- [3] S. V. Fernandes and M. S. Ullah, ‘Use of Machine Learning for Deception Detection from Spectral and Cepstral Features of Speech Signals’, *IEEE Access*, vol. 9, pp. 78925–78935, 2021, doi: 10.1109/ACCESS.2021.3084200.
- [4] F. M. Marcolla, R. de Santiago, and R. L. S. Dazzi, ‘Novel lie speech classification by using voice stress’, in *ICAART 2020 - Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, SciTePress, 2020, pp. 742–749. doi: 10.5220/0009038707420749.
- [5] E. P. Fathima Bareeda, B. S. Shajee Mohan, and K. V. Ahammed Muneer, ‘Lie Detection using Speech Processing Techniques’, in *Journal of Physics: Conference Series*, IOP Publishing Ltd, May 2021. doi: 10.1088/1742-6596/1921/1/012028.
- [6] S. Mihalache and D. Burileanu, ‘Using Voice Activity Detection and Deep Neural Networks with Hybrid Speech Feature Extraction for Deceptive Speech Detection’, *Sensors*, vol. 22, no. 3, Feb. 2022, doi: 10.3390/s22031228.
- [7] H. C. Chou, Y. W. Liu, and C. C. Lee, ‘Automatic Deception Detection Using Multiple Speech and Language Communicative Descriptors in Dialogs’, *APSIPA Trans Signal Inf Process*, 2021, doi: 10.1017/ATSIP.2021.6.
- [8] H. Fu, H. Yu, X. Wang, X. Lu, and C. Zhu, ‘A Semi-Supervised Speech Deception Detection Algorithm Combining Acoustic Statistical Features and Time-Frequency Two-Dimensional Features’, *Brain Sci*, vol. 13, no. 5, May 2023, doi: 10.3390/brainsci13050725.
- [9] ‘Rada Mihalcea: Downloads’. Accessed: Jun. 30, 2023. [Online]. Available: <https://web.eecs.umich.edu/~mihalcea/downloads.html>

- [10] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, 'Deception detection using real-life trial data', in *ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction*, Association for Computing Machinery, Inc, Nov. 2015, pp. 59–66. doi: 10.1145/2818346.2820758.
- [11] Theodoros Giannakopoulos and Aggelos Pikrakis, 'Audio Analysis: A MATLAB Approach', 2014. Accessed: Mar. 15, 2023. [Online]. Available: <http://booksite.elsevier.com/9780080993881>
- [12] H. Jeon, Y. Jung, S. Lee, and Y. Jung, 'Area-Efficient Short-Time Fourier Transform Processor for Time–Frequency Analysis of Non-Stationary Signals', *Applied Sciences (Switzerland)*, vol. 10, no. 20, pp. 1–10, Oct. 2020, doi: 10.3390/app10207208.
- [13] D. Yehya Mohammed, 'Overlapped Speech and Music Segmentation Using Singular Spectrum Analysis and Random Forests', 2017.
- [14] Homayoon Beigi, 'Fundamentals of Speaker Recognition', Yorktown Heights, NY, USA, 2011. doi: 10.1007/978-0-387-77592-0.
- [15] A. Koduru, H. B. Valiveti, and A. K. Budati, 'Feature Extraction Algorithms to Improve the Speech Emotion Recognition Rate', *Int J Speech Technol*, vol. 23, no. 1, pp. 45–55, Mar. 2020, doi: 10.1007/s10772-020-09672-4.
- [16] M. B. Er, 'A Novel Approach for Classification of Speech Emotions Based on Deep and Acoustic Features', *IEEE Access*, 2020, doi: 10.1109/ACCESS.2020.3043201.
- [17] M. Narayan Mohanty, 'Emotional Speech Recognition Using Optimized Features', 2017. [Online]. Available: <https://www.researchgate.net/publication/321364769>
- [18] D. Y. Mohammed, K. Al-Karawi, and A. Aljuboori, 'Robust Speaker Verification by Combining MFCC And ENTROCY in Noisy Conditions', *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 4, pp. 2310–2319, Aug. 2021, doi: 10.11591/EEI.V10I4.2957.
- [19] C. Albon, 'Machine Learning with Python Cookbook : Practical Solutions from Preprocessing to Deep Learning.