The Most Important Features of Lie Detection Using Voice Stress

Fadi K. Al-Dhaher[*], Duraid Y. Mohammed [**], Mohammed Khalaf [***]

[*] College of Engineering, Al-Iraqia University, Baghdad, Iraq
faadii.kamal@gmail.com
https://orcid.org/0009-0006-7774-8537

[**] College of Engineering, Al-Iraqia University, Baghdad, Iraq
duraidyehya19@gmail.com
https://orcid.org/0000-0002-9586-1983

[***] General Directorate of Education Anbar, Ramadi 31001, Iraq
Computer Science Department, Al-Maarif University College, Ramadi 31001, Iraq
m.i.khalaf@uoa.edu.iq
https://orcid.org/0000-0003-0559-455X

Abstract

This study proposes a lie detection module that leverages audio features in both the time and frequency domains to scrutinize key features and speech patterns for more precise dishonesty detection. The research focuses on addressing the limitations of traditional methods and suggests practical alternatives, aiming to contribute to the improvement of existing understanding and system architectures in the field of deception detection. where Detecting deception is critical in various fields, such as law enforcement, national security, and personal relationships. While traditional methods like polygraph exams are criticized for their reliability, emerging real-time technologies like voice stress analysis and speech processing techniques present alternatives. This research aims to enhance lie detection systems by exploring the potential of audio features, offering a more nuanced approach to identifying dishonesty, and overcoming the limitations of current methods. The study introduces a lie detection algorithm utilizing a real-world dataset collected with a handheld microphone, replicating authentic situations. Feature selection employs the random forest technique, focusing on the most significant features for training and testing. The algorithm achieved a 79% accuracy rate through rigorous examination, with Mel-frequency cepstral coefficients (MFCC) identified as the crucial feature for lie detection. This underscores the method's potential effectiveness in real-time fraud detection. However, further research is necessary to confirm its consistency across various datasets, situations, and demographic groups.

Keywords- *Voice stress, Lie speech, Random Forest (RF), Deception detection, Machine learning*.

I. Introduction

Lie detection has become increasingly significant in recent years, with potential applications across diverse fields. In law enforcement and criminal investigations, identifying deception plays a crucial role in solving cases and ensuring public safety. However, the impact of lie detection extends beyond these realms to include national security, workplace ethics, and personal relationships, where the repercussions of dishonesty can be substantial. Traditionally, lie detection methods have relied on processes like polygraph testing, which has faced criticism for its perceived lack of reliability and validity [1]. However, recent developments in technology and research have led to the creation of new and more innovative approaches to detect deception, such as brain imaging [2], machine learning [3], and voice stress analysis [4].

In lie detection research, some frequently employed machine learning algorithms encompass support vector machines, decision trees, neural networks, and random forests are used to evaluate aspects of speech or physiological responses, such as heart rate or skin conductance, to discover lying patterns. These new methods provide the possibility for better and non-invasive lie detection, but they also raise critical ethical problems concerning privacy, informed consent, and potential harm to participants. Despite these limitations, lie detection research continues to advance, with academics exploring creative techniques to spot deceit in real-time. By utilizing recent technological breakthroughs and depending on ideas from disciplines such as psychology and computer science, lie detection researchers aim to construct more accurate and trustworthy algorithms that can be employed in numerous contexts. The potential benefits of such approaches are enormous, not only in law enforcement and criminal investigations but also in other industries where honesty and trust are essential components of successful interactions and outcomes [5].

http://doi.org/10.58564/IJSER.3.1.2024.149

II. Related Work

Hannah and Adam developed a sequential neural network and various machine-learning models dedicated to lie detection, relying solely on acoustic cues in speech. They utilized a balanced dataset comprising recordings of both lying and truthful speech, collected from a two-person lying game. y extracting mel-frequency cepstral coefficients (MFCC), energy envelopes, and pitch contours, the researchers trained a majority-voting ensemble learning classifier. This ensemble comprised a Gradient Boosting Classifier (GBC), a Support Vector Machine (SVM), and a Stochastic Gradient Descent (SGD), each specifically trained on MFCC and energy features. The most successful model achieved a maximum accuracy of 55.8% in identifying lies [1].

Huang-Cheng and others introduce an autonomous deception detection method that extensively incorporates domain knowledge into deceptive analysis. To enhance the automated detection of deception in dialogues, the researchers conducted a comprehensive analysis encompassing acoustics, textual information, implicatures involving non-verbal behaviors, and conversational temporal dynamics. Employing this approach on the Daily Deceptive Dialogues corpus of the Mandarin (DDDM) database, their proposed method achieved cutting-edge performance in deception recognition, attaining an unweighted accuracy recall score of 80.61%. Subsequent investigations unveiled the significance of specific auditory parameters, such as loudness and MFCC, along with textual features, as crucial indicators for detecting deceptive behaviors. Notably, the study revealed that the deceptive behaviors of individuals can be identified by scrutinizing the activities of interrogators during the conversational period.[6].

Fathima and others introduce an effective lie detection methodology centered on a non-invasive approach, specifically focusing on capturing the subject's spoken expressions. They extract discriminative and pertinent features from the speech data and construct classifiers based on Support Vector Machines (SVM) to differentiate between truthful and deceptive statements. Their research aims to leverage psycho-neural properties and the reliance on speech signals to anticipate and identify instances of deception in individual speech utterances, utilizing features such as MFCC and mean MFCC. To streamline the feature set while minimizing computational complexity and costs, Principal Component Analysis (PCA) is employed. Their experiments demonstrate an overall classification accuracy of 81% for lying and 78% for truth classes when utilizing polynomial and Gaussian kernels, respectively [5].

Sinead and her colleagues delve into the distinctions between truthful and deceptive speech, examining four distinct nonlinear speech variables. The research captures the speaker's responses in a police investigation, particularly under stress, with recordings of two genuine and two deceitful responses at various times of the day. The study focuses on analyzing audio recordings from various sessions, with an emphasis on extracting cepstral features and spectral energy characteristics. Specifically, the investigation centers on Mel frequency cepstrum coefficients, delta cepstrum, time-difference cepstrum, energy of the Bark band energy, delta energy, and time-difference energy features.

Deception identification is carried out through classification techniques, including the Levenberg-Marquardt method and the long-short-term memory classification technique. The evaluation of accuracy involves nine specific training and testing combinations derived from three distinct sessions. This assessment incorporates the cepstral and spectral energy features that have been extracted.

To optimize overall performance, principal component analysis is implemented to reduce the dimensionality of the features. This reduction aims to enhance the efficiency of the deception identification process across the implemented classification techniques and various training-testing combinations [3].

Serban and Dragos present a deep neural network (DNN) approach tailored for automatic speech recognition, with a specific focus on vocal activity detection (VAD). The study explores various DNN architectures, including multilayer perceptron (MLPs), recurrent neural networks (RNNs), and convolutional neural networks (CNNs). Notably, CNNs demonstrate superior performance in the context of VAD. To enhance efficiency, the authors incorporate additional processing techniques such as hysteretic thresholding, minimum duration filtering, and bilateral extension.

The experimental setup involves training and evaluating the systems using subsets of the CENSREC-1-C database, which is subjected to diverse simulated ambient noise conditions. Further testing extends to a different CENSREC-1-C dataset featuring real ambient noise and a segment of the TIMIT database.

The authors elaborate on the adaptation and integration of the final VAD system into an utterance-level deceptive speech detection (DSD) processing pipeline. This integration utilizes the Kolmogorov-Smirnov ranking of features for feature selection, incorporating features such as MFCC, Energy, and Pitch. Optimal DSD performance is achieved through a unique hybrid CNN-MLP network, which combines algorithmically and naturally formed speech features. This hybrid network attains an unweighted accuracy (UA) of 63.7% on the RLDD database and 62.4% on the RODeCAR database [7].

Fatma introduces a novel lie detection approach that revolves around an Enhanced Recurrent Neural Network (ERNN) designed with a Long Short-Term Memory (LSTM) architecture. The LSTM's hyperparameters are meticulously fine-tuned using Fuzzy Logic. The analysis focuses on MFCC characteristics extracted from each database. Subsequently, the LSTM model is created and applied for training, using a training input dataset with corresponding outputs. The neural network training utilizes a database of audio recordings generated from interviews with a randomly selected group. Impressively, the proposed ERNN achieves an accuracy of 77.3%. This finding holds significant relevance for voice stress analysis, indicating the potential to detect patterns in the voices of individuals experiencing stress [8].

Felipe and others construct a neural network to assess a person’s voice and classify his speech as reliable or not. In order to attain the aims, a recurrent neural network of LSTM architecture was created based on a framework already utilized in other studies, and by adjusting parameters, different outcomes were reached in the testing. The features that were extracted in the present study are MFCC, Pitch, and jitter. A database of audio recordings was built to execute the neural network training based on an interview with a randomly chosen group. Considering all the neural network base models built, the one that demonstrated significance presented a precision of 72.5% of the data samples. For the sort of issue under focus, which is voice stress analysis, the result is statistically important and shows that it is possible to uncover patterns in the voice of people who are under stress [4].

To address the gap in research concerning the detection of deceptive speech in the local Chinese context, Wenjing Wang and colleagues have successfully developed a Chinese deceptive speech database. The study involved 124 participants (71 females) from Jiangxi University of Traditional Chinese Medicine. Participants were instructed to provide both genuine and false descriptions of videos within a gaming environment, while their facial expressions and audio data in the description section were recorded.

The dataset obtained consisted of 439 false audios (273 from females) and 451 honest audios (280 from females). Sound features, namely Mel-frequency cepstral coefficients (MFCCs), Jitter, and fundamental frequency (F0), were individually extracted using Python. Subsequently, Weka was employed to segregate the deceptive and honest audio features. The results revealed that the use of RIFCC features in the mixed dataset yielded the most effective classification performance. Furthermore, the male dataset demonstrated the highest correct rate across all three classifiers when utilizing the Jitter feature, while the female dataset achieved the highest correct rate in all three classifiers under the F0 feature [9].

Hongliang Fu and collaborators present an innovative semi-supervised approach for detecting deception in speech by integrating acoustic statistical features with time-frequency two-dimensional features. The methodology involves the development of a hybrid semi-supervised neural network, combining a semi-supervised autoencoder network (AE) with a mean-teacher network. In this process, static artificial statistical features are input into the semi-supervised AE to extract robust and advanced features. Concurrently, three-dimensional (3D) mel-spectrum features are fed into the mean-teacher network to capture features rich in time-frequency two-dimensional information.

The amalgamation of these features is further strengthened by the application of a consistency regularization method, effectively addressing overfitting concerns and enhancing the model's generalization ability. Through experimental evaluations conducted on a self-built corpus specifically designed for deception detection, the proposed algorithm achieves a notable recognition accuracy of 68.62%. This surpasses the baseline system by 1.2%, demonstrating significant improvements in detection accuracy. The results underscore the efficacy of the hybrid semi-supervised neural network in effectively leveraging diverse features for enhanced speech deception detection performance [10].

This work contributes to the field of real-time lie detection through audio analysis. First, we created a unique dataset in a genuine, real-world environment using a wireless microphone, capturing authentic audio signals. To refine the speech quality, we implemented a noise reduction algorithm, enhancing audio clarity and intelligibility. Subsequently, we extracted a comprehensive set of features, encompassing both time and frequency domains, including zero crossings, root mean square, jitter, energy, Spectral Centroids (Ci), Spectral Entropy (SE), Mel-Frequency coefficients (MFCCs), Pitch, and Spectral Roll-Off. Notably, we employed the Feature Importance method with random forest (RF) for feature selection, identifying critical characteristics for lie detection.

III. Research elaboration

The proposed system includes six stages, as shown in Figure 1.

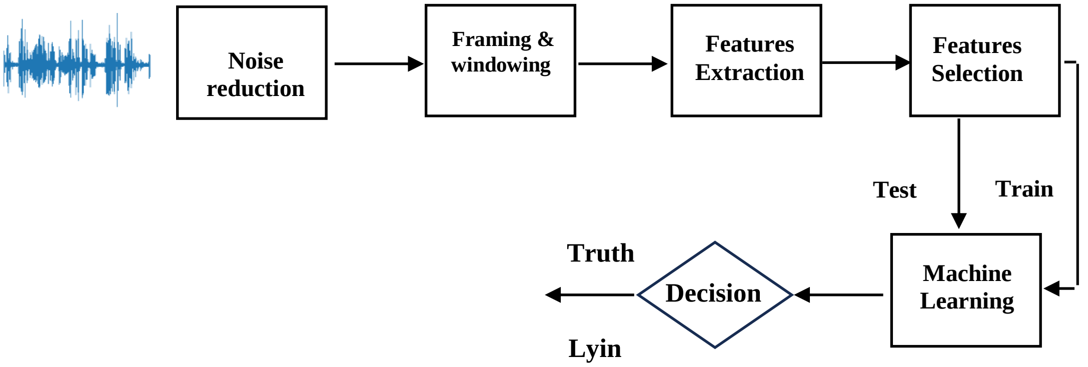


Figure 1 Block diagram of Lie detection system

A. Dataset

This study generated a real-time dataset by capturing sounds in a natural environment using a wireless microphone. The audio collection process involved two distinct scenarios with a predefined set of questions, prompting participants to provide both genuine and deceptive responses. To minimize interference from background noise, a spectral subtraction approach was applied for background noise reduction.

Following the recording phase, the audio data was segmented into clips, resulting in 153 samples for deceptive speech and 161 samples for truthful speech. The average duration of the audio snippets in the dataset was approximately 23 seconds. Furthermore, misleading speech had an average duration of nearly 10.5 seconds, while genuine speech averaged around 12.5 seconds.

The dataset encompasses recordings from a diverse set of speakers, featuring four unique female speakers and three unique male speakers, primarily ranging in age from 20 to 60 years old. The sample rate (sr) of the audio stream in this dataset was 22050 Hz.

To validate the suggested method with the real-time recording, this research utilized a text-dependent real-time dataset, where participants answered specific, predefined questions. by recording sounds in a real-world environment using a wireless microphone. The audio collection technique includes two different scenarios with a specified set of questions.

B. Noise Reduction

Noise reduction, also known as speech enhancement, is designed to remedy the degradation of speech caused by additive noise. The main goal of voice augmentation is to enhance the quality and intelligibility of speech that has been negatively impacted by external factors. Enhanced quality is particularly beneficial as it can alleviate listener fatigue. To mitigate noise or enhance speech, algorithms commonly termed noise reduction algorithms are employed. These algorithms work to attenuate or suppress background noise to a specific extent. Noise can take various forms and exist in different environments, whether it be on the street with factors like passing cars or construction work, in a vehicle with sounds such as engine noise or wind, or even in an office where noise sources may include PC fan noise or air ducts. These examples highlight the diverse nature of noise experienced in everyday life[11]. in this work, Spectral subtractive algorithm is used to reduce background noise and it is used only on the real-time dataset.

- *Spectral subtraction Algorithm*

Historically, the spectral-subtractive approach stands out as one of the earliest methods developed for noise reduction. furthermore, it is the simplest improvement strategy for noise reduction[11].In spectral subtraction, the assumption is made that the initial milliseconds of the signal, typically characterized by silence without speech, primarily consist of additive noise. Additionally, gaps of silence between speech segments are considered to contain additive noise only. The approach involves calculating the average of the noise, extracted from these silent portions, and subtracting it from the speech signal in the frequency domain. This subtraction process aims to enhance the informative content of the speech signal by reducing the impact of the identified additive noise.

The basics of spectral subtraction are Assuming the additive noise of an audio signal in the time domain, after framing and windowing the signal, the Fourier transform utilizing the Fast Fourier transform (FFT) is applied to transfer the signal into the frequency domain.

Following estimated noise, spectral subtraction is achieved by subtracting an estimate of the noise spectrum from the noisy speech spectrum.

The enhanced signal is produced by executing the inverse Fast Fourier transform of the predicted signal spectrum, utilizing the phase information obtained from the noisy signal. Notably, the algorithm is relatively simple, necessitating only a forward and an inverse Fourier transform [11], [12]. Figure 2 A block diagram illustration of spectral subtraction.

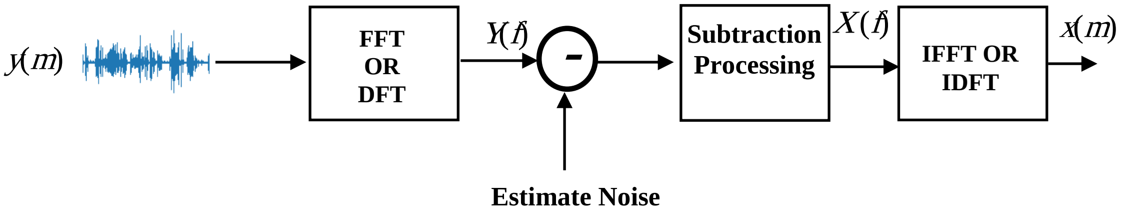


Figure 2 A block diagram representation of spectral subtraction[12]

Where y(m), Y(f), X (f), and x(m) are the signal, the Fourier transforms of the noisy signal y(m), Fourier transforms of the original signal, and x(m) the enhancement signal.

Table 1 presents the average signal-to-noise ratio (SNR) values for the dataset both before and after the application of noise reduction techniques. The SNR values are presented in two categories: before and after noise reduction. The "before" values signify the original SNR levels in the raw audio data, while the "after" values reflect the SNR improvements achieved through the noise reduction process.

Table 1 Average SNR before and after noise reduction for Real time dataset

Real time Dataset	
State	**SNR**
Before	-0.093 dB
After	2.165 dB

Also, the following strategy is proposed and published by Mohammed et al. (2015) [13] here for detecting noise spectra and then speech enhancement:

- The audio file is split into one-second time segments to make distribution more efficient.
- Segment energy is calculated using Equation 1 to identify silent segments depending on a pre-set threshold for segments that have energy less than the statistical threshold (this is confirmed based on 1000 samples from the benchmark database, which is deployed in this study).

$$E = (10log_{10}\frac{1}{N}\sum\nolimits_{n=1}^{N}|(X(n)|^{2})/M \qquad (1)$$

where *n*, *N*, and *M* are time index, window length and the highest peak value respectively.

- Voiced parts are studied to generate a speech timestamp and for further processing, whereas the non-speech choice is mapped straight to the non-speech timestamp without the need for further study.
- For overlapped soundtracks, assuming that the noisy signal contains clean speech and noise based on the spectral subtractive approach concept, see the following Equation 2.

$$y(m) = s(m) + x(m) \qquad (2)$$

Where y(m) noisy signal, s(m) clean speech and x(m) noise

- Based on Equation 6-2, the clean speech can be calculated as given in Equation 2

$$s(m) = x(m) - y(m) \qquad (3)$$

- Then, calculate the discrete time FFT for both sides, which could be represented in Equation 4

$$S(m) = X(m) - Y(m) \qquad (4)$$

- The estimation of speech spectra and noise phase is required to calculate clean speech through the polar form, which is defined in Equation 5.

$$S(i) = \left|X(i) - |d(i)|\right|e^{(j\emptyset d(i))} \qquad (5)$$

where *Øy(i)* represents the noisy speech phase and *i frame index.*

- Finally, the estimated clean time domain signal is calculated by applying the Inverse Fast Fourier Transform IFFT.

C. Framing and Windowing

In various applications, audio signals undergo analysis through a method known as short-term (or short-time) processing. This approach involves dividing the audio signal into overlapping short-term frames (windows) for frame-by-frame analysis. Essentially, the signal is segmented into these frames, allowing for more detailed examination on a frame-by-frame basis [14]. The length of the frames plays a vital role because it establishes the frequency quality of the spectrum. A longer frame leads to greater frequency resolution at the expense of diminishing the quality of temporal resolution. On the other hand, shorter windows produce a more detailed representation in the time domain but, overall, lead to poor frequency resolution, and their duration usually ranges from 10 to 50 ms [14]. For that reason, the frame length was set at 40 ms in this work.

Additionally, the hop size, which affects the overlap between successive frames, was set at 50% of the frame size, as seen in Figure 3. The choice of hop length has an effect on the degree of overlap between subsequent frames. A short hop length increases the overlap, indicating that succeeding frames share more data. Conversely, a larger hop length minimizes the overlap between frames [14].

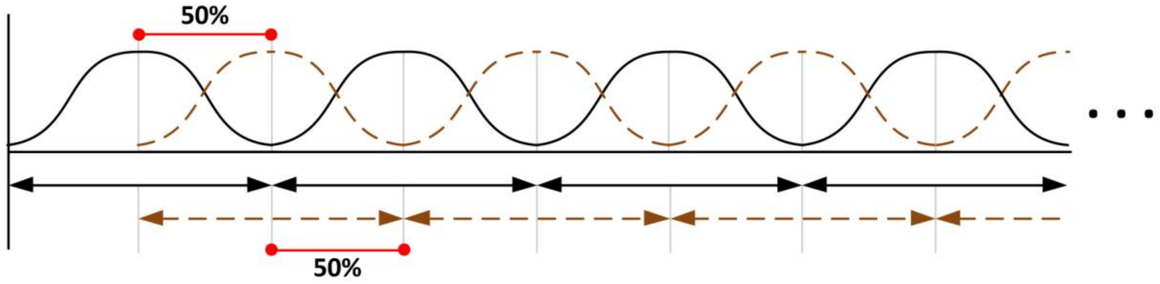


Figure 3 Frames Overlap with 50% of frame size[15]

Additionally, the windowing operation is employed to mitigate the edge effects of the frame. Various window types exist, with the Hamming window being a commonly used choice due to its low spectral leakage. This window type is versatile for general-purpose applications, offering a balance between low distortion and high signal recovery capability. However, it may result in slightly reduced resolution. In this research, the Hamming window has been employed as a filtering window to reduce edge discontinuities [16].

D. Features Extraction

Feature extraction plays a crucial role in audio analysis, serving as a fundamental component. It is a critical processing step in tasks related to pattern recognition and machine learning. The primary objective is to extract a set of features from the dataset of interest. These features are intended to convey informative value regarding the desired attributes of the source data. Furthermore, feature extraction has the potential to function as a data rate reduction operation, aiming to base analysis algorithms on a relatively limited number of informative features [14] .

In the current study, the recovered features were categorized into two groups: those obtained from the temporal domain and those generated from the frequency domain. Frequency domain elements comprised Mel frequency cepstrum coefficients (MFCC), Pitch, Spectral Centroid (Ci), Spectral Entropy (SE), and Spectral Rolloff.

And, Time domain features encompassed Zero crossings (ZCR), jitter, centroids, Root Mean Square (RMS), and Energy. They were employed to assess the distinctive patterns between authentic and fraudulent speech. For each frame, audio features were extracted, resulting in a matrix (m x n), where m represented the number of frames and n signified the number of features. The size of this matrix was (113048*37) characteristics.

Also, the process of averaging frames was implemented to enhance the fidelity of the audio signal by mitigating the impact of noise and other distortions. This approach contributes to improving the reliability and accuracy of various audio analysis techniques, such as speech recognition, speaker identification, and deception detection. However, careful consideration is essential when choosing the appropriate averaging method, as different techniques can influence the resulting feature values and their interpretation.

In this study, the initial step involved averaging all audio samples belonging to the same class, resulting in a single averaged audio sample. Subsequently, the moving average method was employed. This technique calculates the mean of the feature values over a sliding time window, effectively smoothing out fluctuations in the feature values. This results in a more stable representation of the feature trends over time. The described calculation method was applied to all the features outlined in the subsequent section.

1- Time Domain Features

In general, the time-domain audio properties are generated directly from the samples of the audio stream. Examples are energy, zero-crossing rate (ZCR), root mean square (RMS), and Jitter. Such characteristics give an easy way to evaluate audio signals; however, it is typically needed to combine them with more advanced frequency-domain features [14].

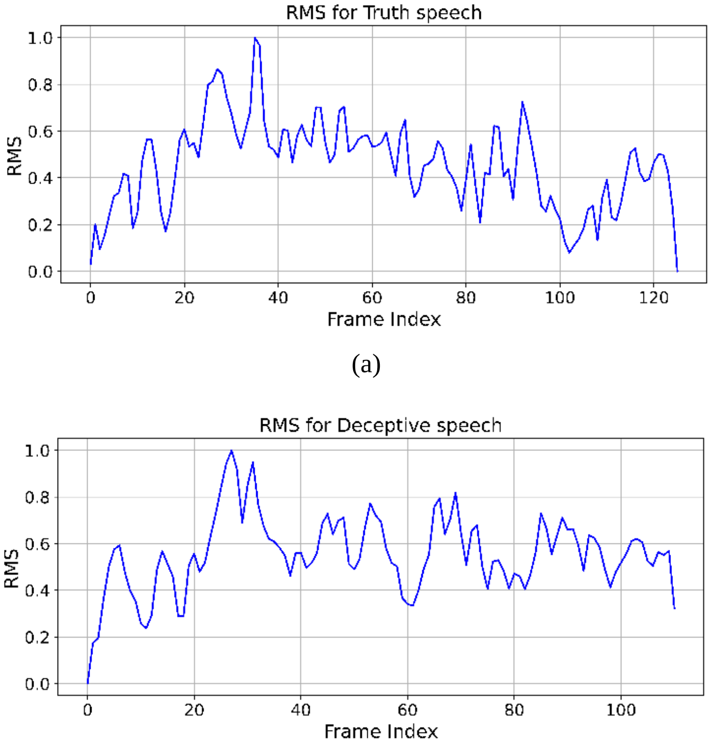
- *Root Mean Square (RMS)*

Typically, The Root Mean Square (RMS) of an audio signal is commonly represented in decibels (dB). A foundational description was provided by Kenny and Keeping in 1962, and it remains a widely used feature. Tzanetakis et al. noted that frames characterized by silence exhibited a lower RMS compared to frames without silence[16]. Also, RMS is the measure of the loudness of an audio signal[17]. which is indicative of variations in stress levels or emotional states. for that employed in this research. And it defined as equation 6.

$$RMS(i) = \sqrt{\frac{1}{L}\sum_{n=1}^{L} f_i(n)^2} \tag{6}$$

In the formula, L is the length of the frame,f_i denote an audio frame and n are the time of index of the audio signal sample [16].

Figure (4) depicts the average RMS calculated from the average total frames of truthful and lying speech, where the x-axis represents the average RMS of samples and the y-axis indicates the number of frames.



(a)

(b)

Figure 4 Average RMS for all speakers: (a) for true speech (b) for lying speech

- *Zero-Crossing Rate (ZCR)*

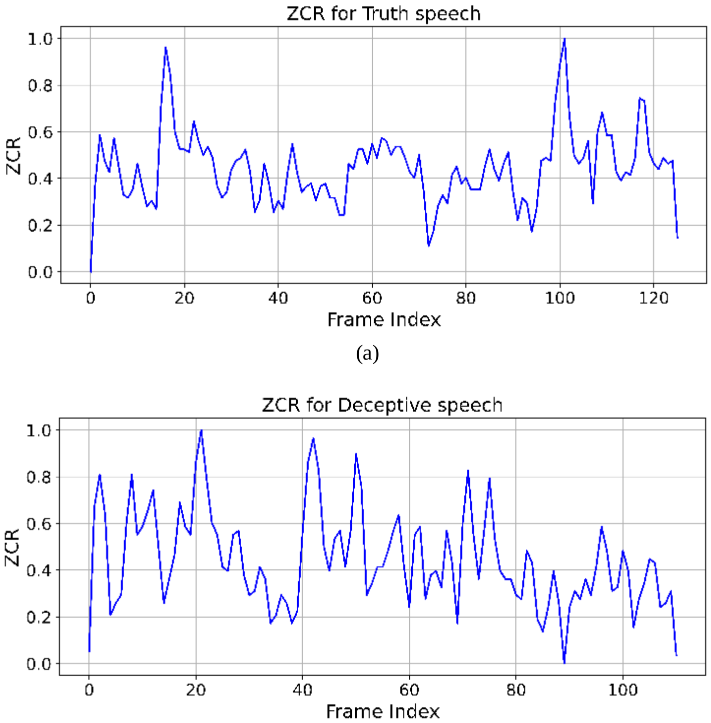
The zero-crossing rate (ZCR) is a measure of how frequently the sign of a signal changes within an audio frame. Put simply, it signifies the count of transitions the signal makes from positive to negative and vice versa, divided by the duration of the frame. Elevated ZCR values are typically observed in segments of the signal characterized by higher levels of noise [14]. Additionally, it demonstrates robustness as a feature, effectively discerning between music and speech or identifying speech from non-speech samples. This uncomplicated computational approach finds widespread application in various research projects, addressing objectives such as speech and music detection, as well as automated speaker recognition [16].
So that is why it is employed in this research, and also because ZCR is a crucial feature for recognizing brief and loud disturbances; it merely identifies minor variations in signal amplitude [18]. ZCR is computed according to equation 7.

$$z(i) = \frac{1}{2N}\sum_{n=1}^{N} |sgn[X_i(n)] - sgn[X_i(n-1)]| \qquad (7)$$

Where n= sequence of the audio sample, i=index of frames, N =frame length and sgn. Is the sign function[14].

$$sgn[X_i(n)] = \begin{cases} 1\,, X_i(n) \geq 0 \\ -1, X_i(n) < 0 \end{cases} \qquad (8)$$

Figure 5 illustrates the average ZCR derived from the average total truth and lying speech frames. It is worth mentioning that these frames represent all speakers. where the x-axis represents the average ZCR of samples and y-axis represent the number of frames.



(a)

(b)

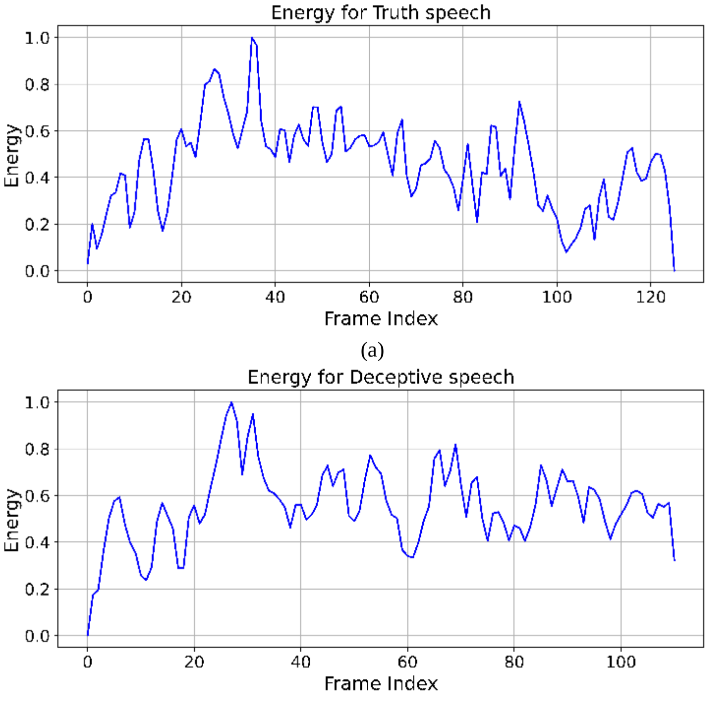
Figure 5 ZCR for all speakers: (a) for truth speech and (b) for lying speech

- *Energy*

Energy in the context of audio signals refers to the intensity of sound perceived by the human ear. The energy of a signal is contingent on the amplitude of the wave. In simpler terms, if the amplitude of the signal is substantial, the sound is perceived as loud [18]. Furthermore, short-term energy is expected to demonstrate notable variability across consecutive speech frames. This suggests that the energy envelope is likely to experience rapid fluctuations between high and low energy levels. This phenomenon can be attributed to the presence of subtle phonemes in speech signals and brief intervals of silence between syllables [14].

For that reason, it was utilized in this investigation, which means it can monitor variations in intensity or volume over time, and these changes in energy may be a symptom of vocal differences between honest and fraudulent speech.
To compute energy, assume that xi(n), n = 1..., WL is the sequence of audio samples for the ith frame, and WL is the frame length. The equation 9 is used to compute the short-term energy. And figure 6 depicts the average Energy for all speakers as determined by the average total frames for truthful and lying speech. where the x-axis indicates the average Energy of samples and the y-axis shows the number of frames[14].

$$E(i) = \sum_{n=1}^{W_L} |X_i(n)|^2 \tag{9}$$



(a)

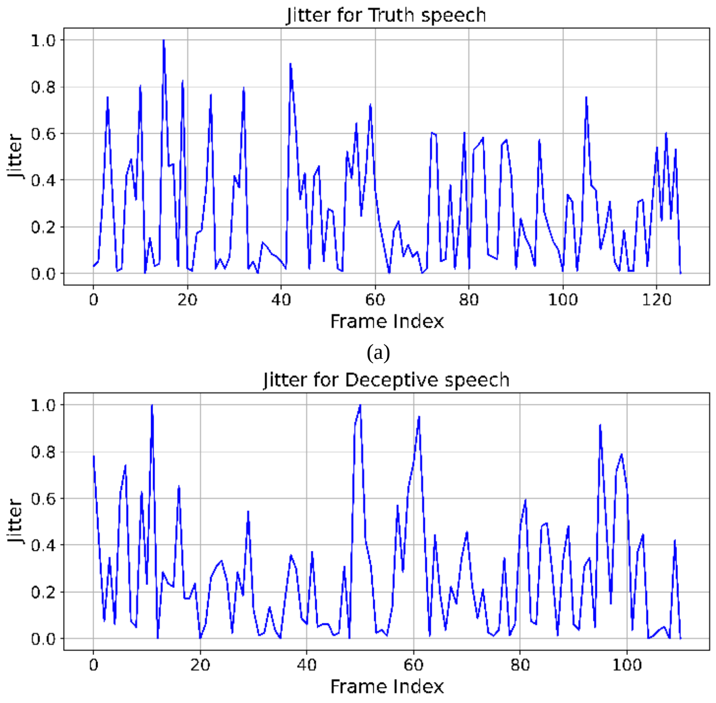
(b)

Figure 6 Energy feature for all speakers: (a) for truth speech (b) for lie Speech

- *Jitter*

There are two sorts of jitter that we are faced, in the processing of speech. Here, they describe them as micro-jitter and macro-jitter.

In our study, we introduce micro-jitter, referring to the minute alterations in the timing of the samples. These small changes may result in an oversight in the signal reconstruction or representation[19]In simpler terms, involuntary voice variations are identified through instinctive shifts in the fundamental frequency over a brief timeframe. Essentially, jitter denotes a disturbance or fluctuation in the pitch of the voice [4]. Figure 7 illustrates the average jitter derived from the average total truth and lying speech frames. It is worth mentioning that these frames represent all speakers. where the x-axis represents the average jitter of samples and y-axis represent the number of frames.



(a)

(b)

Figure 7 illustrates the average Jitter calculated of average total frames (a) truth speech and (b) lying speech

2- *Spectral Domain Features*

Spectral domain feature، also referred to as frequency domain features, provide insights into the spectral distribution of the signal. These characteristics are unveiled through the discrete Fourier Transform, which divides the output into two components. The first component, magnitude, measures intensity and indicates the frequency's contribution to the input signal. The second component reflects the phase of the signal. Additionally, the Discrete Cosine Transform represents a short-time signal as a weighted sum of cosines with real values. The Fast Fourier Transform, capitalizing on computational efficiency in the DFT equation, stands as an excellent method for computing DFT outcomes [16]. The features that were used in this work are MFCC, Spectral Entropy, Spectral rolloff, Spectral Centroid and Pitch.

- *Mel-Frequency Cepstrum Coefficients (MFCCs)*

MFCC, one of the most crucial audio features, finds application in various audio domains such as speech recognition, speaker recognition, and sound categorization. The parameters of MFCC, as defined by Davis and Mermelstein in 1980, have been shown to effectively capture essential auditory information. This feature has demonstrated success across a diverse array of categorization systems [16]. The technique via which the MFCC is determined can be stated as follows:

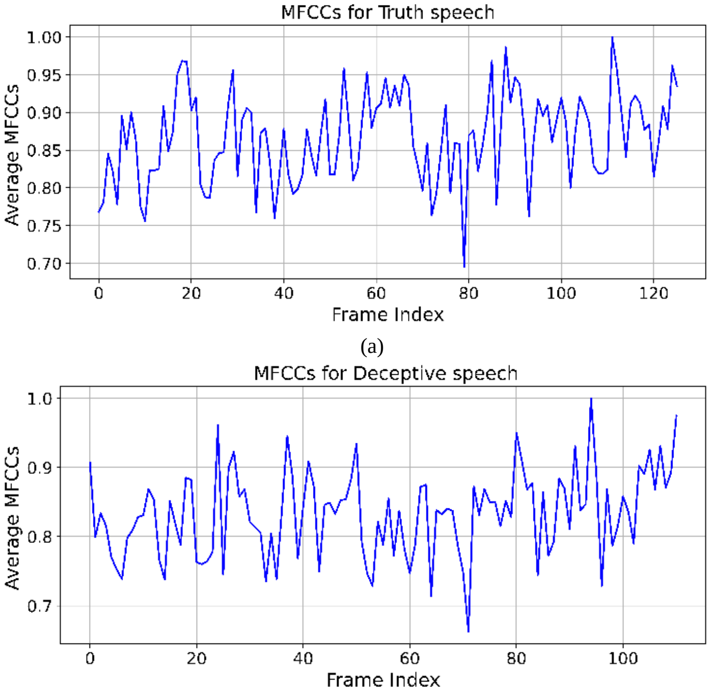
- For every frame, the Fast Fourier Transform (FFT) is computed to convert the time-domain data into the frequency domain.
- The magnitude spectrum is generated by obtaining the absolute value.
- The mel frequency scale, a pitch measurement employed to convert frequency in hertz into the equivalent value in mel ,is measured according to Equation 10.

$$Mel = 1127.0 log\left(1 + \frac{f(HZ)}{700}\right) \qquad (10)$$

- The reduced spectrum is evaluated by applying a triangular Mel filter bank.
- Ultimately, MFCC is obtained by conducting the Discrete Cosine Transform (DCT) of the logarithmically reduced energy spectrum.

$$C_{co} = \sum_{j=1}^{N_f} \log(E_j)\, cos\left[(j - \frac{1}{2})\frac{i\pi}{N_f}\right] \qquad (11)$$

Where $E_j$corresponds to the spectral energy estimated within the range of the jth Mel filter، Nf indicates the total number of Mel triangular filters in the bank, Nc is the total number of cepstral coefficients and C_{co}which are retrieved from each window frame. The default value for Nc is 12. , but in this study, the number of coefficients was set to 23. Figure 8 illustrates the average MFCCs calculated for average total frames for truth and lying speech. It is worth to noting that these averaged framed represents all speakers.



(a)

(b)

Figure 8 Average Mfccs feature for all speakers: (a) for truth speech (b) for lie speech

- *Spectral Entropy (SE)*

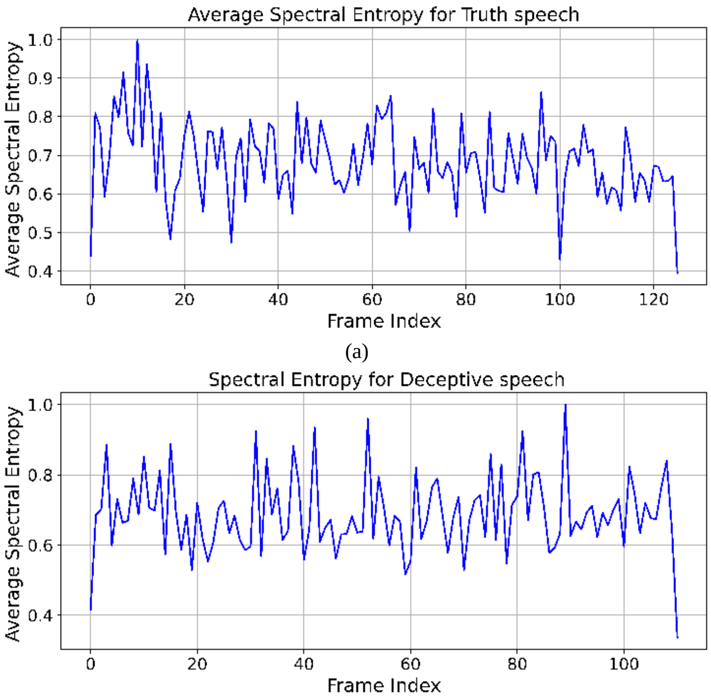
Spectral entropy, a prominent audiofeature, finds application in various fields such as automatic voice recognition, speech detection, and emotion recognition. It has been employed to differentiate between clean and noisy speech, with pure speech exhibiting a lower degree of spectral entropy [16]. The calculation involves initially dividing the spectrum of the short-term frame into L sub-bands (bins).The energy $E_f$of the f^{th}sub-band where f = 0, . . ., L − 1, is then normalized by the total spectral energy, that is nf where is determined according to equation 12 [16].

$$nf = \frac{E_f}{\sum_{f=0}^{L-1} E_f} \qquad (12)$$

Ultimately, the computation of normalized spectral energy entropy (*nf*) it computed according to equation 8 .

$$H = -\sum_{f=0}^{L-1} n_f . log_2(n_f) \qquad (13)$$

Figure 9 displays the average (SE) calculated of average total frames for truth and lying speech. It is worth remembering that these averaged frames reflect all speakers.



(a)

(b)

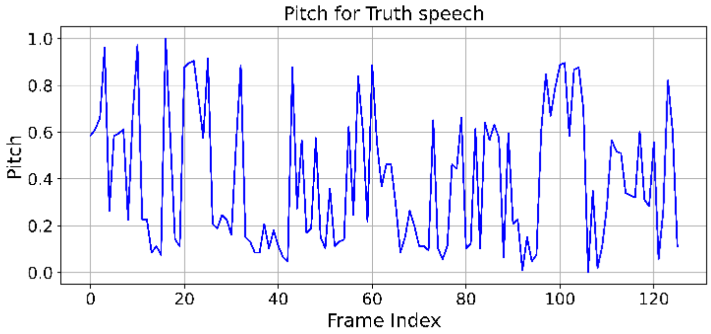
Figure 9 Average (SE) feature for all speakers: (a) for truth speech (b) for lie speech

- *Pitch*

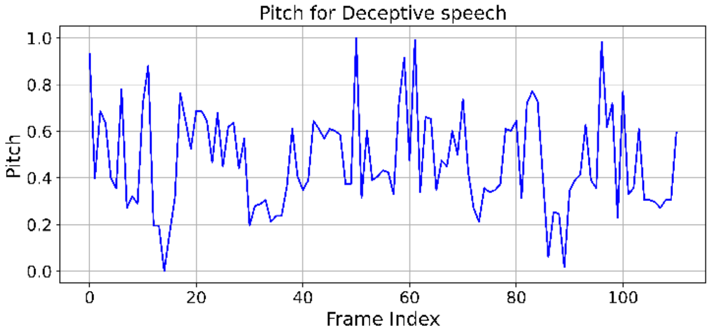
Pitch, represented by the fundamental frequency (F0), is a fundamental element crucial for detecting harmonics. Its significance extends to the segmentation process, as well as the analysis and synthesis of both speech and music. Generally, only speech and harmonic music exhibit a distinctly defined pitch [16].

Additionally, fundamental frequency (f0) serves as a representation of the pitch of a sound, enabling its characterization as high or low. Moreover, it can convey information about the amplitude of the sound, allowing the distinction between strong and weak sounds [4].

Figure 10 illustrates the average Pitch derived by the average total frames for truth and lying speech. It is worth remembering that these average frames reflect all speakers.



(a)



(b)

Figure 10 Pitch feature for all speakers: (a) for truth speech (b) for lie speech

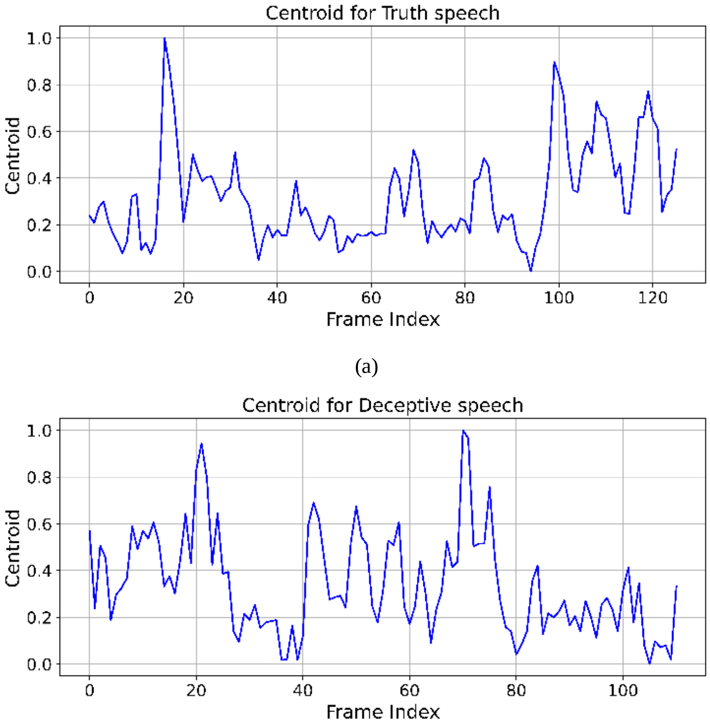
- *Spectral Centroid (Ci)*

The spectral centroid, often defined as the center of gravity of the spectrum magnitude, plays a crucial role in identifying the point in the spectrum where the highest energy is concentrated within the audio frame. Consequently, Ci serves as a fundamental measure of the spectral shape of the frame, providing an essential evaluation of its brightness.[16]. The calculation of the spectral centroid (SC) for the ith frame can be determined as outlined in Equation 14.[14].

$$C_i = \frac{\sum_{k=1}^{wf_L} kX_i(k)}{\sum_{k=1}^{wf_L} X_i(k)} \quad (14)$$

Where wf_Lbe the number of coefficients that are used in the computation, let Xi(k), k = 1, . . . represent the magnitude of the DFT coefficients of the ith audio frame.

Figure 11 displays the average (Ci) obtained from the average total frames for truth and lying speech. It is worth remembering that these average frames reflect all speakers.



(b)

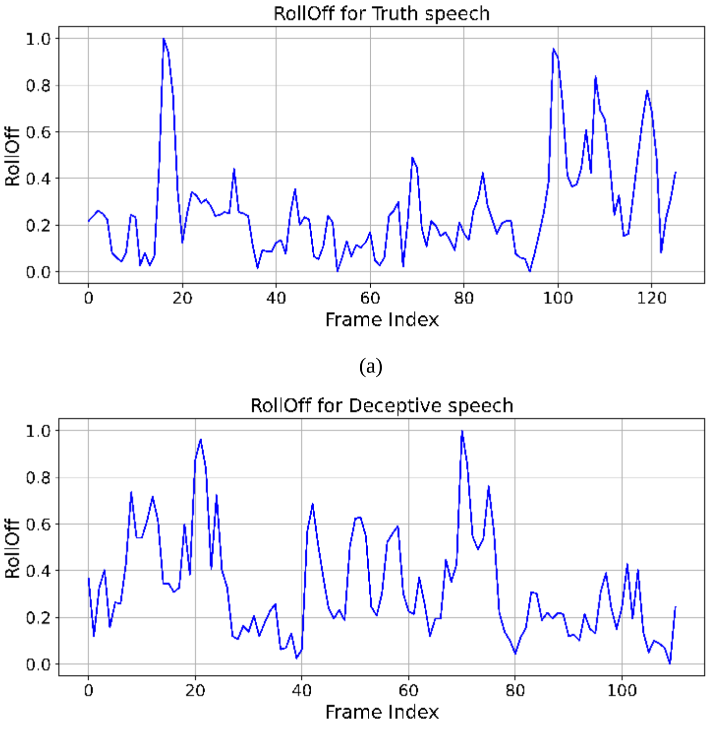
Figure 11 (Ci) feature for all speakers: (a) for truth speech (b) for lie speech

- *Spectral Rolloff*

This particular feature is characterized as the frequency below which a designated proportion (usually around 90%) of the magnitude distribution of the spectrum is concentrated. It serves as a spectral shape descriptor for an audio signal and proves valuable in discerning between voiced and unvoiced sounds [14]. Additionally, if the energy of a speech signal containing emotional information is found within a specified range of frequencies, the spectral rolloff feature can be estimated using Equation 15.

$$\sum_{k=1}^{m} X_i(k) = C \sum_{k=1}^{Wf_L} X_i(k) \qquad (15)$$

m^{th} DFT coefficient corresponds to the spectral rolloff of the i^{th} frame and C is the adopted percentage (user parameter).
Figure 12 presents the average Rolloff calculated from the average total frames for truth and lying speech. It is crucial to mention that these average frames represent all speakers.



(a)

(b)

Figure 12 RollOff feature for all speakers: (a) for truth speech (b) for lie speech

E. Features Selection

Feature selection procedures aim to eliminate non-useful information, thereby reducing the complexity of the final model. The ultimate objective is to create a streamlined model that is faster to compute, with minimal or no compromise in prediction accuracy. To achieve such a model [20]. In simpler terms, feature selection is the process of choosing specific features from the extracted set to eliminate redundancy and irrelevant information. This helps reduce processing time, especially when dealing with extensive datasets that would otherwise require more time for computation [18].
In this research, the feature importance approach employing random forest (RF) was applied as a feature selection. Well, we train a random forest model incorporating all characteristics. Then, we utilize this model to find the most relevant features. Next, we construct a new feature matrix that comprises just these features. Figure (13) show the important feature for real time dataset.

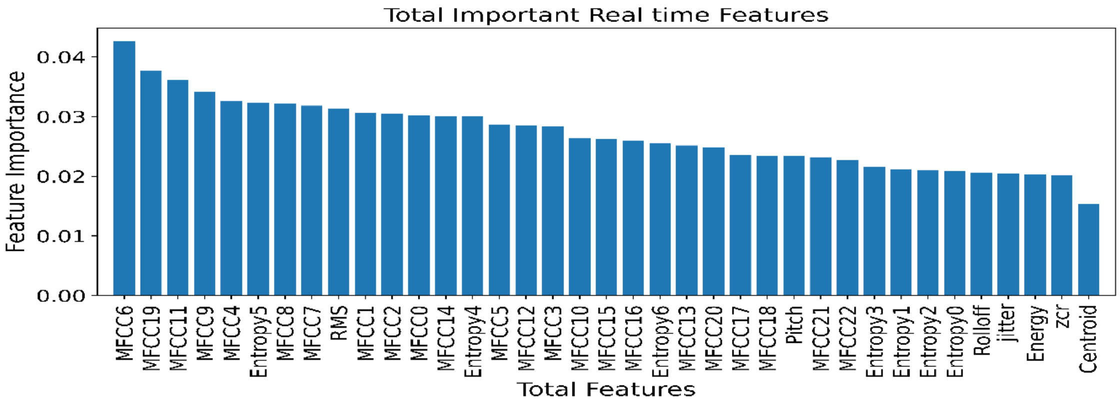


Figure 13 Total important real time features

F. Machine Learning

Machine Learning is an artificial intelligence technology that enhances the process of feature extraction to discern meaningful classifications within a dataset. The two primary methods of classification are unsupervised classification, often referred to as clustering, and supervised classification, also known as discrimination. These approaches are applied in various fields, including physics, mathematics, statistics, engineering, artificial intelligence, computer science, and the social sciences [16].

- Random Forest

The primary notion of random Forests is ensembles of slightly different trees formed by training on random training subsets. To achieve a more comprehensive and accurate goal with a lower error rate and increased noise robustness than what a single module could provide, an ensemble technique was employed. This technique involved integrating a set of existing modules, each addressing the same classification problem [16].

G. Training and Testing

In this section, we will delve into the training and testing procedures of the model. Following the feature extraction process and the application of feature selection techniques, the data was prepared for the model's training and testing phases. The data were labelled, with a value of 1 assigned to deceptive speech and 0 assigned to truthful speech. The resulting matrix had dimensions of 113,048 rows and 37 columns, representing the extracted features and their corresponding labels. The chosen classifiers for this experiment were random forests. The dataset was partitioned into a training set and a testing set, with a ratio of 0.8 for training and 0.2 for testing.

The training set, constituting 80% of the dataset, was employed for model training, while the remaining 20%, referred to as the testing set, was utilized to evaluate the performance of the trained models. Throughout the training process, Random Forest (RF) generated 100 estimators, representing individual decision trees. The predictions of these trees were then amalgamated to formulate the final classification decision.

Moreover, to tackle the problem of overfitting, where a model excels on the training data but struggles with generalization to new data, cross-validation was implemented. Cross-validation involves evaluating the model on various subsets of the data, known as folds. In this particular investigation, a 5-fold cross-validation approach was employed. This method aids in assessing the model's ability to generalize by testing its performance on unseen data, thereby providing a more robust evaluation of its effectiveness.

IV. Experimental results

The trained random forest classifier was assessed on the test set using varying numbers of chosen features .and got this result as displayed in table 2.

Table 2. Show the accuracy of model using random forest

Number of important features	Accuracy
10	0.726890756
15	0.76302521
20	0.784166298
25	**0.791685095**
30	0.787527643
35	0.78659885

Also, Figure 14 shows that the best results were obtained by using classifiers. Were the x-axis representing the number of features and y-axis represent accuracy of classifiers**.**

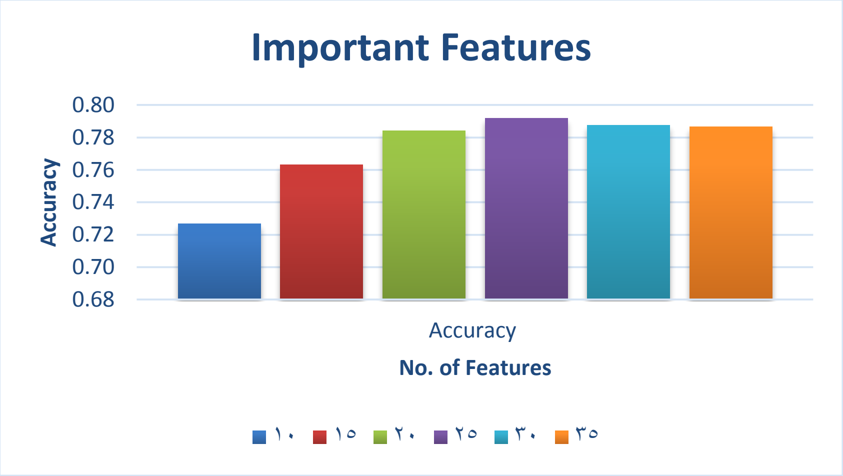


Figure 14 Accuracy of the proposed model

The results indicate a trend of increasing accuracy with the growing number of selected features. The model achieved its highest and most accurate performance when utilizing 25 features. This implies that the additional information captured by these features contributes positively to the classification performance. To provide further insights into the model's performance, the confusion matrix was employed.

The confusion matrix provides a comprehensive evaluation of the model. The performance of the trained random forest classifier was assessed across various numbers of selected features, with the corresponding accuracy values presented in Table 1. To delve deeper into the model's performance, a confusion matrix was generated for the optimal number of chosen features (25 features). After 25 characteristics were selected, the model achieved an accuracy of 0.79. The relevant confusion matrix is given in Table 3.

Table 3. Show the confusion matrix for the lie and truth speech

Predicted Actual	True Speech	Lie Speech
True Speech	41.88	10.45
Lie Speech	10.38	37.29
Performance Measurements at 25 features		
Recall	0.80	0.78
Precision	0.80	0.78
F1 Score	0.80	0.78

V. CONCLUSION

In conclusion, this study introduces a novel approach to deception detection, focusing on individual spoken utterances and leveraging voice stress analysis. It addresses the limitations of existing systems, particularly those relying on psychological and behavioral measurements like polygraph testing, which have faced criticism due to reliability and validity concerns. While emerging technologies like brain imaging and machine learning show promise, they are still in early research stages and require further testing for robust deception identification.

The proposed methodology underwent validation using a "real-time data" dataset, capturing sounds in a real-world environment through a wireless microphone. The audio data was segmented into 153 samples for lying speech and 161 samples for truthful speech. The investigation involved extracting temporal domain and spectral information from audio signals, utilizing overlapping frames, and implementing a Random Forest (RF)-based classification system. The experimental results reveal an overall classification accuracy of 79% in distinguishing between lie and truth classes. Moreover, the data suggest that deceptive speech displays higher unpredictability compared to truthful speech, as illustrated in the provided feature figures.

These findings highlight the potential utility of the proposed approach in detecting dishonesty in individual speech utterances, with MFCC emerging as a crucial feature. However, further research and testing are imperative to assess its effectiveness across diverse datasets, real-world scenarios, and various demographic groups, ensuring its robustness in different contexts.

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.comhttps://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] 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.

[7] 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.

[8] F. M. Talaat, “Enhanced Deep Neural Network for Lie Detection by using Voice Stress,” 2021. Accessed: Sep. 20, 2022. [Online]. Available: https://www.researchgate.net/publication/366606445_Deep_Neural_Network_for_Lie_Detection_Based_on_Voice_Stress

[9] X. S. H. F. B. W. and T. Y. W. Wang, “‘Machine Learning based Deceptive Speech Detection,’ 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 2023, pp. 1-6,”, doi: 10.1109/ICSCDS56580.2023.10104721.

[10] 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, pp. 1–22, May 2023, doi: 10.3390/brainsci13050725.

[11] P. C. Loizou, “SPEECH ENHANCEMENT Theory and Practice Second Edition,” 2013.

[12] Saeed V. Vaseghi, P. Rayner, B. Milner, C. Ho, and A. Chen, “Advanced Digital Signal Processing and Noise Reduction,” 2000.

[13] D. Y. Mohammed, P. J. Duncan, M. M. Al-Maathidi, and F. F. Li, “A system for semantic information extraction from mixed soundtracks deploying MARSYAS framework,” in *Proceeding - 2015 IEEE International Conference on Industrial Informatics, INDIN 2015*, Institute of Electrical and Electronics Engineers Inc., Sep. 2015, pp. 1084–1089. doi: 10.1109/INDIN.2015.7281886.

[14] Theodoros Giannakopoulos and Aggelos Pikrakis, “AUDIO ANALYSIS: A MATLAB Approach,” 2014. Accessed: Mar. 15, 2023. [Online]. Available: htpp://booksite.elsevier.com/9780080993881

[15] 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.

[16] D. Yehya Mohammed, “OVERLAPPED SPEECH AND MUSIC SEGMENTATION USING SINGULAR SPECTRUM ANALYSIS AND RANDOM FORESTS,” Manchester ,School of Computing, Science and Engineering University of Salford, 2017.

[17] 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.

[18] 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.

[19] Homayoon Beigi, “Fundamentals of Speaker Recognition,” Yorktown Heights ,NY , USA, 2011. doi: 10.1007/978-0-387-77592-0.

[20] A. Zheng and A. Casari, “Feature Engineering for Machine Learning PRINCIPLES AND TECHNIQUES FOR DATA SCIENTISTS,” 2018.