Revista ARACÊ
ISSN: 2358-2472

DECIPHERING AUDIO DEEPFAKES: AN INVESTIGATION BY FREQUENCY ANALYSIS

https://doi.org/10.56238/arev6n3-375

Submitted on: 29/10/2024 **Publication date:** 29/11/2024

Maria Stella Simões Piccolo[1], João Henrique Gião Borges[2] and Fabiana Florian[3]

ABSTRACT
This research examines the emerging technology of audio DeepFakes, which presents itself as a contemporary challenge. To address this issue, ElevenLabs was used for voice cloning, along with scripts based on Deep Learning and Machine Learning for the creation of spectrograms and evaluation of patterns in the frequencies of the voices. The findings indicated that the evaluation of the graphs of each voice is efficient to differentiate the artificial voice from the real one.

Keywords: Artificial Intelligence. DeepFakes. Deep Learning. Audio DeepFakes. ElevenLabs. Spectrogram.

[1] Undergraduate student of the Information System Course at the University of Araraquara - UNIARA Araraquara-SP
E-mail: mssp1717@outlook.com
[2] Advisor
Professor of the Information System Course at the University of Araraquara-UNIARA. Araraquara –SP
E-mail: jhgborges@uniara.edu.br
[3] Co-advisor
Professor of the Information Systems Course at the University of Araraquara – UNIARA. Araraquara – SP
E-mail: fflorian@uniara.edu.br

Revista ARACÊ
ISSN: 2358-2472

INTRODUCTION

In the contemporary scenario, characterized by technological advances and the proliferation of Artificial Intelligence (AI), a new technology emerges within the *DeepFakes* complex: *Audio DeepFakes*. Although its development was initially to bring an improvement in the daily routine, with the use of *audiobooks,* this technology has been used and exploited for a manipulation of reality, representing a threat to public safety.

DeepFakes are AI products that merge, combine, replace, or even overlay or replace audio and images to create fake files. This article focuses on the study of Audio *DeepFakes*, which has the ability to generate fake voices and clone existing voices, and its analysis involves several parameters, including frequency, intensity, and other sound attributes.

This article aims to study the operation of Audio *DeepFakes* by comparing the frequencies of the voices that have been cloned and the real voice. The use of ElevenLabs tools to generate counterfeit and cloned voices and Spectrograms for a graphical visualization of the frequency of voices, contributed to the accurate identification of Audio *DeepFakes*, distinguishing one voice from another.

According to the UOL Notícias website (2024), the advance in the use of *DeepFakes*, especially in audio manipulation, generates growing concerns about their consequences. Many people still do not have enough understanding of this topic, which can lead to problematic situations, such as coups and, in recent cases, manipulation in elections. On this site some uses of *DeepFakes are reported*, in a specific case one of the targets was Joe Biden, current president of the United States, who experienced this problem of *audio DeepFakes* in an election in the United States, where an audio simulating his voice was circulated, illustrating the threat to the integrity of the democratic process.

In view of these challenges, there is a need to understand Audio *DeepFakes* and to develop methods of analysis or improve existing ones, allowing the possibility of discerning authentic and counterfeit voices, contributing to an attempt to reduce the risks associated with misuse.

According to Almutairi, Z., & Elgibreen, H. (p.2, 2022). "The detection of audio DeepFakes has therefore become an active area of research with the development of advanced Deep Learning (DL) techniques and methods. However, with such advancements, current DL methods are struggling, and further investigation is needed to understand in which area of audio DeepFakes detection needs further development. In

Revista ARACÊ
ISSN: 2358-2472

addition, a comparative analysis of current methods is also important, and, to our knowledge, a review of imitated and synthetically generated audio detection methods is lacking in the literature."

The hypothesis of this article is that the use of the ElevenLabs and Spectrograms tools will allow to efficiently distinguish some issues that will guide the research are that in each voice there are different levels of frequencies, through frequency it is possible to perform a comparative analysis between each voice. To test this hypothesis, a tool such as Spectrograms will be used, which serves as a graph that demonstrates the frequencies of voices through sound wave drawings.

To start this research, it is first necessary to understand what *DeepFakes is?* How does *Audio DeepFakes* work? How is voice frequency analyzed? Will the Spectrogram be different between a real voice and a cloned voice? To answer such questions, different articles were read about what this *DeepFakes* technology is, it will be studied how this *Audio DeepFakes* works, there will be the use of the ElevenLabs tool to clone the voices, to carry out the analysis, a script will be developed that will receive these audio files and represent them in the form of Spectrograms, That is, in a graphic form, where waves representing frequencies will be formed to make the comparison.

LITERATURE RESEARCH

This section covers the concepts of Artificial Intelligence, *Deep Learning*, *DeepFakes*, *Audio DeepFakes*, (2.3) Voice Frequency, (2.4) ElevenLabs' Potential in Voice Cloning, (2.5) Spectrogram and Frequency Relationship.

ARTIFICIAL INTELLIGENCE (AI)

To have a better understanding of the topic of this article, it is first necessary to understand its structure, as already reported, *DeepFakes* a technology formed through artificial intelligence using a methodology called *Deep Learning*, with this in mind, how can we describe AI?

According to Jaime Simão Schiman (2021), "The AI domain is characterized by a collection of models, techniques, and technologies (search, reasoning and knowledge representation, decision mechanisms, perception, planning, natural language processing, uncertainty treatment, machine learning) that, alone or grouped, solve problems of this nature.".

Author Priscila Mello Alves (2020), on the other hand, says that "The definitions of AI found in the scientific literature, as a discipline of human knowledge, are categorized, either empirically with the formulation of hypotheses and experimental confirmations or theoretically involving mathematical calculations. From the perspective of the empirical category, there is a perspective of systems thinking like human beings, enabling learning from experiences, while in the theoretical approach, logical actions are expected that include the ability to deduce and infer about new relationships".

In view of these two definitions, it can be seen that AI is a technology, whose definitions vary greatly, but that at its base, we can call it a technology that encompasses several models, which in its surroundings can behave like human, in the sense that it can adhere to learning techniques such as *machine learning* or *deep learning,* cognition and decision-making such as neural networks, among others, becoming a way to optimize processes, automate functions, and reduce complexity.

Deep Learning –

According to IBM's website, "*Deep learning* is a subset of machine learning, which is essentially a neural network with three or more layers. These neural networks attempt to simulate the behavior of the human brain, although far from matching its capacity, allowing it to "learn" from large amounts of data. While a neural network with a single layer can still make rough predictions, additional hidden layers can help optimize and refine accuracy."

> Deep Learning currently offers an important set of methods to analyze signals such as audio and speech, visual content, including images and videos, and even textual content." (Ponti Moacir and Costa Gabriel, 2017, 1, p. (63-93), "How Deep Learning works").

With the descriptions of these authors, it is noted that AI is not a consolidated technology by itself, it encompasses several others for certain functions, in the case of this article, *Deep Learning* will be something present, since we will have to carry out an analysis regarding the recorded and created voices.

DEEPFAKES

> "The term DeepFakes refers to synthetic media in which images or sounds captured from certain people are replaced by those of others through advanced machine learning and Artificial Intelligence (AI) techniques, with the purpose of manipulating visual and/or sound content, with enormous potential for falsifying reality. Celebrities

> and politicians have been the preferred targets of these manipulations, which have become exponentially popular even through free cell phone applications." (Fonseca, p.106, FANAYA, February/2021).

"DeepFakes are essentially false identities created with deep learning through massive use of data" (Spencer, Michael K., translation: Gabriela Leite, 2019, other words).

According to the CNN website (2024), *DeepFakes* occur when artificial intelligence (AI) fuses, combines, replaces, or overlays audios and images to create fake files in which people can be placed in any situation, saying phrases never said or taking actions never taken. The content can be humorous, political or even pornographic. There are countless possibilities: face swapping, voice cloning, lip-syncing to an audio track different from the original, among others. The technique commonly distorts the perception of an individual in a given situation.

By following some definitions, we understand that *DeepFakes* come from the junction of *"Fake News"* with *Deep Learning* technology through artificial intelligence, where through its use, we can create synthetic content, which can be image, video and audio, as explained in the theme, where the stored data is used by a *Machine Learning* and *Deep Learning algorithm.*

Deepfakes de áudio

Among all this technology advancement, increasing use of artificial intelligence, and consequently new emergence of *DeepFakes, Audio* DeepFakes *emerges, which is the central focus of this article.*

According to Almutairi, Z., & Elgibreen, H. (p-3,2022). AI-synthesized tools have recently been developed with capabilities to generate compelling voices. However, while these tools were introduced to help people, they were also used to spread misinformation around the world using audio, and their malicious misuse led to the fear of *"Audio DeepFakes."*

> "Audio DeepFakes focus on generating the target speaker's voice, using deep learning techniques to portray the speaker saying something that was not said. Fake voices can be generated using text-to-speech (TTS) or speech conversion (VC)." (Masood, M., Nawaz., Malik, K.M., Ali, J. & Irtaza A. Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward. p (1-54), Cornell University, 2021.)

Revista ARACÊ
ISSN: 2358-2472

Audio *DeepFakes* works through a technology where it is possible to clone your voice, through a voice conversion that is based on some encoders, such as neural network that will compress the input data into a compact and internal representation, and ends up learning to decompress this data from this representation to restore the original data, In this way, it learns to present data in a compact format while highlighting important information.

Focusing on studies on Audio *DeepFakes*, we noticed the use of advanced artificial intelligence techniques, specifically *Deep Learning* and *Machine Learning*, to clone a person's speech in a natural and convincing way. The platform used for voice cloning and the steps for cloning will be described below:

VOICE FREQUENCY

> To produce the voice, first, the air being exhaled must pass through the vocal cords and vibrate them (try to speak during inspiration and see how difficult it is). The vocal cords are musculomembranous structures that form an inverted V (...). If the vocal folds are relaxed, air will pass freely without producing sound. When we want to speak, the brain sends nerve commands to the muscles that control the tension of the vocal cords and the air will pass vibrating through the folds. The vibrations are very fast, whose frequency is more than 100 times per second. While the vibration occurs, several structures (mouth, throat and nasal passage) resonate producing several harmonic waves and amplifying the sound. (Nishida, M. S., Weber, S. A. T., Oliveira, F. A. K. de, & Troll, J. Human Voice).

Understanding the functioning of the voice, we have the perception that the sound we emit can be distinguished according to the movements of the vocal folds, which can generate some different characteristics. In a voice analysis, we can distinguish one voice from others, by pitch, intensity, and amplitude. At the height we check if the sound is high or low, which is determined by the frequency of the wave, the timbre is distinguished through the sound waves and the intensity is measured by the volume, where a more intense, strong or weak sound comes from a factor called wave amplitude, where the greater the amplitude of a wave, the greater the pressure it will exert in the air, which causes our eardrums to vibrate more intensely.

It is worth mentioning that we do not hear our voice in the same way as someone else, because those who are listening to us do not hear us through the airway, so our own voice does not undergo pressure changes, which concludes that the recorded voice is considered the true representation of our voice.

Revista ARACÊ
ISSN: 2358-2472

ELEVENLABS' POTENTIAL IN VOICE CLONING

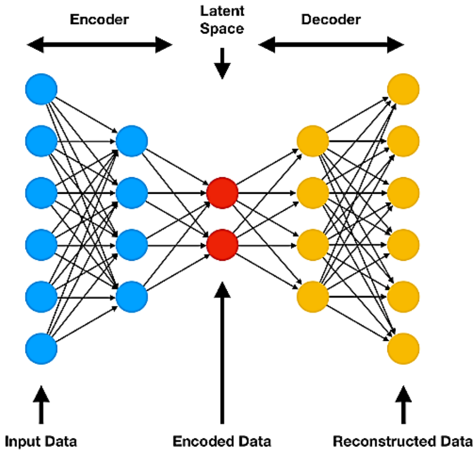
The use of tools such as ElevenLabs contributed a lot to this article, due to its ability to provide the possibility of voice cloning, in view of its growth, many people started to use it for trolling games, creation of synthetic voices and even for scam attempts. Although ElevenLabs is a paid platform, it provides free features such as creating synthetic voices.

According to information from the Kaspersky Daily website (2023). "Voice conversion is based on automatic encoders, a type of neural network that compresses the input data (part of the encoder) into a compact internal representation and then learns to decompress it from that representation (part of the decoder) to restore the original data.

In this way, the model learns to present the data in a compressed format while highlighting the most important information."

As suggested by the image available on this same site. Found in figure 1.

Figure 1: Automatic encoder diagram:

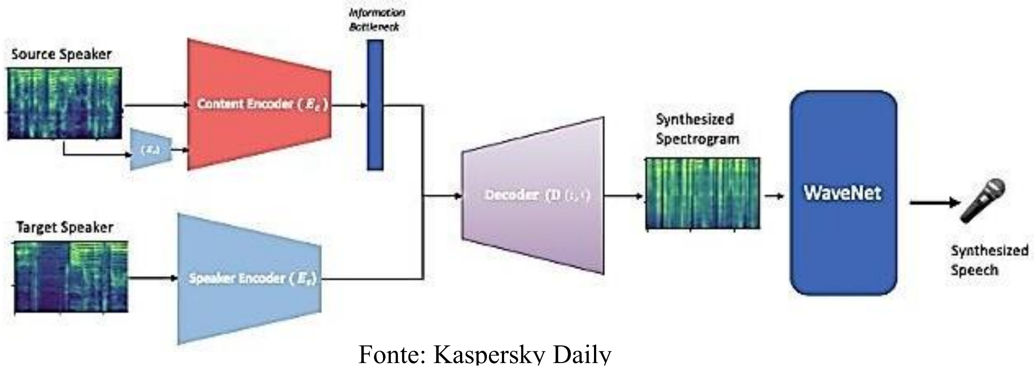


Fonte: Kaspersky Daily (2023)

The process of creating these *DeepFakes* usually involves at least two audio recordings that are fed into the model. The second recording is converted to look like the first, after which the content encoder determines what was said from this first recording. The encoder extracts the characteristics of the voice, such as timbres, intonation, accents, frequencies, amplitudes from the second recording, these representations are combined to produce the result through the decoder.

Revista ARACÊ
ISSN: 2358-2472

Figure 2- *DeepFakes voice generation process*:



Fonte: Kaspersky Daily

Platforms such as ElevenLabs, among others that offer this possibility of cloning voices, use these processes, along with the use of artificial intelligence algorithms such as *Machine Learning* and *Deep Learning*.

SPECTROGRAM AND FREQUENCY RATIO

Considering that *Audio DeepFakes* has this ability to clone voice, create synthetic audio and understanding that our voice produces a certain frequency, to represent it there is a need to analyze it and the spectrogram is a technique that enables this graphic visualization.

According to the website, Your Physicist, 2024, the Spectrogram is a technique that combines Fourier analysis with time to visualize the spectral characteristics of a signal over time. It graphically represents the intensity of each frequency component as a function of time, providing a clear visual representation of changes in spectral content along the signal.

This technique is widely used in areas such as audio analysis, speech processing, vibration monitoring, and non-stationary signal analysis. The Spectrogram allows you to identify transient changes or short-lived events, as well as provide information about the dominant frequency characteristics at different times of the signal.

This article aims to use the Spectrogram as one of the tools to analyze the frequencies emitted by the audios, in order to have a perspective of the differences, for a better detection of fake audios and cloning.

DEVELOPMENT

In this section it presents: Use of the ElevenLabs platform for voice cloning, conversion of the MP3 file to WAV and a script was developed to generate the frequency

ISSN: 2358-2472

spectrogram, another script to learn the spectrogram patterns and finally a script to analyze the differences between a real voice and a cloned voice.

ELEVENLABS

According to the ComunitIA website (2023), ElevenLabs is an innovative technology company whose main product is audio generated by artificial intelligence. It brings advanced tools for creators and publishers looking to enhance their storytelling by delivering captivating, emotive, and realistic voices.

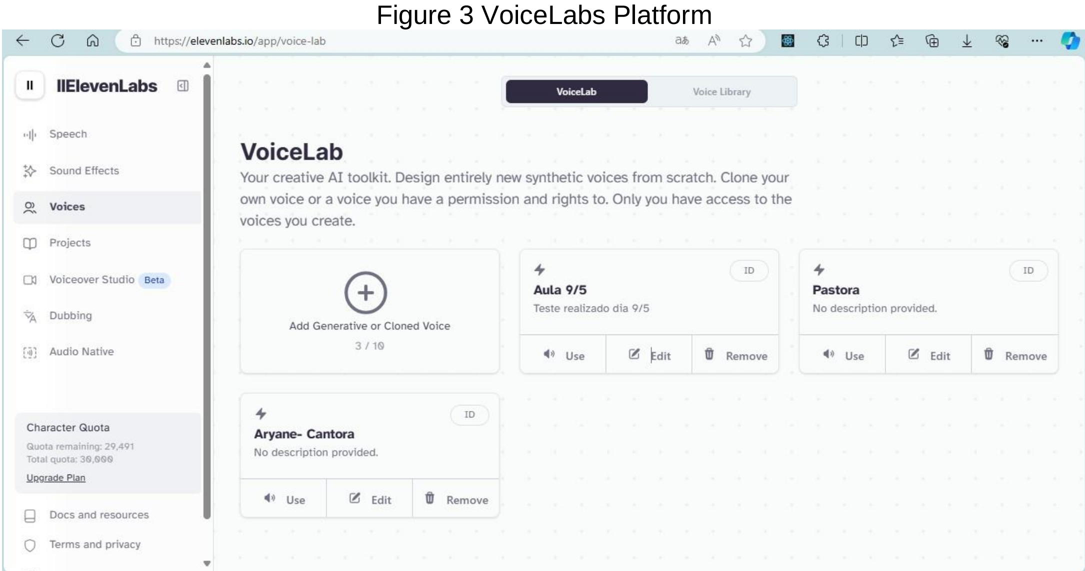
In this work, the ElevenLabs website, the VoiceLabs tool, was used for the cloning of the voices, The process involves recording the desired voices, writing a phrase for the cloned voice, analyzing and improving it to make the audio more natural, and finally, downloading the generated audio.

PROCESS OF CREATING AUDIO *DEEPFAKES* THROUGH THE ELEVENLABS PLATFORM

Data Collection

On the ElevenLabs website, under VoiceLab found in the "Voices" section. To start cloning, one of the previously created libraries was selected and the audios were added. Shown in figure 3.

Figure 3 VoiceLabs Platform



Source: ElevenLabs (2024)

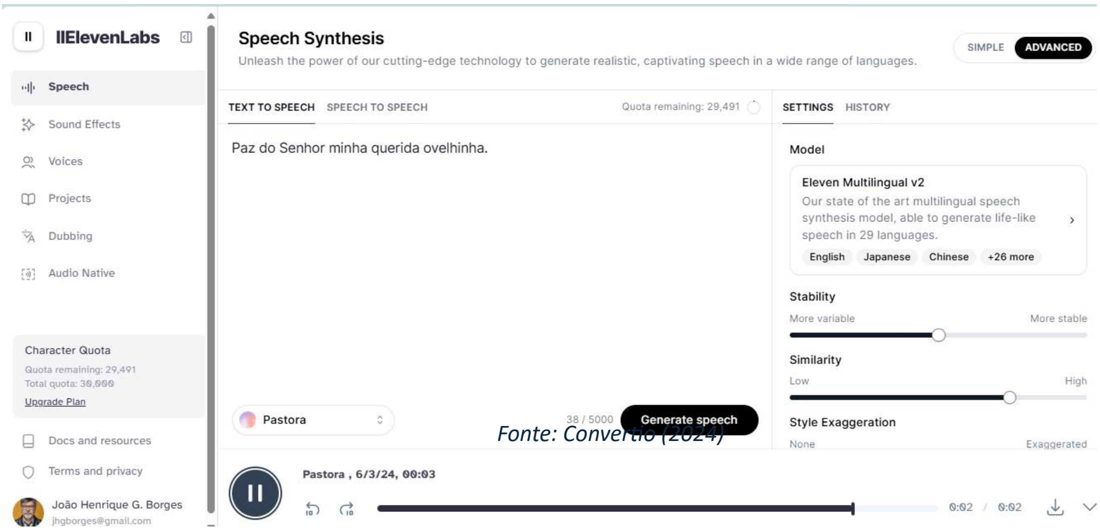
Revista ARACÊ
ISSN: 2358-2472

When you select the voice titled "Pastor", the following interface is displayed. Found in figure 4 and figure 5.

Figure 4- Adding the audios of the pastor's voice

Source: ElevenLabs (2024)

Figure 5 – Inserting a text to be generated with this voice:



Source: ElevenLabs (2024)

After clicking "Generate Speech" shown in Figure 5, the audio is generated. This audio is generated through the *Deep Learning technique* to process the text and then generate the realistic audio, as files with the original voice were inserted, this volume of data was processed and trained to capture nuances, pauses and the natural rhythm of speech. You can adjust the audio so that the voice becomes even more like the natural one. After adjusting, we click on the download button.

Converting MP3 to WAV

After generating the cloned voice, an algorithm created by the author is used to receive the audio files and generate a spectrogram. But for that, it will be necessary to

Revista ARACÊ
ISSN: 2358-2472

convert the MP3 file to the WAV extension. For this, a website called Convertio was used, which does this for free. As shown in Figure 6.

Figure 6 - MP3 Converter for WAV:

Fonte: Convertio

GENERATING SPECTROGRAM THROUGH WAV AUDIO

Figure 7 has the function of generating a graph called spectrogram, by means of the MP3 file converted to WAV extension. In view of the objective of the algorithm, it is necessary to understand its operation, which is based on some steps such as: Importing libraries; Audio file reading; Normalization of audio data; Calculation of the Fourier transform and finally the plotting of the frequency spectrum. This code is found in Figure 7.

Figure 7 – Spectrogram Script

E Espectograma   Version control   main

```
import numpy as np
from scipy.io import wavfile
import matplotlib.pyplot as plt

sample_rate, audio_data = wavfile.read('Voz-Clonada-pastora (1).wav')
audio_data = audio_data.astype(float) / 2**15

fft_data = np.fft.fft(audio_data)
freq = np.fft.fftfreq(len(audio_data), 1/sample_rate)

plt.plot( *args: freq, np.abs(fft_data))
plt.xlabel('Frequência (Hz)')
plt.ylabel('Amplitude')
plt.title('Espectro de Frequência - Voz Clonada Pastora')
plt.show()
```

Source: Author (2024)

ISSN: 2358-2472

IMPORTING FROM LIBRARIES

The imported libraries are numpy, scipy.io.wavfile, matplotlib.pyplot. Shown in Figure 8.

Figure 8 – Importing the libraries

```
import numpy as np
from scipy.io import wavfile
import matplotlib.pyplot as plt
```

Source: Author (2024)

Each library has a specific functionality, with *Numpy* being used for efficient numerical operations, scipy.io.wavfile for reading and writing WAV files, and *matplotlib.pyplot* for creating graphs.

FILE READING

In this part of the algorithm, a reading of the file is performed, using the following concepts: sample_rate; audio_data, wavfile. read(). As shown in figure 9.

Figure 9 – File reading

```
sample_rate, audio_data = wavfile.read('Voz-Clonada-pastora (1).wav')
```

Source: Author (2024)

The wavfile.read() function reads a WAV file and returns two pieces of information: the sample rate (sample_rate) and the audio data (audio_data).

The variable sample_rate – It is the sample rate of the audio, indicating how many samples per second were captured, and the audio_data is an array containing the audio data.

NORMALIZATION OF AUDIO DATA

In this step of the code, the conversion of the MP3 extension to WAV and the normalization of the audio data is carried out. Shown in Figure 10.

Figure 10 – Normalization of audio data and conversion of extensions

```
audio_data = audio_data.astype(float) / 2**15
```

Source: Author (2024)

Revista ARACÊ
ISSN: 2358-2472

The variable audio_data. astype(float) converts the audio data to float type and audio_data / 2**15 normalizes this audio data. In WAV files with 16 bits per sample, the values range from -32768 to 32767. Dividing by 2152^ {15}215 (32768) normalizes the data to the range -1 to 1.

CALCULATION OF THE FOURIER TRANSFORM

In this part of the code, the Fourier transform calculation is performed, whose applicability is to divide something into several sine waves. The Fourier Transform is a powerful tool for converting a signal from the time domain to the frequency domain, allowing the identification of frequency components that are not directly visible in the original signal. This part is found in Figure 11.

Figure 11 – Calculation of the Fourier transform

```
fft_data = np.fft.fft(audio_data)
freq = np.fft.fftfreq(len(audio_data), 1/sample_rate)

```

Source: author (2024)

In this code np.fft.fft (audio_data) calculates the Fourier Transform of the audio data, transforming the signal from the time domain to the frequency domain. The variable fft_data contains the coefficients of the transform.

np.fft.fftfreq(len(audio_data), 1/sample_rate), generates a list of frequencies corresponding to the coefficients of the Fourier transform where len(audio_data) is the number of samples, and 1/sample_rate is the time interval between samples.

FREQUENCY SPECTRUM PLOTTING

This final part of the code is where the frequency spectrum plotting part is developed. Shown in Figure 12.

Figure 12 – Frequency Spectrum Plotting

```
plt.plot( *args: freq, np.abs(fft_data))
plt.xlabel('Frequência (Hz)')
plt.ylabel('Amplitude')
plt.title('Espectro de Frequência - Voz Clonada Pastora')
plt.show()

```

Source: Author (2024)

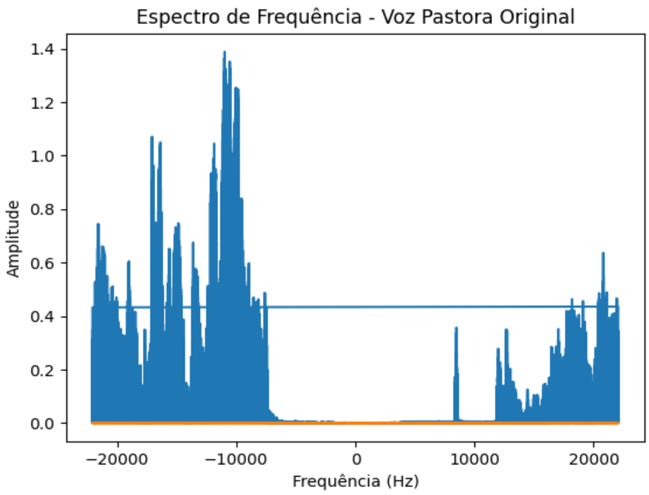
Revista ARACÊ
ISSN: 2358-2472

The plt.plot function (freq. np.abs(fft_data)) plots the frequency spectrum, using the frequencies (freq) on the x-axis and the magnitude of the Fourier transform coefficients (np.abs(fft_data)) on the y-axis. The variable plt.xlabel ('Frequency (Hz)') sets the x-axis label to "Frequency (Hz)". The variables plt.ylabel ('Amplitude') sets the y-axis label to "Amplitude". The plt.title function ('Frequency Spectrum - Cloned Shepherdess Voice'): Defines the title of the graph and finally the plt.show() function will display the graph.

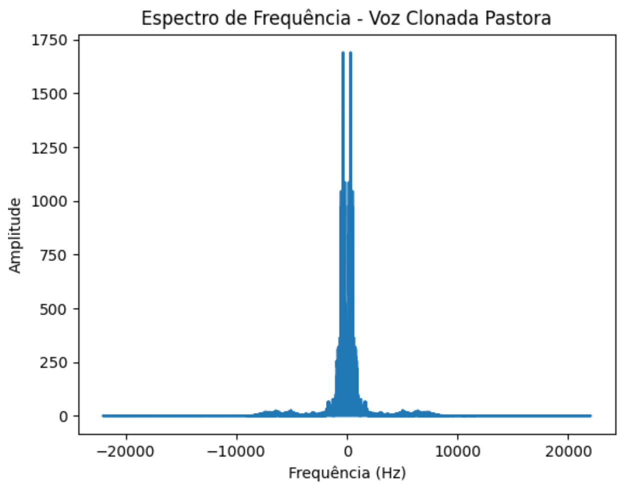
SPECTROGRAM GRAPH

By performing the above procedures, and generating the graphs with the cloned voice and with the original voice, the following results were obtained, Figure 13 and Figure 14.

Figure 13 and Figure 14 – Graph Frequency spectrum – Original voice and cloned voice of the Shepherdess



Source: Author (2024)



Source: Author (2024)

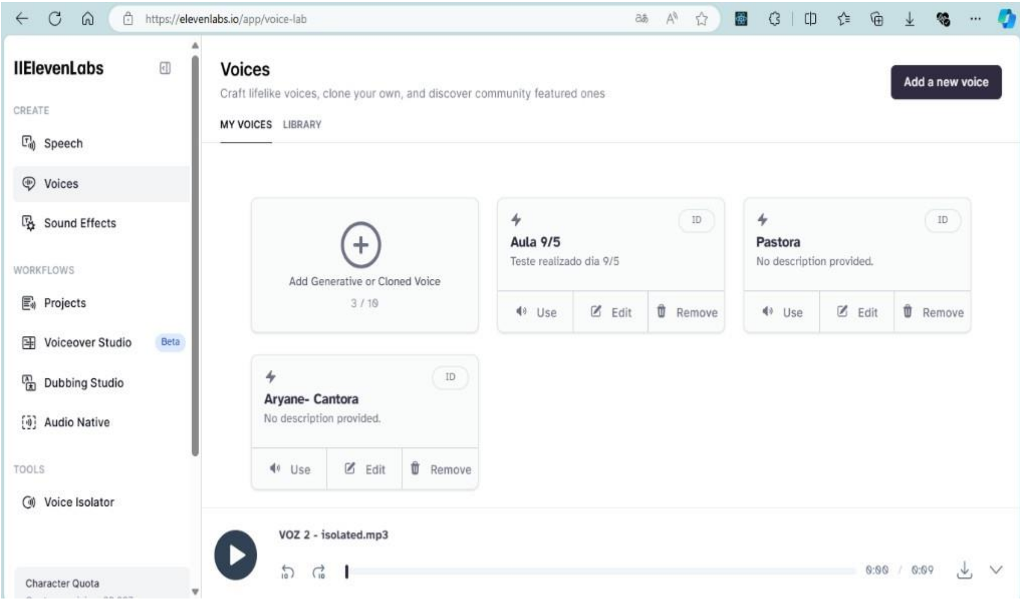
Considering these two images (Figures 13 and 14), we can make the following analysis, evaluating the frequencies of the original voice, as shown in figure 13, a greater diversity is noted in what we call harmonic components, containing moderate amplitude peaks distributed over wide frequency ranges, both positive and negative. The cloned voice has a highly concentrated spectrum, with two prominent peaks very close to 0 Hz and the amplitudes larger. This means that in this case cloning reduced the harmonic richness of the voice, leaving its energy condensed in a restricted frequency range.

Revista ARACÊ
ISSN: 2358-2472

TESTING VOICE TIMBRES

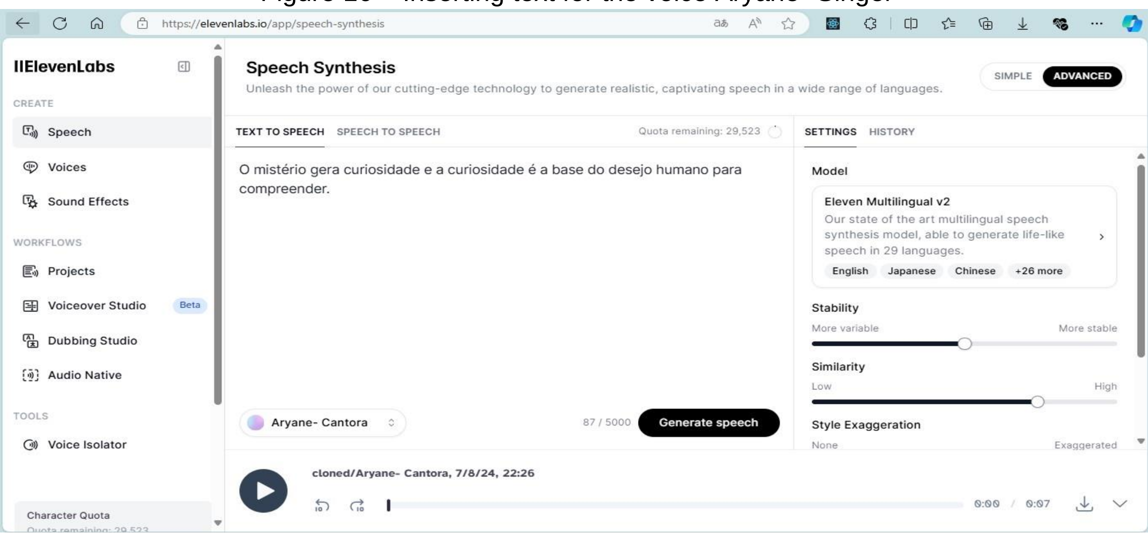
In topic 3.2.1 data collection, in figure 3 VoiceLabs, we used an audio called "pastora", now the voice entitled "Aryane - Singer" will be used in Figure 15 and then the text was inserted in Figure 16.

Figure 15 - Aryane Voice - Singer



Source: ElevenLabs (2024)

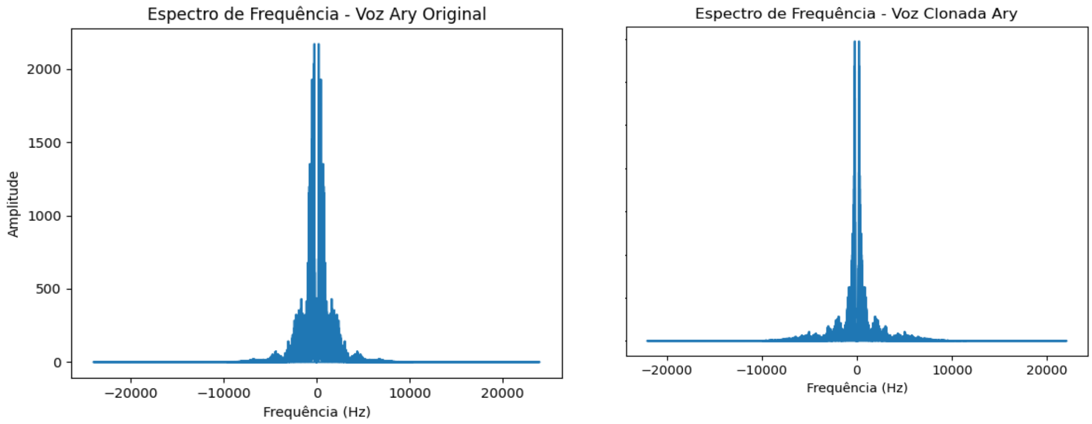
Figure 16 – Inserting text for the voice Aryane -Singer



Source: ElevenLabs (2024)

Revista ARACÊ
ISSN: 2358-2472

Figure 17 and Figure 18 – Original and Cloned Voice, respectively.



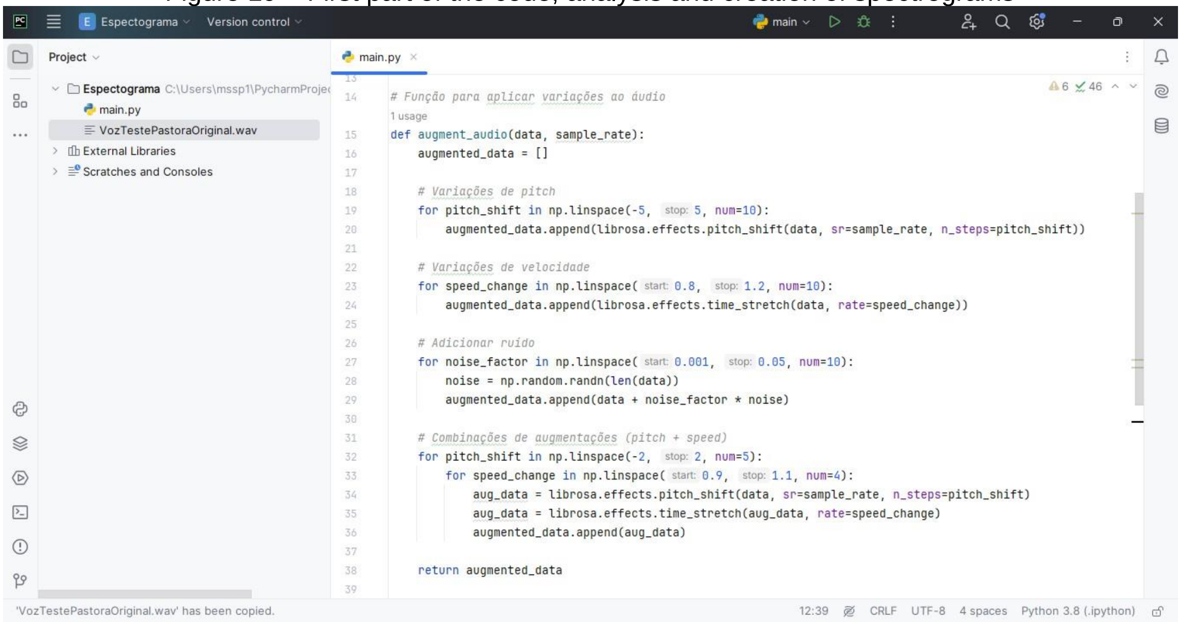
In the analysis of figures 17 and 18, we noticed that both images demonstrate that most of the energy is concentrated in low frequencies, close to 0 Hz, giving the perspective that the audio signals are similar in terms of basic frequency content. Figure 18 (Ary cloned voice) indicates that it has been simplified or has a lower dynamic range compared to figure 17 (Ary Original Voice), another point is that the amplitude peaks present in figure 18 (Ary cloned voice), replicate the main characteristics of the original voice, but are lost in fidelity or dynamic range, as indicated by the smaller amplitude and the limitations represented in the frequencies.

CODE FOR SPECTROGRAM PATTERN ANALYSIS

After cloning the voices, converting them to WAV format and generating the spectrograms several times, the need arose to create a script to perform the sound pattern analysis and generate new spectrograms. The figures below (19-21) show the full script:

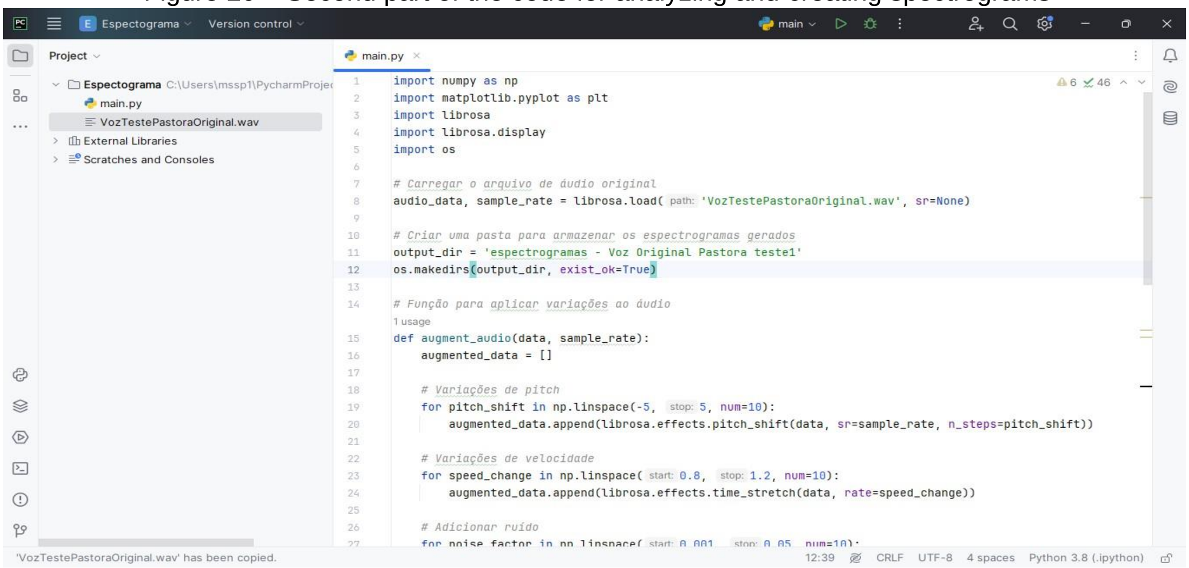
Revista ARACÊ
ISSN: 2358-2472

Figure 19 – First part of the code, analysis and creation of spectrograms



Source: Author (2024)

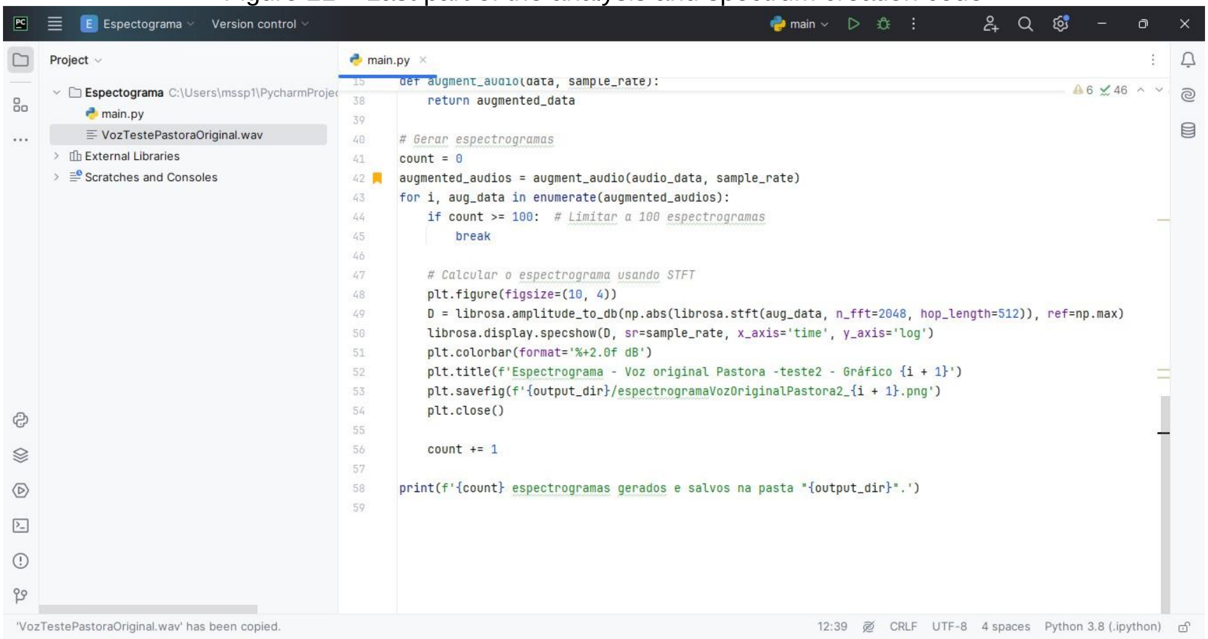
Figure 20 – Second part of the code for analyzing and creating spectrograms



Source: Author (2024)

Revista ARACÊ
ISSN: 2358-2472

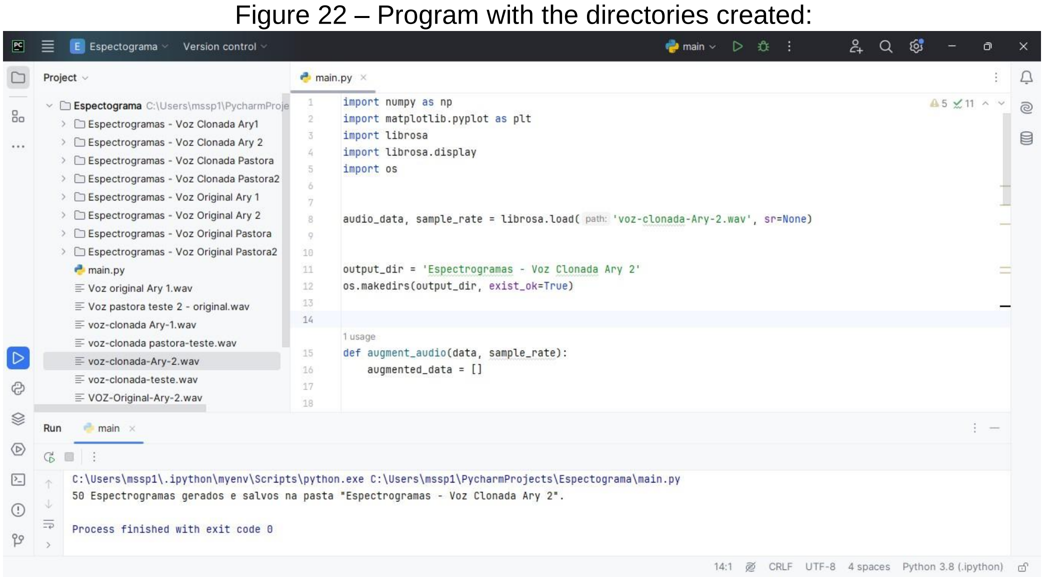
Figure 21 – Last part of the analysis and spectrum creation code



Source: Author (2024)

After placing the audios of the original voices and the cloned voices of the "Pastora and Cantora Aryane", some directories were created for the respective audios, with about 50 spectrogram images in each directory. Figure 22, below, illustrates this process:

Figure 22 – Program with the directories created:



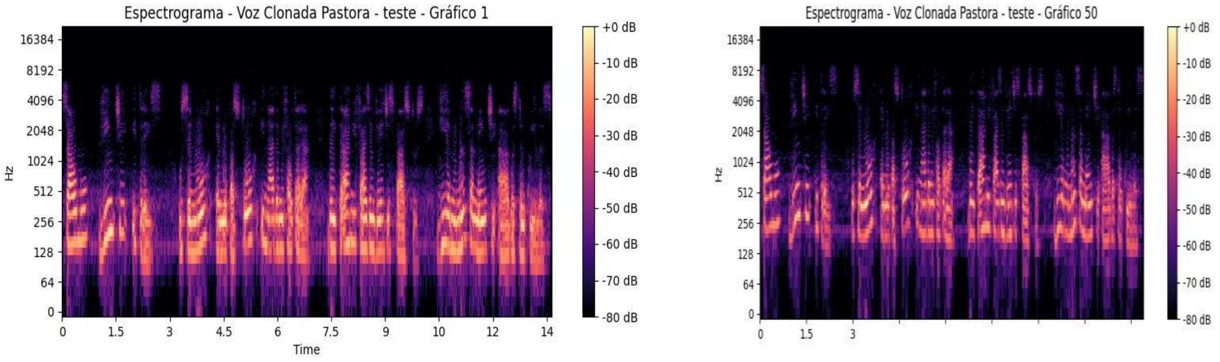
Source: Author (2024)

It is noted that with each audio processed, a directory with several spectrogram images is created. These spectrograms are generated based on the transformations applied

Revista
ARACÊ
ISSN: 2358-2472

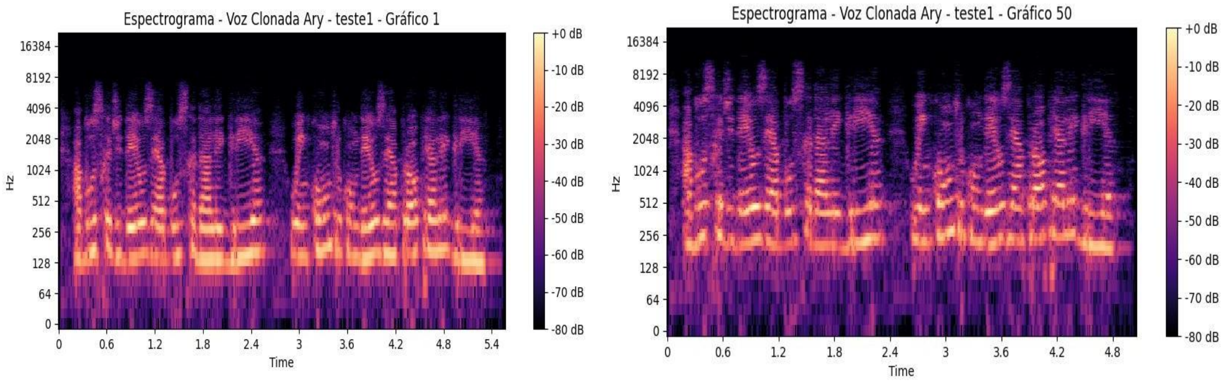
to the audio signal, and are visually represented by frequency variations over time. As shown in figures 23 and 24.

Figure 23 and Figure 24 – Spectrograms of the original voice of the "Pastora"



These graphs visualize the strength of an audio signal's frequencies over time. The differences found are present in total time, where figure 23, spectrogram of the original voice of the shepherdess, covers around 14 seconds, while figure 24, spectrogram of the original voice of the shepherdess, covers around 22 seconds. In the visual characteristics it presents the colored stripes, similar, indicating that it is the same audio. However, it has small differences in the distributions of the highest frequencies around 4096 Hz and above, and in the density of the bands in some regions.

Figure 27 and Figure 28 – Spectrograms of the original voice "Aryane Cantora"



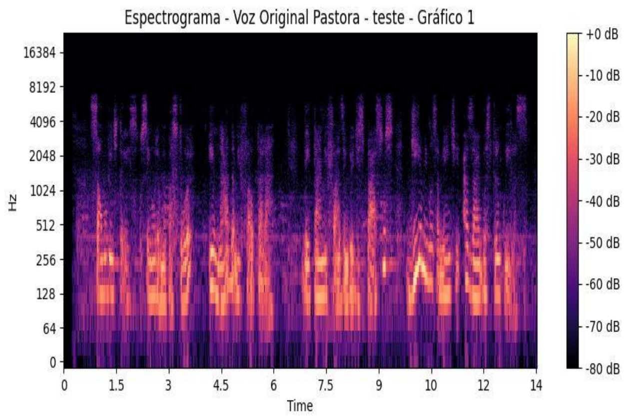
In figures 27 and 28, these graphs present a color scale indicating the signal strength in dB (decibels), where light tones represent greater intensity and dark tones, lower intensity. The dominant frequencies in the two graphs are similar, they are at 128 Hz and 1024 Hz, with harmonic components at higher intervals, which is common and fundamental in human speech. In amplitude and energy, the two figures represent areas with greater intensity concentrated between 128 Hz and 1024 Hz. Something interesting found in the

Revista ARACÊ
ISSN: 2358-2472

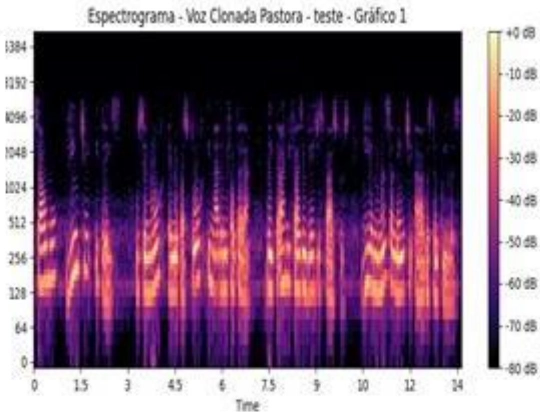
repeating components is that there are repeating patterns that apparently occur cyclically, where it is found in phonemes or syllables in the voice signal.

COMPARISON OF SPECTROGRAMS

Figure 30 and Figure 31 – Spectrograms original and cloned voice of the "Pastora"



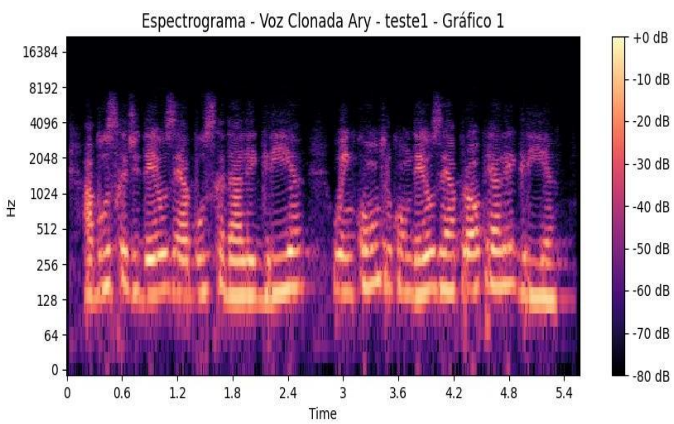
Source: Author (2024)



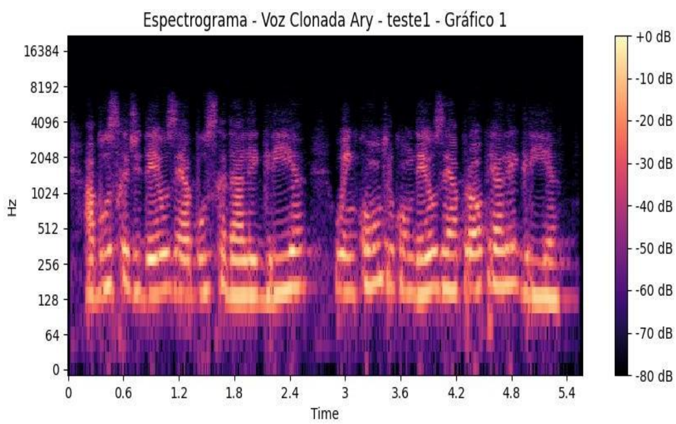
Source: Author (2024)

When analyzing figures 30 and 31, it is noted that both images have very similar frequency patterns, where they have small variations because it is an original voice and the other cloned. The main frequency bands range from 64 Hz to 4096 Hz, containing a similar intensity and visual pattern. Other differences present are that in figure 30, which is the original voice, the "energy" present is slightly more uniform, while figure 31, cloned voice, has a slight variation in harmonics, the differences in intensities and details of frequency patterns may suggest that the cloned voice presents slight distortions.

Figure 32 and Figure 33 – Spectrograms original and cloned voice of "Aryane Cantora"



Source: Author (2024)



Source: Author (2024)

Revista ARACÊ
ISSN: 2358-2472

The two figures show similar patterns, demonstrating the effectiveness of voice cloning, this effectiveness occurs mainly because it manages to maintain the harmonic characteristics of the original recording. The most apparent differences are found in the small variations in intensity in some frequency ranges and short duration of the cloned voice.

SPECTROGRAM-BASED VOICE DIFFERENTIATION SCRIPT

Figure 34 – 39 – Full Code Images

Figure 34 - Source: Author (2024) Figure 35 - Source: Author (2024)

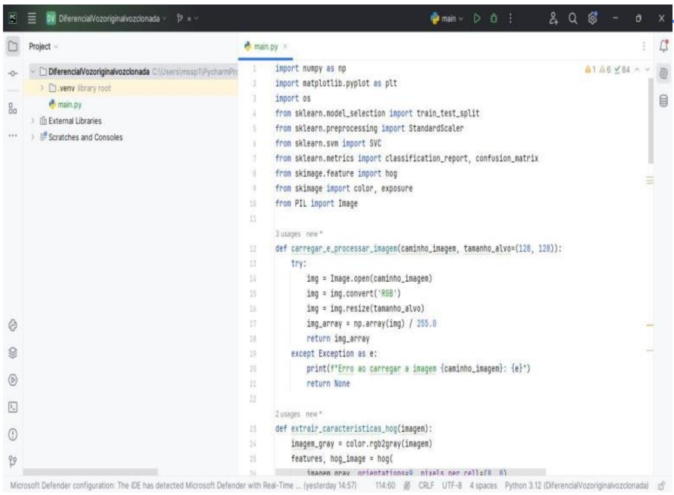
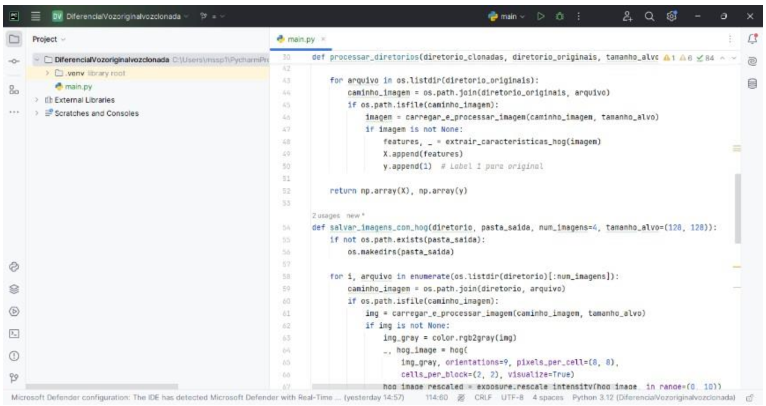


Figure 36 - Source: Author (2024) Figure 37 - Source: Author (2024)



Revista ARACÊ
ISSN: 2358-2472

Figure 38 - Source: Author (2024)

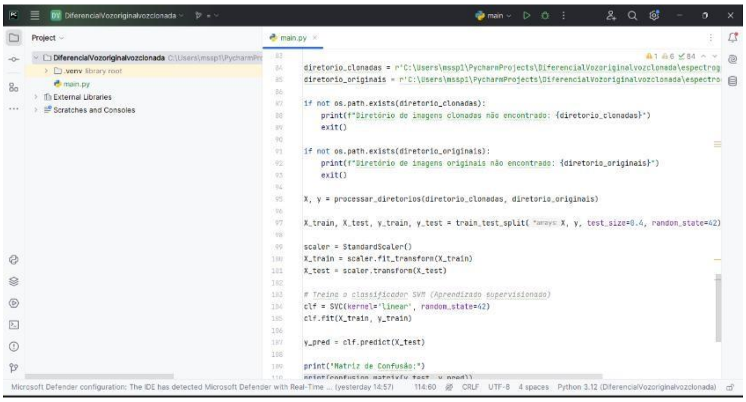


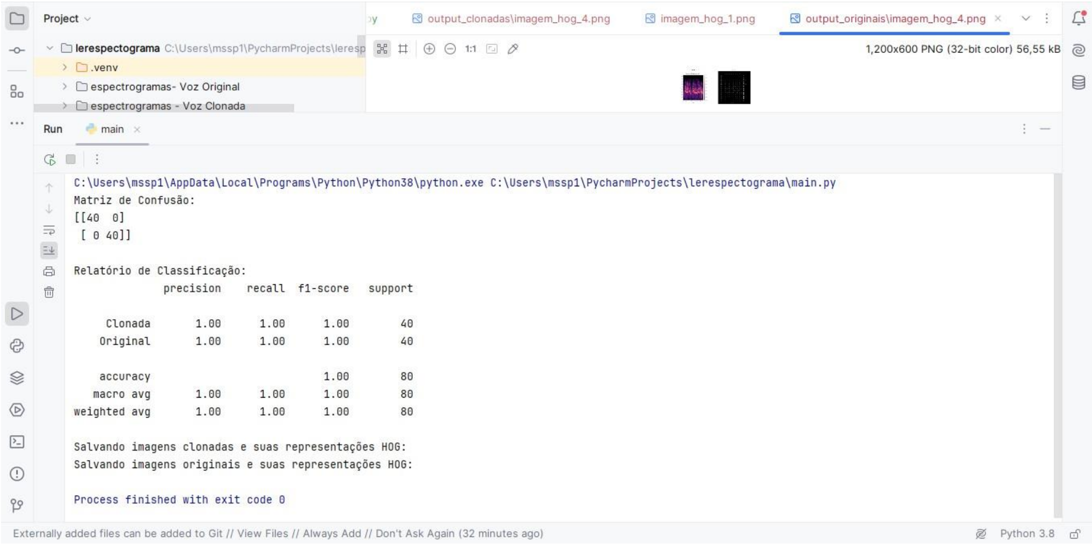
Figure 39 - Source: Author (2024)

RESULTS

With the use of the script found in topic 5.2 and shown in figures 34-39, the following results are obtained, as suggested in figures 40 - 44.

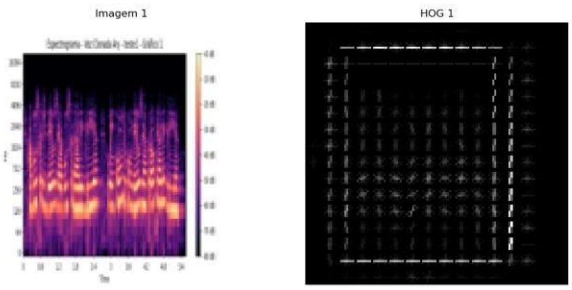
Revista ARACÊ
ISSN: 2358-2472

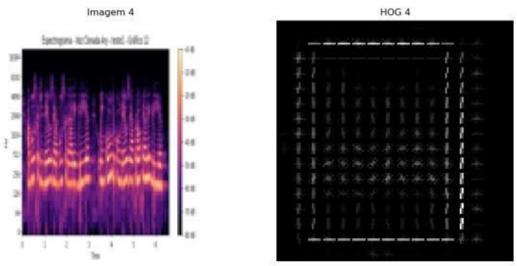
Figure 40 – Result of the script.



Source: Author (2024)

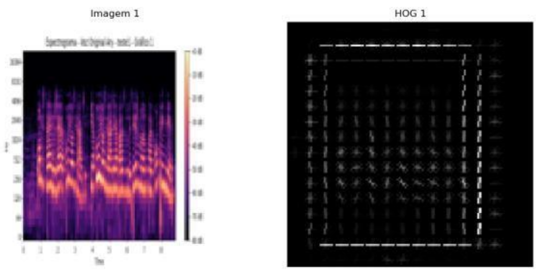
Figure 41 and Figure 42 – HOG Image – Cloned Voice



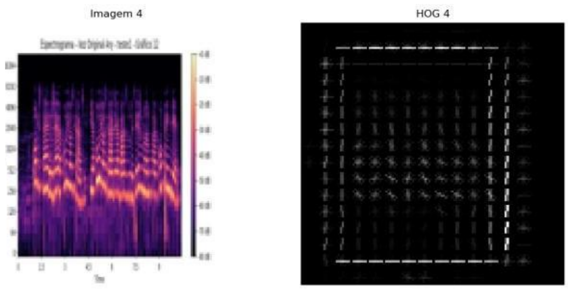


Source: Author (2024) Source: Author (2024)

Figure 43 and Figure 44 – HOG Image – Original Voice



Source: Author



Source: Author

Figure 40 returns a confusion matrix, which is basically a table where it is possible to visualize the performance of the model in relation to the classes, cloned and original. In the script it returns [40 0], which means that the model correctly classified 40 samples of the "Cloned" class correctly, and [0 40], which means that the model classified 40 samples of the "Original" class without making any mistakes.

Revista ARACÊ
ISSN: 2358-2472

Below the confusion matrix, a classification report was also returned, where the metrics have a value of 1.00 for both classes, that is, the performance was considered perfect. The evaluated aspects were *Precision* : 1.00 for "Cloned" and 1.00 for "Original", *Recall* : 1.00 for both classes, which indicates that all samples of each class were correctly identified. *F1-Score:* 1.00 for both classes, indicating a perfect harmony between precision and *recall* and finally in total accuracy obtained a result of 100% in a total of 80 samples.

Figures 41 to 44 represent images in HOG (*Histogram of Oriented Gradients),* which is a feature extraction technique that captures texture and contour information from the images used in a computer vision. This may indicate that the system is recording the representations for possible later analysis, or reuse in new tests, i.e., it is learning the patterns of the images.

CONCLUSION

This research analyzed the ability to differentiate simulated voices from authentic voices through the analysis of spectrograms and artificial intelligence tools, such as ElevenLabs. By creating scripts for the evaluation of spectral patterns and applying Deep Learning and Machine Learning methods, we were able to distinguish significant differences between the frequencies of real voices and synthesized voices.

The findings indicate that the application of spectrograms and the identification of attributes such as the HOG (Histogram of Oriented Gradients) are efficient in distinguishing voices. The confounding matrix and categorization metrics demonstrate the accuracy and consistency of the model, achieving 100% accuracy when recognizing the "Cloned" and "Original" categories. These results suggest that frequency evaluation techniques and vocal patterns can provide an extra level of protection against fraud and manipulation by audio DeepFakes.

This study shows that computer analysis of spectrograms can be useful resources in the identification of false voices. In a context where DeepFakes technology is increasingly employed in problematic situations, such as fraud, it is crucial to create efficient techniques to identify such manipulations, in order to safeguard the integrity of information and the security of the population.

Revista ARACÊ
ISSN: 2358-2472

REFERENCES

1. Almutairi, Z., & Elgibreen, H. (2022). A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions. Algorithms, 15(5), 155. https://doi.org/10.3390/a15050155. Acesso em: 17/04/2024.

2. Alves, P. (2020). Inteligência e Redes Neurais. IPEA. Acesso em: 21/04/2024. Disponível em: Inteligência Artificial e Redes Neurais - Centro de Pesquisa em Ciência, Tecnologia e Sociedade.

3. Fanaya, P. F. (2021). Deepfake e a realidade sintetizada. TECCOGS – Revista digital de tecnologias cognitivas, 23, 104–118. Acessado em: 21/04/2024. Disponível em: Vista do Deepfake e a realidade sintetizada.

4. IBM Brasil. (n.d.). O que é Deep Learning? Acesso em: 20/04/2024. Disponível em: IBM.

5. CNN. (2024). Saiba o que é deepfake, técnica de inteligência artificial que foi apropriada para produzir desinformação. Acesso em: 02/06/2024. Disponível em: CNN Brasil.

6. ComunitIA. (n.d.). ElevenLabs. Acesso em: 02/06/2024. Disponível em: ComunitIA.

7. Convertio. (n.d.). Conversor de áudio. Acesso em: 03/06/2024. Disponível em: Convertio.

8. IIElevenLabs. (n.d.). ElevenLabs. Acessado em: 02/06/2024. Disponível em: ElevenLabs.

9. Kaspersky Daily. (2023). Não acredite em tudo o que ouve: deepfakes de voz. Acesso em: 23/04/2024. Disponível em: Kaspersky.

10. Masood, M., Nawaz, M., Malik, K. M., Ali, J., & Irtaza, A. (2021). Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward. Cornell University. Acessado em: 23/04/2024. Disponível em: arxiv.org.

11. Nishida, M. S., Weber, S. A. T., Oliveira, F. A. K. de, & Troll, J. (n.d.). Voz Humana. Nadi. Acesso em: 23/04/2024. Disponível em: UNESP.

12. Ponti, M., & Costa, G. (2017). Como funciona a Deep Learning (pp. 63–93). ISBN 978-85-7669-400-7. Acesso em: 21/04/2024. Disponível em: USP.

13. Sichman, J. (2021). Inteligência Artificial e sociedade: Avanços e Riscos. SciELO Brasil, 35, 37–49. Acesso em: 20/04/2024. Disponível em: SciELO.

14. Spencer, M. K. (2019). DeepFake, a mais recente ameaça distópica. Outras Palavras. Acessado em: 21/04/2024. Disponível em: Outras Palavras.

15. UOL Notícias. (2024). Deepfake: uso de inteligência artificial em eleições na Argentina e nos Estados Unidos. Acesso em: 21/04/2024. Disponível em: UOL Notícias.

16. Your Physicist. (n.d.). 4 tipos mais comuns de técnicas de análise espectral. Acesso em: 23/04/2024. Disponível em: Your Physicist.