Revista ARACÊ
ISSN: 2358-2472

DECIFRANDO DEEPFAKES DE ÁUDIOS: UMA INVESTIGAÇÃO PELA ÁNALISE DE FREQUENCIA

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

Data de submissão: 29/10/2024 **Data de publicação:** 29/11/2024

Maria Stella Simões Piccolo
Graduanda do Curso de Sistema de Informação da Universidade de Araraquara - UNIARA
Araraquara-SP
E- mail: mssp1717@outlook.com

João Henrique Gião Borges
Orientador
Docente do Curso de Sistema de Informação da Universidade de Araraquara-UNIARA. Araraquara – SP
E-mail: jhgborges@uniara.edu.br

Fabiana Florian
Coorientadora
Docente do Curso de Sistema de Informação da Universidade de Araraquara – UNIARA. Araraquara – SP
E-mail: fflorian@uniara.edu.br

RESUMO
Esta investigação examina a tecnologia emergente dos DeepFakes de áudio, que se apresenta como um desafio contemporâneo. Para tratar deste assunto, empregou-se o ElevenLabs para a clonagem de voz, juntamente com scripts fundamentados em Deep Learning e Machine Learning para a criação de espectrogramas e avaliação de padrões nas frequências das vozes. Os achados indicaram que a avaliação dos gráficos de cada voz é eficiente para diferenciar a voz artificial da verdadeira.

Palavras-chave: Inteligência Artificial. DeepFakes. Deep Learning. DeepFakes de Audio. ElevenLabs. Espectograma.

Revista ARACÊ
ISSN: 2358-2472

1 INTRODUÇÃO

No cenário contemporâneo, caracterizado pelos avanços tecnológicos e pela proliferação da Inteligência Artificial (IA), surge uma nova tecnologia dentro do complexo de *DeepFakes*: as *DeepFakes* de Áudio. Embora seu desenvolvimento tenha sido inicialmente paratrazer uma melhora na rotina cotidiana, com o uso dos *audiobooks,* essa tecnologia tem sido usada e explorada para uma manipulação da realidade, representando uma ameaça para a segurança pública.

As *DeepFakes*, são produtos da IA, que fundem, combinam, substituem ou até mesmosobrepõe ou substitui áudios e imagens para criar arquivos falsos. Este artigo, tem como foco oestudo da *DeepFakes* de Áudio, que possui a capacidade de gerar vozes falsas e a clonar vozesexistentes e sua análise envolve diversos parâmetros, incluindo frequência, intensidade e outrosatributos sonoros.

Este artigo tem o objetivo de estudar o funcionamento das *DeepFakes* de Áudio comparando as frequências das vozes que foram clonadas e da voz real. A utilização das ferramentas ElevenLabs para geração das vozes falsificadas e clonadas e dos Espectrogramas para uma visualização gráfica da frequência das vozes, contribuíram na identificação precisa das *DeepFakes* de Áudio, distinguindo uma voz da outra.

De acordo com o site UOL notícias (2024), o avanço na utilização das *DeepFakes*, especialmente na manipulação de áudios, gera preocupações crescentes sobre suas consequências. Muitas pessoas ainda não possuem compreensão suficiente sobre essa temáticao que pode levar a situações problemáticas, como golpes e em casos recentes manipulações naseleições. Nesse site relata-se alguns usos de *DeepFakes*, em um caso especifico um dos alvos foi Joe Biden, atual presidente dos Estados Unidos, que passou por esse problema de *DeepFakes*de áudio em uma eleição nos Estados Unidos, onde um áudio simulando sua voz foi circulado,ilustrando a ameaça à integridade do processo democrático.

Tendo em vista esses desafios, percebe-se a necessidade da compreensão das *DeepFakes* de Áudio e de desenvolver métodos de análise ou aprimorar as já existentes permitindo a possibilidade de discernir vozes autenticas e falsificadas, contribuindo para umatentativa de diminuir os riscos associados aos usos indevidos.

Segundo Almutairi, Z., & Elgibreen, H. (p.2, 2022). “A detecção de DeepFakes de áudiotornou-se, portanto, uma área ativa de pesquisa com o desenvolvimento de técnicas avançadase métodos de Deep Learning (DL). No entanto, com tais avanços, os métodos atuais de DL estão enfrentando dificuldades, e uma investigação adicional é necessária para entender em qualárea da detecção de DeepFakes de áudio precisa de mais desenvolvimento. Além disso, uma análise comparativa dos

Revista ARACÊ
ISSN: 2358-2472

métodos atuais também é importante, e, até onde sabemos, falta na literatura uma revisão dos métodos de detecção de áudio imitados e gerados sinteticamente (Tradução nossa).”

A hipótese deste artigo é que a utilização das ferramentas ElevenLabs e Espectrogramas permitirão distinguir com eficiência algumas questões que nortearão a pesquisa são que em cada voz há diferentes níveis de frequências, por meio da frequência pode-se realizar uma análise comparativa entre cada voz. Para testar essa hipótese, será utilizado ferramenta como Espectrogramas que serve como um gráfico que demonstra as frequências das vozes por meio de desenhos de ondas sonoras.

Para iniciar essa pesquisa é necessário primeiro entender o que é *DeepFakes?* Como funciona a *DeepFakes* de Áudio? Como se analisa a frequência de voz? O Espectrograma irá ser diferente entre uma voz real e a voz clonada? Para responder tais questionamentos foi realizado uma leitura de diferentes artigos que abordam sobre o que é essa tecnologia *DeepFakes*, será estudado como funciona essa *DeepFakes* de Áudio, haverá a utilização da ferramenta ElevenLabs para realizar a clonagem das vozes, para a realização da análise, será desenvolvido um script que receberá esses arquivos de áudios e os representará em forma de Espectrogramas, ou seja, em uma forma gráfica, onde serão formadas ondas que representam as frequências, para realizar a comparação.

2 PESQUISA BIBLIOGRÁFICA

Esta seção aborda os conceitos de Inteligência Artificial, *Deep Learning*, *DeepFakes*, *DeepFakes* de áudio, (2.3) Frequência de voz, (2.4) Potencial da ElevenLabs na clonagem de voz, (2.5) Relação do espectrograma e da frequência.

2.1 INTELIGÊNCIA ARTIFICIAL (IA)

Para ter uma compreensão melhor do tema deste artigo, é necessário primeiro entender a sua estrutura, como já relatado, a *DeepFakes* uma tecnologia formada através da inteligência artificial usando uma metodologia chamada de *Deep Learning*, tendo isso em vista,como podemos descrever a IA?

Segundo Jaime Simão Schiman (2021), “O domínio de IA se caracteriza por uma coleção de modelos, técnicas e tecnologias (busca, raciocínio e representação de conhecimento,mecanismos de decisão, percepção, planejamento, processamento de linguagem natural, tratamento de incertezas, aprendizados de máquinas) que, isoladamente ou agrupadas, resolvemproblemas de tal natureza.”.

Já a Autora Priscila Mello Alves (2020), diz que “As definições de IA encontradasna literatura científica, enquanto disciplina do conhecimento humano, são categorizadas, sendoou empiricamente com formulação de hipóteses e confirmações experimentais ou teoricamente envolvendo cálculos

Revista ARACÊ
ISSN: 2358-2472

matemáticos. Na ótica da categoria empírica há perspectiva de os sistemas pensarem como seres humanos, possibilitando o aprendizado com experiências, enquanto no enfoque teórico esperam-se ações lógicas que incluem capacidade de dedução e inferência sobrenovas relações".

Tendo em vista essas duas definições, percebe-se que a IA, é uma tecnologia, cujadefinições variam muito, mas que em sua base, podemos denominar como uma tecnologia queabrange diversos modelos, que em seu entorno pode comportar-se como humana, no sentido que pode aderir técnicas de aprendizagem como *machine learning* ou *deep learning*, cogniçãoe tomadas de decisão como rede neurais, dentre outras, tornando -se uma maneira de otimizar processos, automatizar funções e diminuir a complexidade.

2.1.1 deep learning – (aprendizado profundo)

Segundo o site da IBM, "*Deep Learning* é um subconjunto de aprendizado de máquina, que é essencialmente uma rede neural com três ou mais camadas. Essas redes neuraistentam simular o comportamento do cérebro humano, embora longe de corresponder a sua capacidade, permitindo que ele "aprenda" com grandes quantidades de dados. Embora uma redeneural com uma única camada ainda possa fazer previsões aproximadas, camadas ocultas adicionais podem ajudar a otimizar e refinar a precisão."

> Aprendizado Profundo oferecem atualmente um importante conjunto de métodos para analisar sinais como áudio e fala, conteúdos visuais, incluindo imagens e vídeos, e ainda conteúdo textual." (Ponti Moacir eCosta Gabriel, 2017, 1, p. (63-93), "Como funciona Deep Learning").

Com as descrições desses autores, nota-se que a IA, não é uma tecnologia consolidada por si só, ela abrange diversas outras para determinadas funções, no caso deste artigo, a *Deep Learning* será algo presente, visto que teremos que realizar uma análise referente as vozes gravadas e criadas.

2.2 DEEPFAKES

> "O termo DeepFakes refere-se às mídias sintéticas nas quais imagens ou sonscapturados de determinadas pessoas são substituídos pelos de outras por meiode técnicas avançadas de aprendizagem de máquina e Inteligência Artificial (IA), com a finalidade de manipular conteúdo visuais e/ou sonoros, com enorme potencial de falseamento da realidade. Celebridades e políticos têm sido os alvos preferenciais dessas manipulações, que têm se popularizado exponencialmente até mesmo por meio de aplicativos gratuitos de celular." (Fonseca, p.106, FANAYA, fevereiro/2021).

Revista ARACÊ
ISSN: 2358-2472

"DeepFakes são, essencialmente, identidades falsas criadas com o deep learning (aprendizado profundo) por meio de um uso maciço de dados" (Spencer, Michael K., tradução:Gabriela Leite, 2019, outras palavras).

Segundo o site CNN (2024), as *DeepFakes* ocorrem quando a inteligência artificial(IA), funde, combina, substitui ou sobrepõe áudios e imagens para criar arquivos falsos em quepessoas podem ser colocadas em qualquer situação, dizendo frases nunca ditas ou assumindo atitudes jamais tomadas. O conteúdo pode ser de caráter humorístico, político ou mesmo pornográfico. São inúmeras as possibilidades: troca de rostos, clonagem de voz, sincronizaçãolabial a uma faixa de áudio diferente da original, entre outras. A técnica comumente distorce apercepção a respeito de um indivíduo em uma determinada situação.

Ao seguir algumas definições entendemos que as *DeepFakes* vêm da junção de *"Fake News"* com a tecnologia *Deep Learning* por meio da inteligência artificial, onde através de seu uso, podemos criar conteúdo sintéticos, podendo ser imagem, vídeo e áudio, como explicito no tema, onde os dados armazenados são utilizados por um algoritmo de *Machine Learning* e *Deep Learning.*

2.2.1 deepfakes de áudio

Dentre todo esse avanço de tecnologia, aumento crescente do uso da inteligência artificial e consequentemente novos surgimentos de *DeepFakes,* surge a *DeepFakes* de Áudio,que é o foco central deste artigo.

Segundo Almutairi, Z., & Elgibreen, H. (p-3,2022). Ferramentas sintetizadas por IA foram recentemente desenvolvidas com capacidades de gerar vozes convincentes. No entanto, embora essas ferramentas tenham sido introduzidas para ajudar as pessoas, elas também foramusadas para espalhar desinformação pelo mundo usando áudio, e seu mau uso mal-intencionado levou ao medo do *"DeepFakes* de Áudio*"* (tradução nossa).

> "DeepFakes de Áudio concentram-se na geração da voz do orador alvo, usando técnicas de aprendizado profundo para retratar o orador dizendo algo que não foi dito. As vozes falsas podem ser geradas usando síntese de texto para fala (TTS) ou conversão de voz (VC)" (tradução nossa). (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.)

A *DeepFakes* de Áudio funciona através de uma tecnologia onde se é possível clonar a sua voz, através de uma conversão de voz que se baseia em alguns codificadores, do tipo rede neural que vai comprimir os dados de entrada em uma representação compacta e interna, e acaba aprendendo a

Revista ARACÊ
ISSN: 2358-2472

descompactar esses dados dessa representação para restaurar os dados originais, dessa maneira ela aprende a apresentar os dados em um formato compacto enquanto destaca informações importantes.

Focando nos estudos sobre *DeepFakes* de Áudio, percebemos o uso de técnicas avançadas de inteligência artificial, especificamente *Deep Learning* e *Machine Learning*, paraclonar a fala de uma pessoa de maneira natural e convincente. A plataforma utilizada para a clonagem de voz e as etapas para a clonagem serão descritas a seguir:

2.3 FREQUÊNCIA DE VOZ

> Para produzir a voz, primeiro, o ar que está sendo expirado deve passar pelascordas vocais e vibrá-las (tente falar durante a inspiração e veja como é difícil). As cordas vocais são estruturas musculomembranosas que formam um V invertido(...). Se as pregas vocais estiverem relaxadas, o ar passará livremente sem produzir som. Quando desejamos falar, o cérebro envia comandos nervosos para os músculos que controlam a tensão das cordas vocais e o ar passará vibrando as pregas. As vibrações são muito rápidas, cujafrequência é de mais 100 vezes por segundo. Enquanto ocorre a vibração, várias estruturas (cavidades da boca, garganta e a passagem nasal) entram em ressonância produzindo várias ondas harmônicas e amplificando o som. (Nishida, M. S., Weber, S. A. T., Oliveira, F. A. K. de, & Troll, J. Voz Humana).

Entendendo o funcionamento da voz, temos a percepção que o som que emitimos pode ser distinguido conforme as movimentações das pregas vocais, o que pode gerar algumascaracterísticas diferentes. Em uma análise de voz, podemos distinguir uma voz dá outras, pela altura, intensidade e amplitude. Na altura verificamos se o som é agudo ou grave, o que é determinado pela frequência da onda, o timbre é distinguido através das ondas sonoras e a intensidade é medida pelo volume, onde um som mais intenso, forte ou fraco vem de um fatordenominado amplitude da onda, onde quanto maior a amplitude de uma onda, maior é a pressãoque ela irá exercer no ar, o que faz com que nossos tímpanos vibrem de maneira mais intensa.

Vale ressaltar que nós não ouvimos nossa voz da mesma maneira que outra pessoa,por conta que quem está nos ouvindo não nos escuta por intermédio da via aérea, logo, a nossaprópria voz não passa por alteração de pressão, o que se conclui que a voz gravada é considerada a representação verdadeira da nossa voz.

2.4 POTENCIAL DA ELEVENLABS NA CLONAGEM DE VOZ

A utilização de ferramentas como ElevenLabs contribuiu muito para este artigo, devido a sua capacidade de proporcionar a possibilidade de clonagem de voz, tendo em vista seu crescimento, muitas pessoas passaram a utilizar para brincadeiras de trolagem, criação de vozes sintéticas e até

Revista ARACÊ
ISSN: 2358-2472

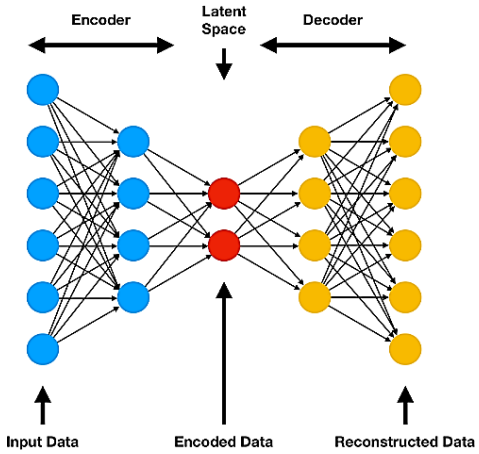
mesmo para tentativas de aplicação de golpes. Apesar de ElevenLabs seruma plataforma paga, ela fornece recursos gratuitos como criação de vozes sintéticas.

De acordo com as informações do site Kaspersky Daily (2023). “A conversão de voz é baseada em codificadores automáticos, um tipo de rede neural que comprime os dados deentrada (parte do codificador) em uma representação interna compacta e, então, aprende a descompactá-los dessa representação (parte do decodificador) para restaurar os dados originais.

Desta forma, o modelo aprende a apresentar os dados em um formato compactadoenquanto destaca as informações mais importantes.”

Como sugere a imagem disponibilizada neste mesmo site. Encontrada na figura 1.

Figura 1: Esquema Codificador Automático:

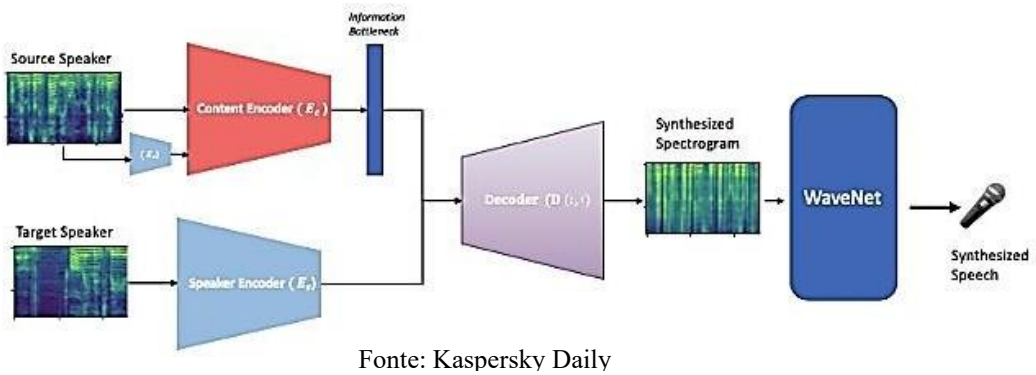


Fonte: Kaspersky Daily (2023)

O processo de criação dessas *DeepFakes*, geralmente envolve ao menos duas gravações de áudios que são alimentadas no modelo. A segunda gravação é convertida para a parecer com a primeira e após isso o codificador de conteúdo determina o que foi dito a partir desta primeira gravação. O codificador extrai as características da voz, como timbres, entonação, sotaques, frequências, amplitudes da segunda gravação, essas representações são combinadas para a produção do resultado por meio do decodificador.

Revista ARACÊ
ISSN: 2358-2472

Figura 2- Processo de geração da voz das *DeepFakes*:



Fonte: Kaspersky Daily

Plataformas como ElevenLabs, entre outras que oferecem essa possibilidade de clonar vozes utiliza-se desses processos, juntamente com a utilização de algoritmos de inteligência artificial como *Machine Learning* e *Deep Learning*.

2.5 RELAÇÃO DO ESPECTROGRAMA E DA FREQUÊNCIA

Tendo em vista que a *DeepFakes* de Áudio, possui essa capacidade de clonagem devoz, criação de áudio sintético e entendendo que a nossa voz produz uma certa frequência, pararepresentá-la surge a necessidade de analisá-la e o espectrograma é uma técnica que possibilitaessa visualização gráfica.

Segundo o site, Your Physicist, 2024, o Espectrograma é uma técnica que combinaa análise de Fourier com o tempo para visualizar as características espectrais de um sinal ao longo do tempo. Ele representa graficamente a intensidade de cada componente de frequência em função do tempo, fornecendo uma representação visual clara das mudanças no conteúdo espectral ao longo do sinal.

Essa técnica é amplamente utilizada em áreas como análise de áudio, processamento de fala, monitoramento de vibrações e análise de sinais não estacionários. O Espectrograma permite identificar mudanças transientes ou eventos de curta duração, além de fornecer informações sobre as características de frequência dominantes em diferentes momentos do sinal.

Este artigo visa utilizar o Espectrograma como uma das ferramentas para analisar as frequências emitidas pelos áudios, visando ter uma perspectiva das diferenças, para uma melhor detecção de áudios fakes e clonagem.

3 DESENVOLVIMENTO

Nesta seção apresenta: Utilização da plataforma ElevenLabs para a clonagem da voz, conversão do arquivo MP3 para WAV e foi desenvolvido um script para gerar o espectrograma de frequência, um outro script para aprender os padrões dos espectrogramas e por último um script para análise das diferenças entre uma voz real e uma voz clonada.

Revista ARACÊ
ISSN: 2358-2472

3.1 ELEVENLABS

Segundo o site ComunitIA (2023), ElevenLabs é uma inovadora empresa de tecnologia que tem como principal produto o áudio gerado por inteligência artificial. Ela traz ferramentas avançadas para criadores e editoras que buscam aprimorar suas narrativas, oferecendo vozes cativantes, emotivas e realistas.

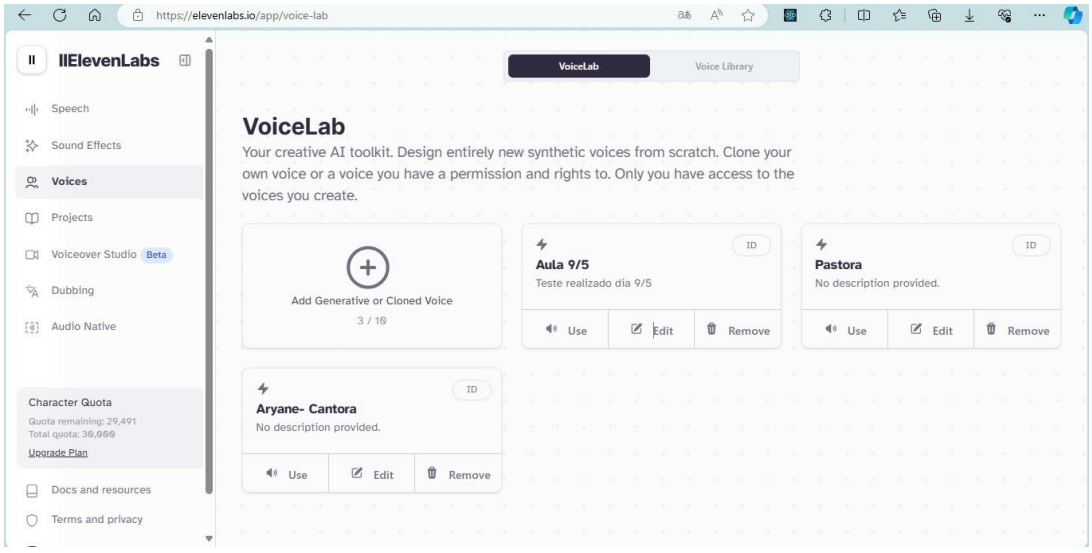
Neste trabalho foi utilizado o site do ElevenLabs, a ferramenta VoiceLabs, para a clonagem das vozes, O processo envolve a gravação das vozes desejadas, a escrita de uma frasepara a voz clonada, análise e aprimoramento para tornar o áudio mais natural, e finalmente, o download do áudio gerado.

3.2 PROCESSO DE CRIAÇÃO DE *DEEPFAKES* DE ÁUDIO PELA PLATAFORMA ELEVENLABS

3.2.1 coleta de dados

No site do ElevenLabs, em VoiceLab encontrado na seção "Voices". Para iniciar aclonagem, foi selecionado uma das bibliotecas previamente criadas e foi adicionado os áudios.Demonstrado na figura 3.

Figura 3 Plataforma VoiceLabs

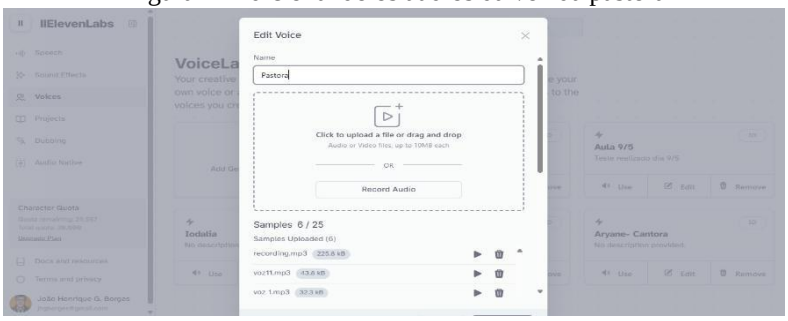


Fonte: ElevenLabs (2024)

Ao selecionar a voz intitulada como “Pastora”, a seguinte interface é exibida. Encontrada na figura 4 e figura 5.

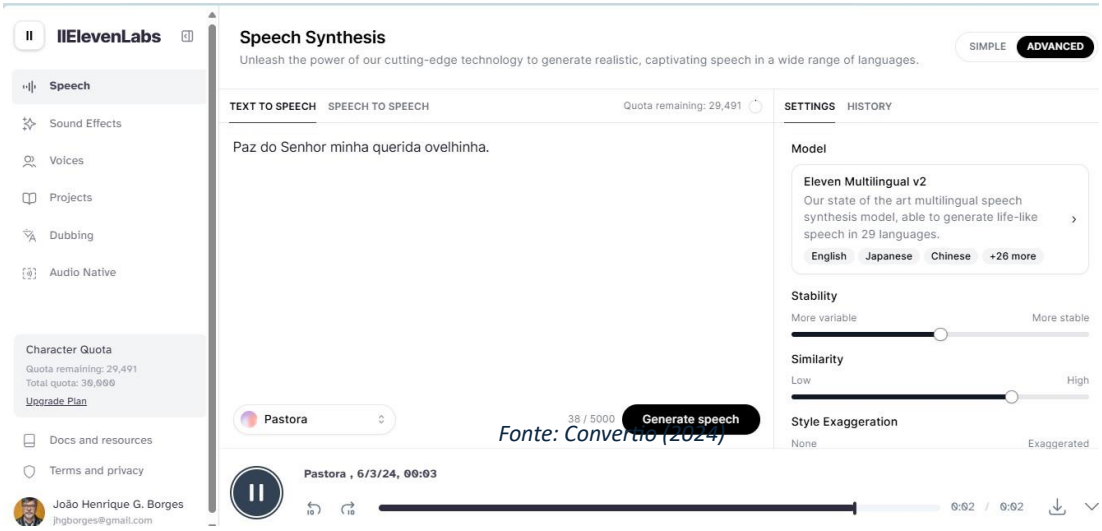
Revista ARACÊ
ISSN: 2358-2472

Figura 4- Adicionando os áudios da voz da pastora



Fonte: ElevenLabs (2024)

Figura 5 – Inserindo um texto para ser gerado com essa voz:



Fonte: ElevenLabs (2024)

Após clicar em "Generate Speech" demonstrado na Figura 5, o áudio é gerado. Esseáudio é gerado através da técnica de *Deep Learning* para processar o texto e então gerar o áudiorealista, como foi inserido arquivos com a voz original, esse volume de dados foi processado etreinado para capturar nuances, pausas e o ritmo natural da fala. É possível ajustar o áudio paraque a voz se torne ainda mais parecida com a natural. Depois de ajustar, clicamos no botão de download.

3.2.2 convertendo mp3 para wav

Depois de gerar a voz clonada, utilizado um algoritmo criado pelo autor que recebeos arquivos de áudio e gera um espectrograma. Mas para isso, será necessário que haja a conversão do arquivo MP3 para a extensão WAV. Para isso foi utilizado um site chamado Convertio, que faz isso de maneira gratuita. Como é demonstrado na figura 6.

Revista ARACÊ
ISSN: 2358-2472

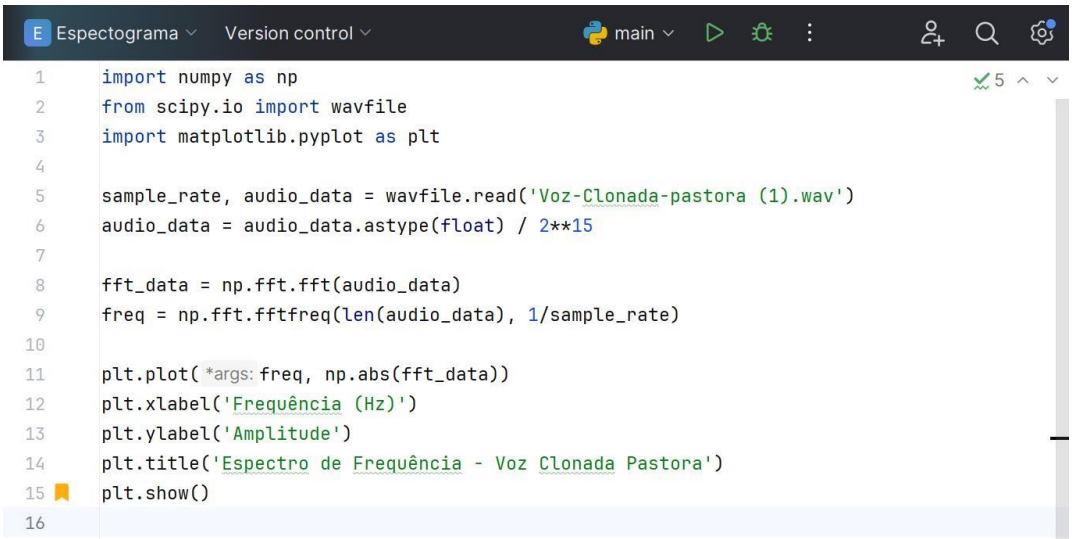
Figura 6 - Conversor de MP3 para WAV:

Fonte: Convertio

4 GERANDO ESPECTROGRAMA ATRAVÉS DO ÁUDIO WAV

A figura 7, tem como função gerar um gráfico denominado espectrograma, por meiodo arquivo MP3 convertido para extensão WAV. Tendo em vista o objetivo do algoritmo, deve-se entender o seu funcionamento, que se baseia em alguns passos como: Importação das bibliotecas; Leitura de Arquivo de áudio; Normalização dos dados do áudio; Cálculo da transformada de Fourier e por fim a plotagem do espectro de frequência. Esse código é encontrado na Figura 7.

Figura 7 – Script Espectrograma



Fonte: Autor (2024)

4.1 IMPORTAÇÃO DAS BIBLIOTECAS

As bibliotecas importadas são numpy, scipy.io.wavfile, matplotlib.pyplot. Demonstradas na Figura 8.

Revista ARACÊ
ISSN: 2358-2472

Figura 8 – Importação das bibliotecas

```
1  import numpy as np
2  from scipy.io import wavfile
3  import matplotlib.pyplot as plt
4
```

Fonte: Autor (2024)

Cada biblioteca tem uma funcionalidade especifica, sendo a *Numpy* utilizada para operações numéricas eficientes, scipy.io.wavfile para ler e escrever arquivos WAVe a *matplotlib.pyplot* criação de gráficos.

4.2 LEITURA DE ARQUIVO

Nessa parte do algoritmo, é realizada uma leitura do arquivo, usando os seguintes conceitos: sample_rate; audio_data, wavfile. read (). Como demonstrada na figura 9.

Figura 9 – Leitura de arquivo

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

Fonte: Autor (2024)

A função wavfile.read() lê um arquivo WAV e retorna duas informações a taxa de amostragem (sample_rate) e os dados do áudio (audio_data).

A variável sample_rate – É a taxa de amostragem do áudio, indicando quantas amostras por segundo foram capturadas e a audio_data é um array contendo os dados do áudio.

4.3 NORMALIZAÇÃO DOS DADOS DO ÁUDIO

Nessa etapa do código, é realizada a conversão da extensão MP3 para WAV e anormalização dos dados do áudio. Demonstrada na Figura 10.

Figura 10 – Normalização dos dados do áudio e conversão das extensões

```
6  audio_data = audio_data.astype(float) / 2**15
7
```

Fonte: Autor (2024)

A variável audio_data. astype(float) converte os dados do áudio para o tipo floate audio_data / 2**15 normaliza esses dados do áudio. Em arquivos WAV com 16 bits por amostra, os valores variam de -32768 a 32767. A divisão por 2152^ {15}215 (32768)normaliza os dados para o intervalo de -1 a 1.

Revista ARACÊ
ISSN: 2358-2472

4.4 CÁLCULO DA TRANSFORMADA DE FOURIER

Nessa parte do código é realizado o cálculo da transformada de Fourier, cuja aplicabilidade é dividir algo em várias ondas senoidais. A Transformada de Fourier é uma ferramenta poderosa para converter um sinal do domínio do tempo para o domínio da frequência, permitindo a identificação de componentes de frequência que não sãovisíveis diretamente no sinal original. Essa parte é encontrada na Figura 11.

Figura 11 – Cálculo da transformada de Fourier

```
8   fft_data = np.fft.fft(audio_data)
9   freq = np.fft.fftfreq(len(audio_data), 1/sample_rate)
10
```

Fonte: autor (2024)

Nesse código np.fft.fft (audio_data) calcula a Transformada de Fourier dos dados do áudio, transformando o sinal do domínio do tempo para o domínio da frequência. Avariável fft_data contém os coeficientes da transformada.

Já a np.fft.fftfreq(len(audio_data), 1/sample_rate), gera uma lista de frequências correspondentes aos coeficientes da Transformada de Fourier onde len(audio_data) é onúmero de amostras, e 1/sample_rate é o intervalo de tempo entre as amostras.

4.5 PLOTAGEM DO ESPECTRO DE FREQUÊNCIA

Nessa parte final do código é onde é desenvolvida a parte da plotagem do espectro da frequência. Mostrado na Figura 12.

Figura 12 – Plotagem do Espectro de frequência

```
11  plt.plot( *args: freq, np.abs(fft_data))
12  plt.xlabel('Frequência (Hz)')
13  plt.ylabel('Amplitude')
14  plt.title('Espectro de Frequência - Voz Clonada Pastora')
15  plt.show()
16
```

Fonte: Autor (2024)

A função plt.plot (freq. np.abs(fft_data)) plota o espectro de frequência, utilizando as frequências (freq) no eixo x e a magnitude dos coeficientes da Transformada de Fourier (np.abs(fft_data)) no eixo y. A variável plt.xlabel ('Frequência(Hz)') define o rótulo do eixo x como "Frequência (Hz)". As variáveis plt.ylabel ('Amplitude') define o rótulo do eixo y como "Amplitude”. Já

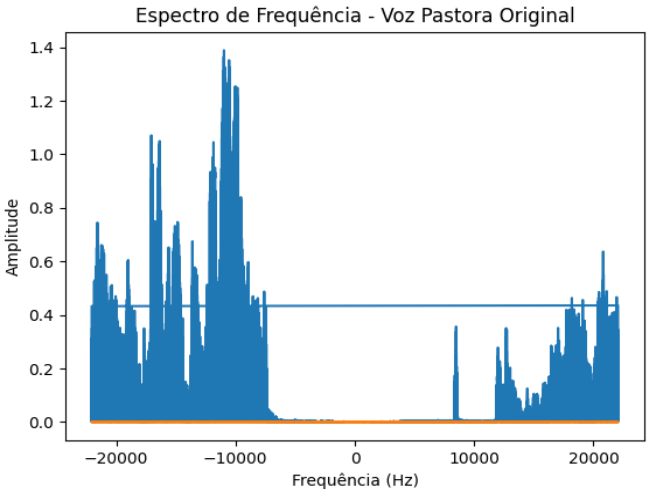
Revista ARACÊ
ISSN: 2358-2472

a função plt.title ('Espectrode Frequência - Voz Clonada Pastora'): Define o título do gráfico e por último a funçãoplt.show() vai exibir o gráfico.

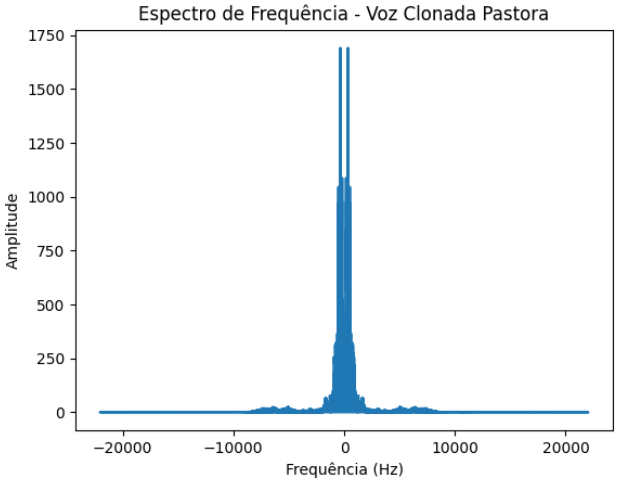
5 GRÁFICO ESPECTROGRAMA

Ao realizar os procedimentos acima, e gerar os gráficos com a voz clonada ecom a voz original, foi obtido os seguintes resultados, Figura 13 e Figura 14.

Figura 13 e Figura 14 – Gráfico Espectro de frequência- Voz original e voz clonada daPastora



Fonte: Autor (2024)



Fonte: Autor (2024)

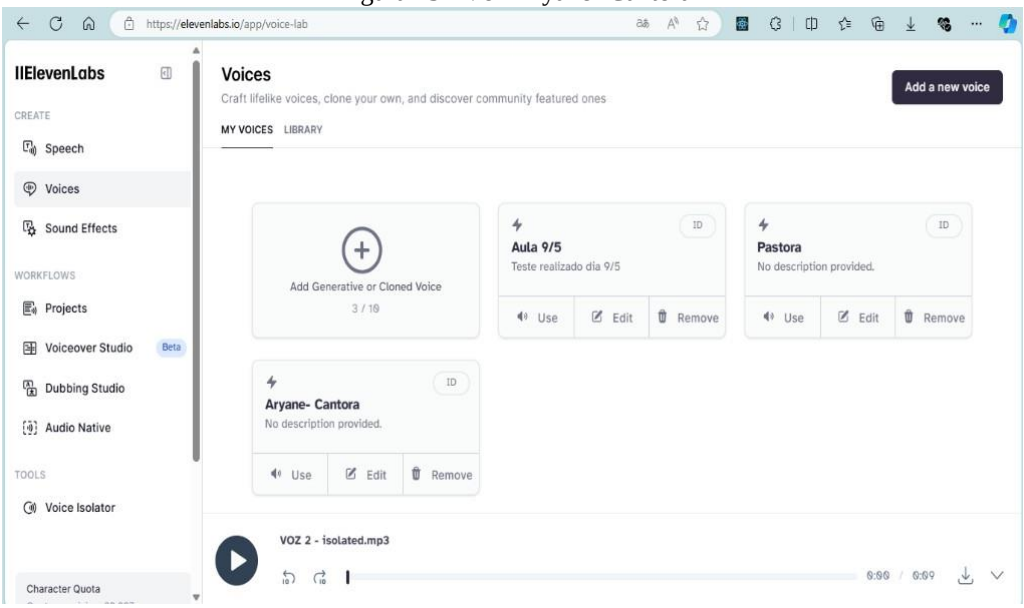
Considerando essas duas imagens (Figuras 13 e 14), podemos fazer a seguinte análise, avaliando as frequências da voz original, como demonstra a figura 13, nota-se uma maior diversidade no que chamamos de componentes harmônicos, contendo picosde amplitude moderados distribuídos por amplas faixas de frequências, sendo elas tantospositivas quanto negativas. A voz clonada apresenta um espectro altamente concentrado, tendo dois picos proeminentes bem próximos de 0 Hz e as amplitudes maiores. Isso significa que nesse caso a clonagem reduziu a riqueza harmônica da voz, deixando de maneira condensada sua energia em uma faixa de frequência restrita.

5.1 TESTANDO TIMBRES DE VOZ

No tópico 3.2.1 coleta de dados, na figura 3 VoiceLabs, utilizamos um áudio denominado “pastora”, agora será utilizado a voz intitulada “Aryane - Cantora”, na Figura 15 e depois foi inserido o texto, na Figura 16.

Revista ARACÊ
ISSN: 2358-2472

Figura 15 - Voz Aryane -Cantora

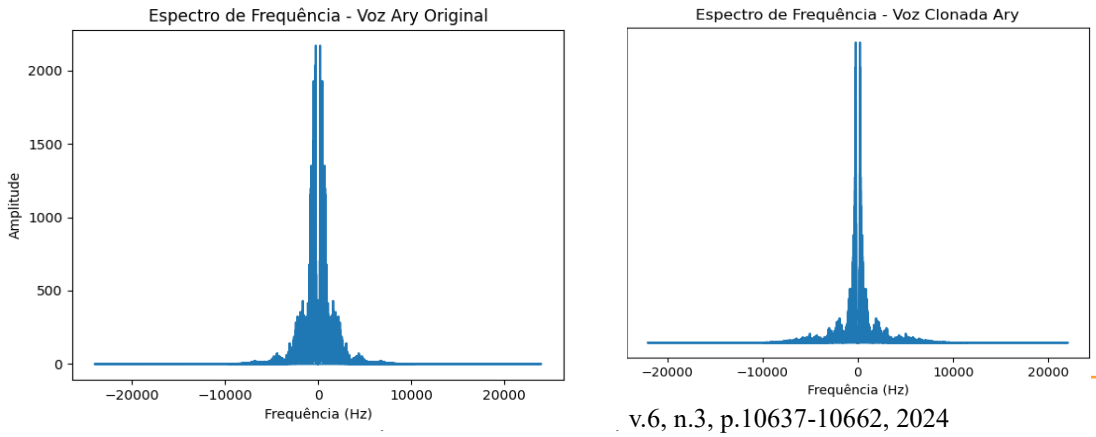


Fonte: ElevenLabs (2024)

Figura 16 – Inserindo texto para a voz Aryane -Cantora

Fonte: ElevenLabs

Figura 17 e Figura 18 – Voz Original e Clonada respectivamente.



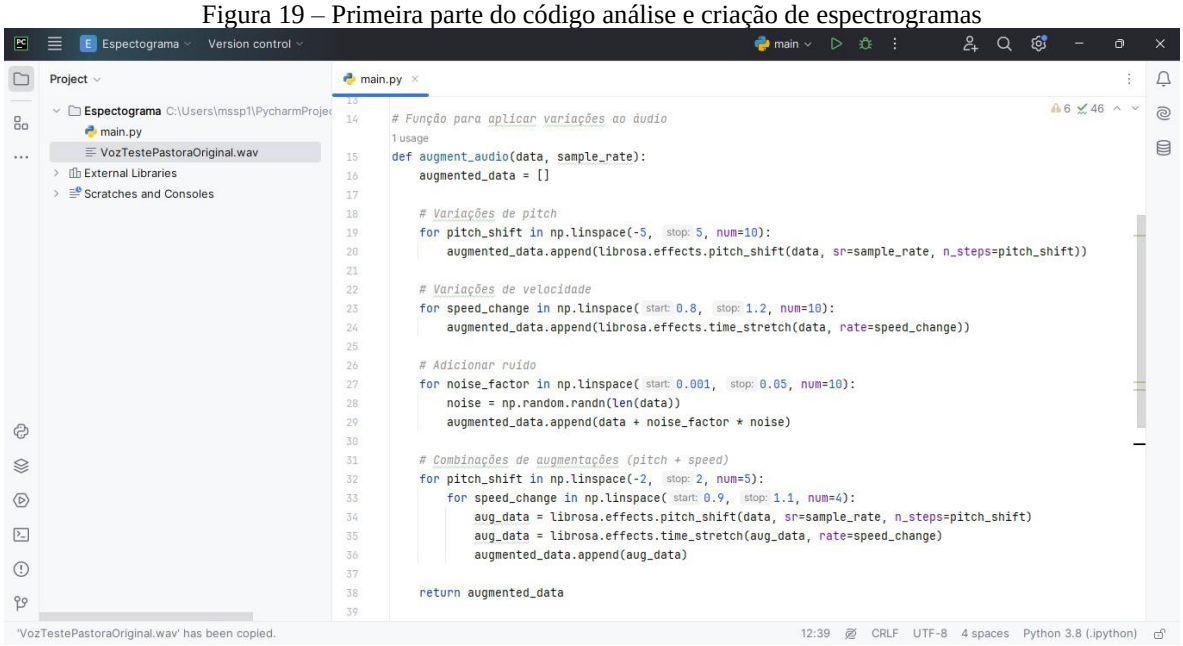
Revista ARACÊ
ISSN: 2358-2472

Na análise das figuras 17 e 18, percebemos que ambas as imagens demonstram que a maior parte da energia está concentrada em frequências baixas, perto de 0 Hz, dando a perspectiva que os sinais dos áudios são similares em questão de conteúdo de frequência básico. A figura 18 (Voz clonada Ary), indica que ela foi simplificada ou possui um alcance dinâmico menor em comparação a figura 17 (Voz Original Ary), umoutro ponto é que os picos de amplitude presentes na figura 18 (Voz clonada Ary), replicam as características principais da voz original, mas se perde na fidelidade ou alcance dinâmico, conforme indicado pela amplitude menor e as limitações representadas nas frequências.

5.2 CÓDIGO PARA ANÁLISE DE PADRÃO DO ESPECTROGRAMA

Após clonar as vozes, converter para o formato WAV e gerar os espectrogramasvárias vezes, surgiu a necessidade de criar um script para realizar a análise de padrão sonora e gerar novos espectrogramas. As figuras abaixo (19-21) mostram o script completo:

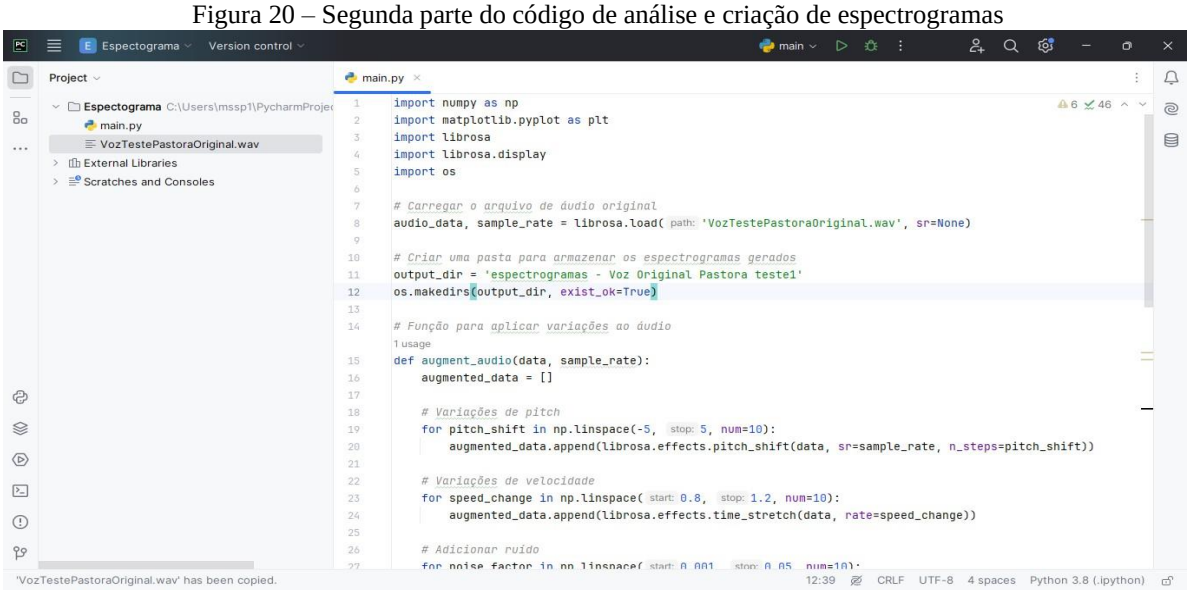
Figura 19 – Primeira parte do código análise e criação de espectrogramas



Fonte: Autor (2024)

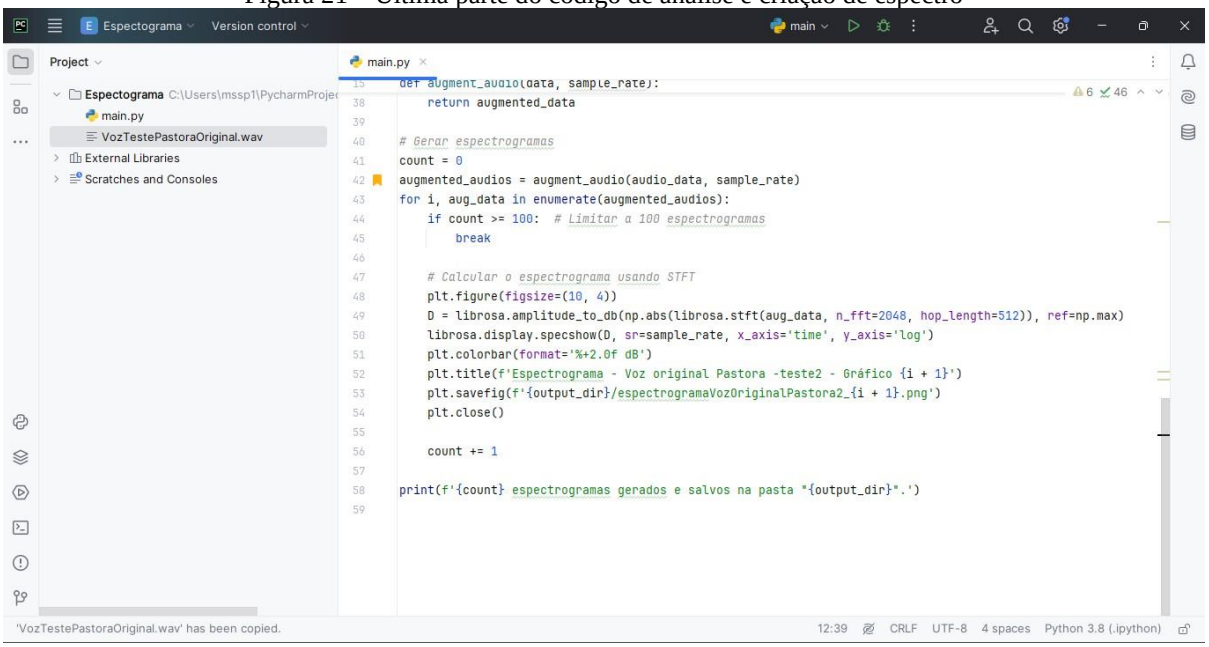
Revista ARACÊ
ISSN: 2358-2472

Figura 20 – Segunda parte do código de análise e criação de espectrogramas



Fonte: Autor (2024)

Figura 21 – Última parte do código de análise e criação de espectro

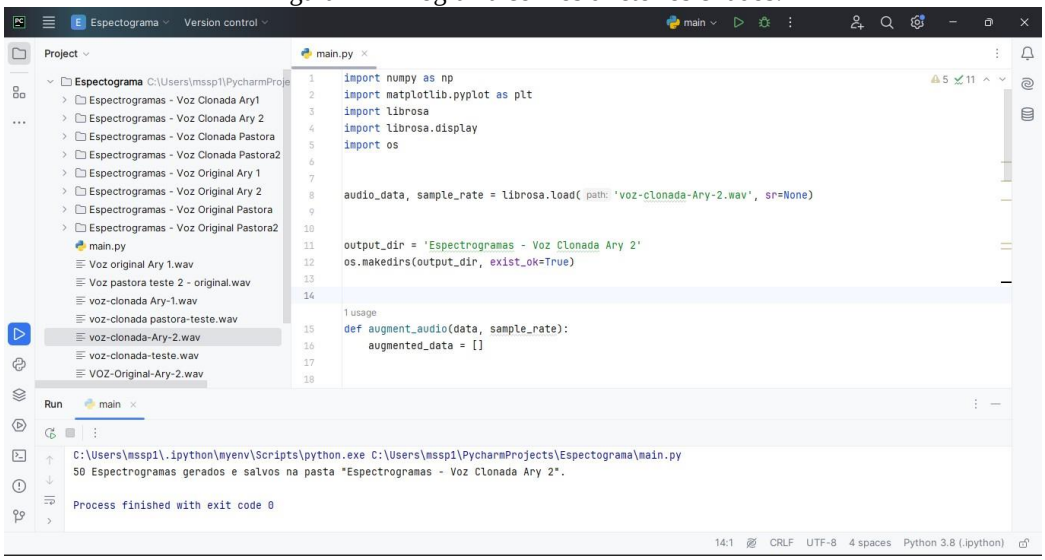


Fonte: Autor (2024)

Depois de colocar os áudios das vozes originais e das vozes clonadas da “Pastorae da Cantora Aryane”, foram criados alguns diretórios para os respectivos áudios, com cerca de 50 imagens de espectrogramas em cada diretório. A Figura 22, abaixo, ilustra esse processo:

Revista ARACÊ
ISSN: 2358-2472

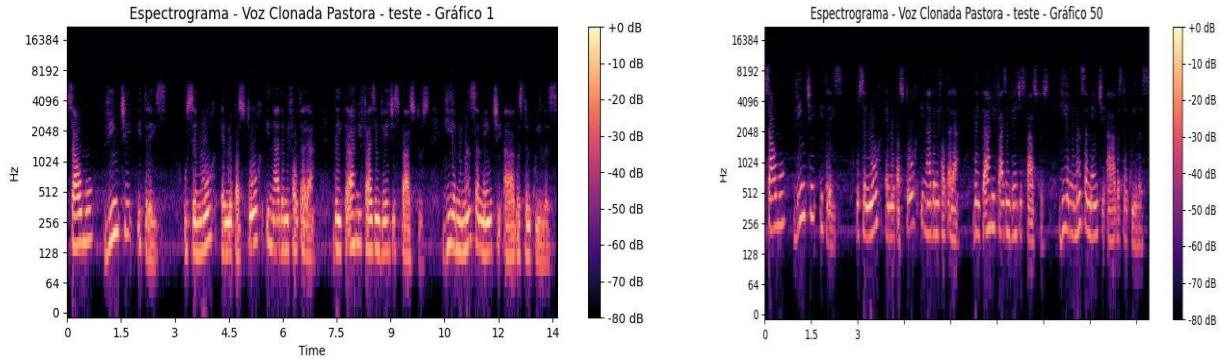
Figura 22 – Programa com os diretórios criados:



Fonte: Autor (2024)

Nota-se que a cada áudio processado, um diretório com diversas imagens de espectrogramas é criado. Esses espectrogramas são gerados com base nas transformações aplicadas ao sinal de áudio, sendo representados visualmente pelas variações de frequência ao longo do tempo. Como demonstrado nas figuras 23 e 24.

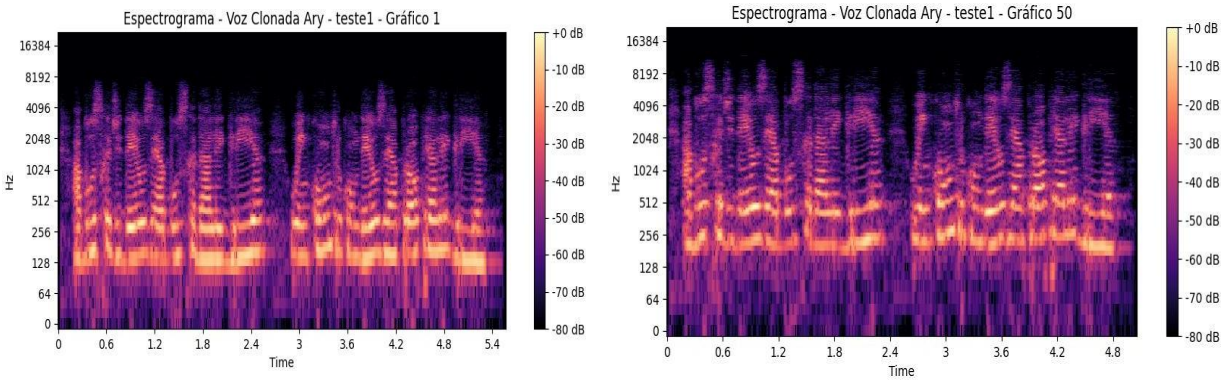
Figura 23 e Figura 24 – Espectrogramas da voz original da “Pastora”



Esses gráficos visualizam a intensidade das frequências de um sinal de áudio aolongo do tempo. As diferenças encontradas estão presentes no tempo total, onde a figura23, espectrograma da voz original da pastora, cobre em torno de 14 segundos, já a da figura 24, espectrograma da voz original da pastora, cobre em torno 22 segundos. Nas características visuais apresenta as faixas coloridas, semelhantes, indicando que se tratado mesmo áudio. Com tudo, apresenta pequenas diferenças nas distribuições dasfrequências mais altas em torno de 4096 Hz e acima, e na densidade das faixas em algumas regiões.

Revista ARACÊ
ISSN: 2358-2472

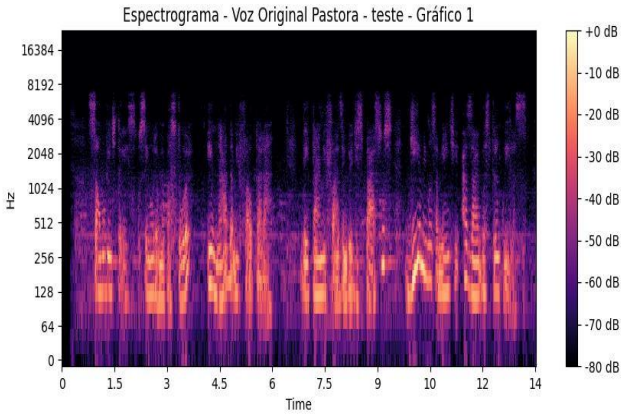
Figura 27 e Figura 28 – Espectrogramas da voz original “Aryane Cantora”



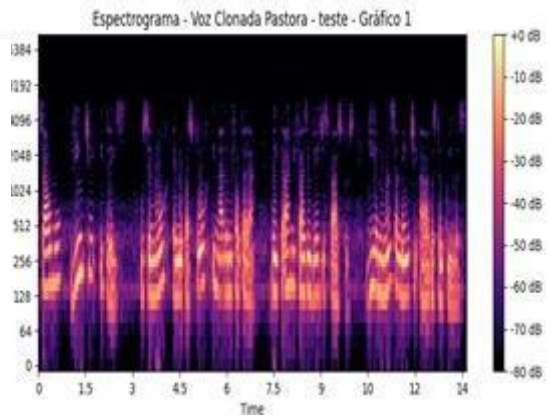
Nas figuras 27 e 28, esses gráficos apresentam escala de cor onde indica a intensidade do sinal em dB (decibéis), onde os tons claros representam maior intensidade e os escuros, menor intensidade. As frequências dominantes nos dois gráficos são similares, elas estão em 128 Hz e 1024 Hz, com componentes harmônicosem intervalos mais altos, o que é comum e fundamentais nas falas humanas. Na amplitude e energia, as duas figuras representam áreas com maior intensidade concentradas entre 128 Hz e 1024 Hz. Algo interessante encontrado nos componentes repetidos é que existem padrões repetidos que aparentemente ocorrem ciclicamente, onde é encontrado em fonemas ou sílabas no sinal de voz.

5.3 COMPARAÇÃO DOS ESPECTROGRAMAS

Figura 30 e Figura 31 – Espectrogramas voz original e clonada da “Pastora”



Fonte: Autor (2024)



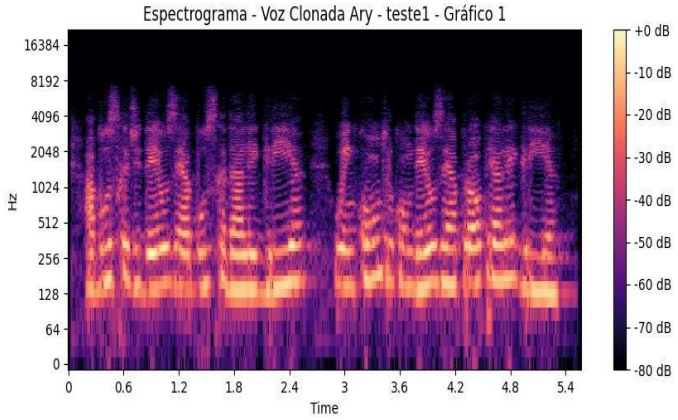
Fonte: Autor (2024)

Ao analisar as figuras 30 e 31, nota-se que ambas as imagens possuem padrõesde frequências bem parecidas, onde possuem pequenas variações por se tratar de uma voz original e a outra clonada.

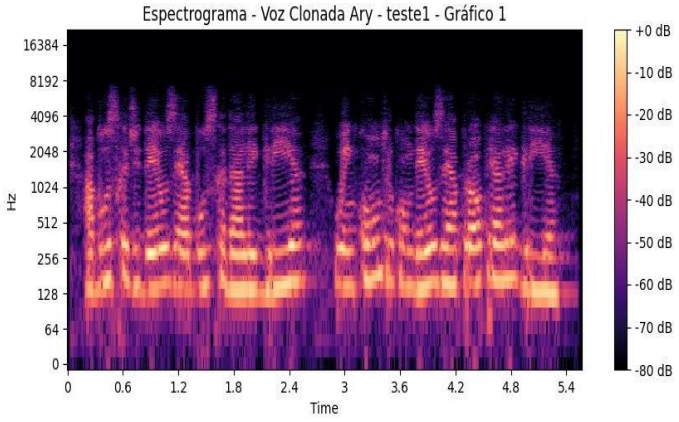
Revista ARACÊ
ISSN: 2358-2472

As faixas de frequências principais, possuem em torno de 64 Hz a 4096 Hz, contendo uma intensidade e um padrão visual similares. Outras diferenças presentes é que na figura 30, que se trata da voz original, a “energia” presenteé ligeiramente mais uniforme, enquanto a figura 31, voz clonada, possui uma variação leve nos harmônicos, as diferenças nas intensidades e nos detalhes de padrões de frequência podem sugerir que a voz clonada apresenta leves distorções.

Figura 32 e Figura 33 – Espectrogramas voz original e clonada da “Aryane Cantora”



Fonte: Autor (2024)



Fonte: Autor (2024)

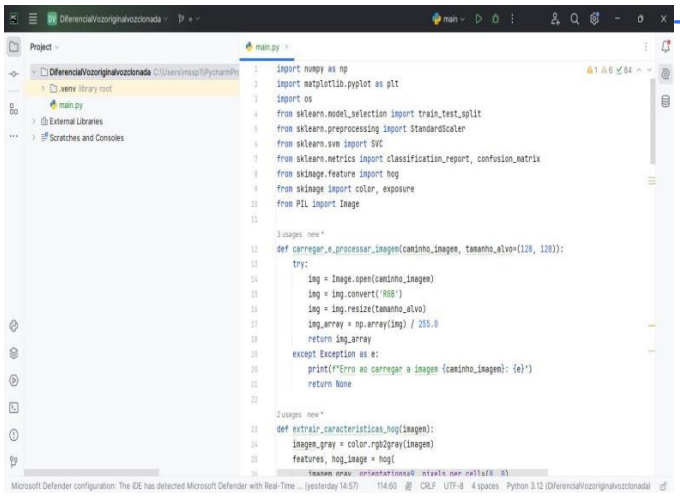
As duas figuras, mostram padrões semelhantes, demonstrando a eficácia da clonagem da voz, essa eficácia ocorre principalmente porque ela consegue manter as características harmônicas da gravação original. As diferenças mais aparentes se encontram nas pequenas variações na intensidade em algumas faixas de frequência e pequena duração da voz clonada.

Revista ARACÊ
ISSN: 2358-2472

5.4 SCRIPT PARA DIFERENCIAÇÃO DE VOZES BASEADOS EM ESPECTROGRAMA

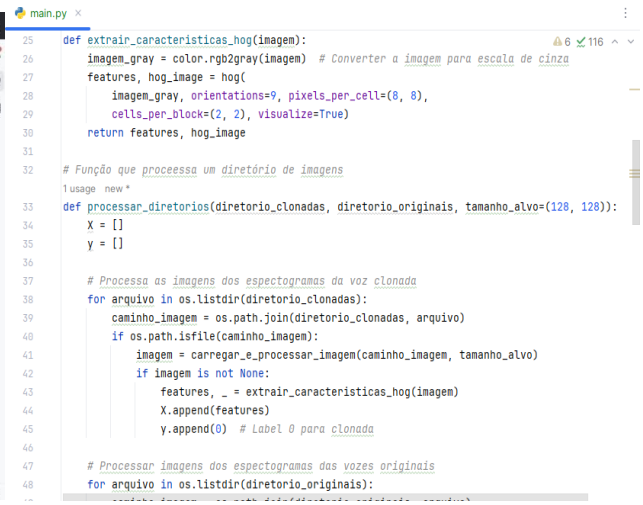
5.4.1 figura 34 – 39 – imagens do código completo

Figura 34 - Fonte: Autor (2024)



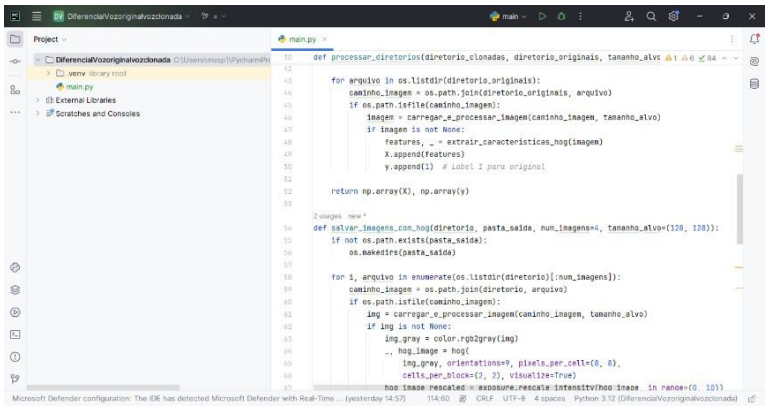
```
import numpy as np
import matplotlib.pyplot as plt
import os
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.svm import SVC
from sklearn.metrics import classification_report, confusion_matrix
from skimage.feature import hog
from skimage import color, exposure
from PIL import Image

def carregar_e_processar_imagem(caminho_imagem, tamanho_alvo=(128, 128)):
    try:
        img = Image.open(caminho_imagem)
        img = img.convert('RGB')
        img = img.resize(tamanho_alvo)
        img_array = np.array(img) / 255.0
        return img_array
    except Exception as e:
        print(f"Erro ao carregar a imagem {caminho_imagem}: {e}")
        return None

def extrair_caracteristicas_hog(imagem):
    imagem_gray = color.rgb2gray(imagem)
    features, hog_image = hog(
```

Figura 35 - Fonte: Autor (2024)



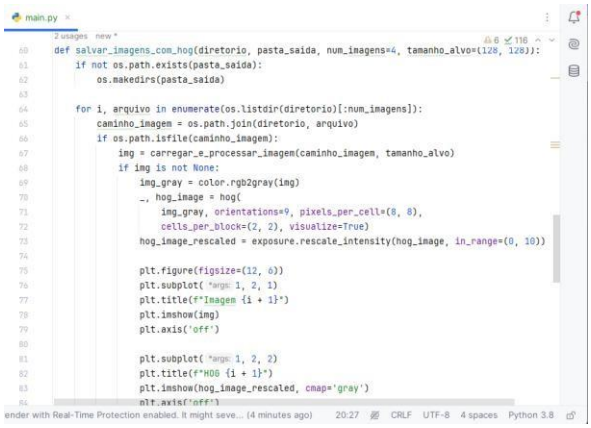
```
def extrair_caracteristicas_hog(imagem):
    imagem_gray = color.rgb2gray(imagem)  # Converter a imagem para escala de cinza
    features, hog_image = hog(
        imagem_gray, orientations=9, pixels_per_cell=(8, 8),
        cells_per_block=(2, 2), visualize=True)
    return features, hog_image

# Função que proceessa um diretório de imagens
def processar_diretorios(diretorio_clonadas, diretorio_originais, tamanho_alvo=(128, 128)):
    X = []
    y = []

    # Processa as imagens dos espectrogramas da voz clonada
    for arquivo in os.listdir(diretorio_clonadas):
        caminho_imagem = os.path.join(diretorio_clonadas, arquivo)
        if os.path.isfile(caminho_imagem):
            imagem = carregar_e_processar_imagem(caminho_imagem, tamanho_alvo)
            if imagem is not None:
                features, _ = extrair_caracteristicas_hog(imagem)
                X.append(features)
                y.append(0)  # Label 0 para clonada

    # Processar imagens dos espectrogramas das vozes originais
    for arquivo in os.listdir(diretorio_originais):
```

Figura 36 - Fonte: Autor (2024)



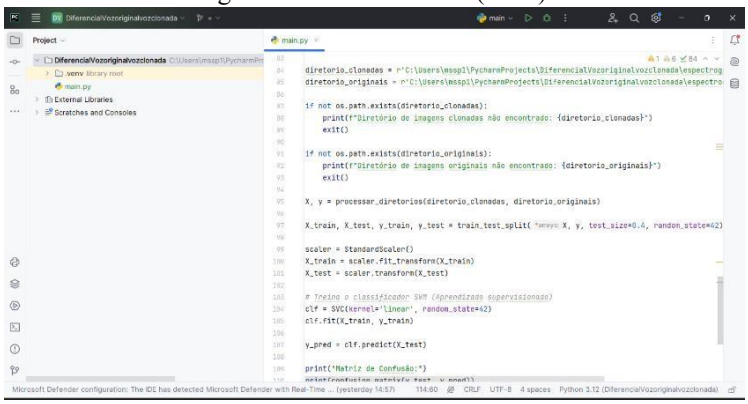
```
def processar_diretorios(diretorio_clonadas, diretorio_originais, tamanho_alvo=(128, 128)):
    for arquivo in os.listdir(diretorio_originais):
        caminho_imagem = os.path.join(diretorio_originais, arquivo)
        if os.path.isfile(caminho_imagem):
            imagem = carregar_e_processar_imagem(caminho_imagem, tamanho_alvo)
            if imagem is not None:
                features, _ = extrair_caracteristicas_hog(imagem)
                X.append(features)
                y.append(1)  # Label 1 para original

    return np.array(X), np.array(y)

def salvar_imagens_com_hog(diretorio, pasta_saida, num_imagens=4, tamanho_alvo=(128, 128)):
    if not os.path.exists(pasta_saida):
        os.makedirs(pasta_saida)

    for i, arquivo in enumerate(os.listdir(diretorio)[:num_imagens]):
        caminho_imagem = os.path.join(diretorio, arquivo)
        if os.path.isfile(caminho_imagem):
            img = carregar_e_processar_imagem(caminho_imagem, tamanho_alvo)
            if img is not None:
                img_gray = color.rgb2gray(img)
                _, hog_image = hog(
                    img_gray, orientations=9, pixels_per_cell=(8, 8),
                    cells_per_block=(2, 2), visualize=True)
                hog_image_rescaled = exposure.rescale_intensity(hog_image, in_range=(0, 10))

                plt.figure(figsize=(12, 6))
                plt.subplot(1, 2, 1)
                plt.title(f"Imagem {i + 1}")
                plt.imshow(img)
                plt.axis('off')

                plt.subplot(1, 2, 2)
                plt.title(f"HOG {i + 1}")
                plt.imshow(hog_image_rescaled, cmap='gray')
                plt.axis('off')
```

Figura 37 - Fonte: Autor (2024)



```
def salvar_imagens_com_hog(diretorio, pasta_saida, num_imagens=4, tamanho_alvo=(128, 128)):
    if not os.path.exists(pasta_saida):
        os.makedirs(pasta_saida)

    for i, arquivo in enumerate(os.listdir(diretorio)[:num_imagens]):
        caminho_imagem = os.path.join(diretorio, arquivo)
        if os.path.isfile(caminho_imagem):
            img = carregar_e_processar_imagem(caminho_imagem, tamanho_alvo)
            if img is not None:
                img_gray = color.rgb2gray(img)
                _, hog_image = hog(
                    img_gray, orientations=9, pixels_per_cell=(8, 8),
                    cells_per_block=(2, 2), visualize=True)
                hog_image_rescaled = exposure.rescale_intensity(hog_image, in_range=(0, 10))

                plt.figure(figsize=(12, 6))
                plt.subplot(1, 2, 1)
                plt.title(f"Imagem {i + 1}")
                plt.imshow(img)
                plt.axis('off')

                plt.subplot(1, 2, 2)
                plt.title(f"HOG {i + 1}")
                plt.imshow(hog_image_rescaled, cmap='gray')
                plt.axis('off')
```

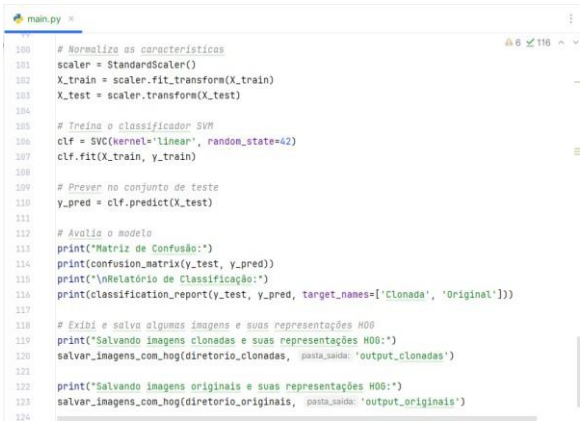
Figura 38 - Fonte: Autor (2024)



```
diretorio_clonadas = r'C:\Users\mssp1\PycharmProjects\DiferencialVozoriginalvozclonada\espectrog
diretorio_originais = r'C:\Users\mssp1\PycharmProjects\DiferencialVozoriginalvozclonada\espectro

if not os.path.exists(diretorio_clonadas):
    print(f"Diretório de imagens clonadas não encontrado: {diretorio_clonadas}")
    exit()

if not os.path.exists(diretorio_originais):
    print(f"Diretório de imagens originais não encontrado: {diretorio_originais}")
    exit()

X, y = processar_diretorios(diretorio_clonadas, diretorio_originais)

X_train, X_test, y_train, y_test = train_test_split( *arrays: X, y, test_size=0.4, random_state=42)

scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

# Treina o classificador SVM (Aprendizado supervisionado)
clf = SVC(kernel='linear', random_state=42)
clf.fit(X_train, y_train)

y_pred = clf.predict(X_test)

print("Matriz de Confusão:")
```

Revista ARACÊ
ISSN: 2358-2472

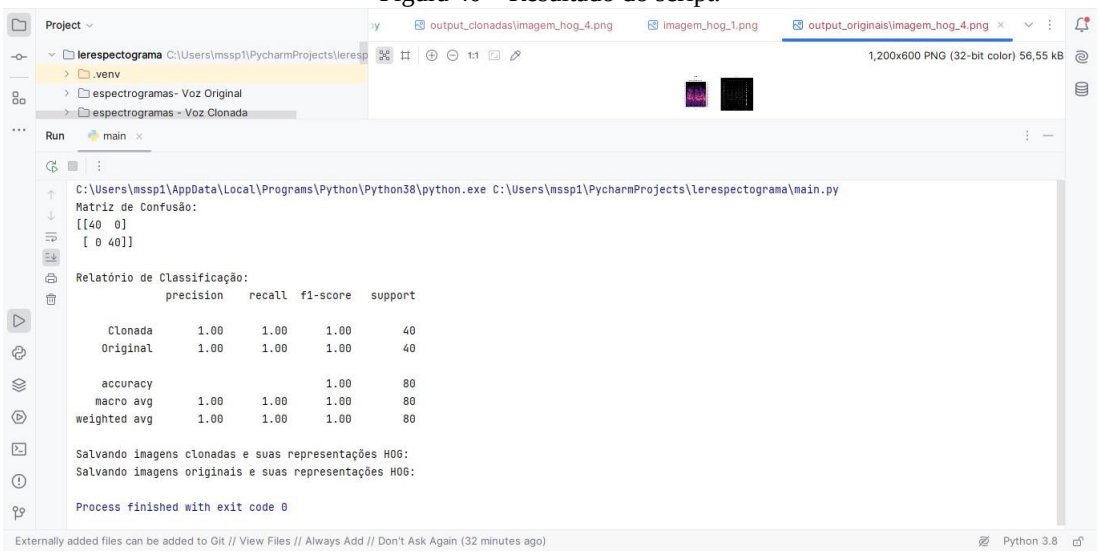
Figura 39 - Fonte: Autor (2024)



6 RESULTADOS

Com o uso do script encontrado no tópico 5.2 e mostrados nas figuras 34-39,obtém-se os seguintes resultados, como sugere as figuras 40 - 44.

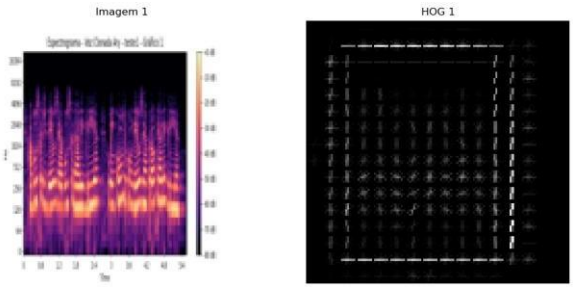
Figura 40 – Resultado do script.



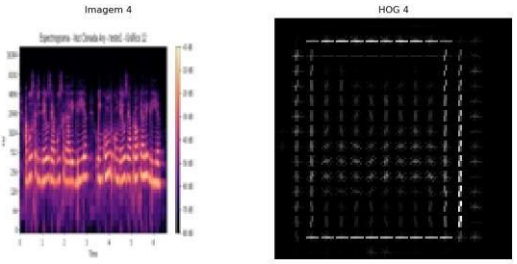
Fonte: Autor (2024)

Revista ARACÊ
ISSN: 2358-2472

Figura 41 e Figura 42 – Imagem HOG – Voz Clonada

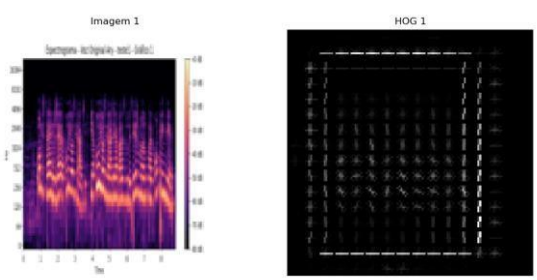


Fonte: Autor (2024)

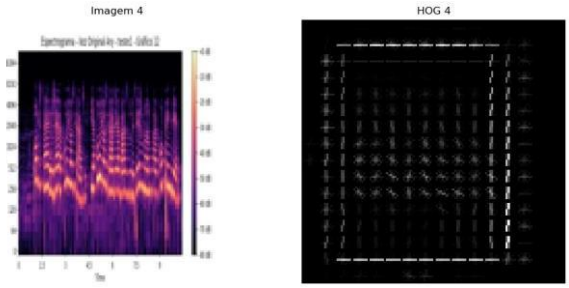


Fonte: Autor (2024)

Figura 43 e Figura 44 – Imagem HOG – Voz Original



Fonte: Autor



Fonte: Autor

A figura 40, nos retorna uma matriz de confusão, que basicamente é uma tabelaonde se é possível visualizar o desempenho do modelo em relação as classes, clonadas e original. No script ele retorna [40 0], que significa que o modelo classificou corretamente 40 amostras da classe “Clonada” corretamente, e [0 40], que significa queo modelo classificou 40 amostras da classe “Original” sem cometer erros.

Abaixo da matriz de confusão, também foi retornado um relatório de classificação, onde as métricas apresentam um valor de 1.00 para ambas as classes, ou seja, o desempenho foi considerado perfeito. Os aspectos avaliados foram *Precision* (Precisão): 1.00 para “Clonada” e 1.00 para “Original”, *Recall* (Revocação): 1.00 para ambas as classes, o que indica que todas as amostras de cada classe foram identificadascorretamente. *F1-Score:* 1.00 para ambas as classes, indicando uma harmonia perfeita entre a precisão e o *recall* e por último na acurácia total obteve um resultado de 100% num total de 80 amostras.

As figuras 41 a 44, representam imagens em HOG (*Histogram of Oriented Gradients)*, que é uma técnica de extração de características que captura informações detextura e contornos das imagens utilizadas em uma visão computacional. O que pode indicar que o sistema está registrando as representações para uma possível análise posterior, ou reutilização em novos testes, ou seja, está aprendendo os padrões das imagens.

Revista ARACÊ
ISSN: 2358-2472

7 CONCLUSÃO

Esta pesquisa analisou a habilidade de diferenciar vozes simuladas de vozes autênticas por meio da análise de espectrogramas e de ferramentas de inteligência artificial, como a ElevenLabs. Por meio da criação de scripts para a avaliação de padrõesespectrais e aplicação de métodos de Deep Learning e Machine Learning, conseguimosdistinguir diferenças significativas entre as frequências de vozes reais e as vozessintetizadas.

Os achados indicam que a aplicação de espectrogramas e a identificação de atributos como o HOG (Histogram of Oriented Gradients) são eficientes na distinção devozes. A matriz de confusão e as métricas de categorização demonstram a exatidão e consistência do modelo, alcançando 100% de precisão ao reconhecer as categorias "Clonada" e "Original". Estes resultados sugerem que técnicas de avaliação de frequência e padrões vocais podem proporcionar um nível extra de proteção contra fraudes e manipulações por DeepFakes de áudio.

Este estudo evidencia que análises computacionais de espectrogramas podem serrecursos úteis na identificação de vozes falsas. Em um contexto em que a tecnologia DeepFakes é cada vez mais empregada em situações problemáticas, como fraudes, é crucial criar técnicas eficientes para identificar tais manipulações, a fim de salvaguardar a integridade das informações e a segurança da população.

Revista ARACÊ
ISSN: 2358-2472

REFERÊNCIAS

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

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

Fanaya, Patrícia Fonseca. Deepfake e a realidade sintetizada. TECCOGS – Revista digital de tecnologias cognitivas, n.23, jan./jun.2021, p. (104-118). Acessado em: 21/04/2024. Disponível em: Vista do Deepfake e a realidade sintetizada (pucsp.br)

IBM Brasil. O que é Deep Learning? Acesso em: 20/04/2024, Disponível em:https://www.ibm.com/br-pt/topics/deep-learning

CNN. Saiba o que é deepfake, técnica de inteligência artificial que foi apropriada para produzir desinformação. Acesso em: 02/06/2024. Disponível em https://www.cnnbrasil.com.br/noticias/saiba-o-que-e-deepfake-tecnica-de-inteligencia- artificial-que-foi-apropriada-para-produzir-desinformacao/

ComunitIA. ElevenLabs. Acesso em: 02/06//2024. Disponível em: https://www.comunitia.com/ferramenta/eleven-labs

Convertio. Conversor de áudio. Acesso em: 03/06/2024. Disponível em: ConverterMP3 em WAV (Online e Gratuito) — Convertio

IIElevenLabs. ElevenLabs. Acessado em: 02/06/2024. Disponível em: AI VoiceGenerator & Text to Speech | ElevenLabs

Kaspersky daily. Não acredite em tudo o que ouve: deepfakes de voz (2023). Acesso em:23/04/2024. Disponível em: Não acredite em tudo o que ouve: deepfakesde voz (kaspersky.com.br).

Masood, M., Nawaz., Malik, K.M., Ali, J. & Irtaza A. (2021) Deepfakes Generationand Detection: State-of-the-art, open challenges, countermeasures, and way forward. Cornell University. Acessado em: 23/04/2024. Disponível em:2103.00484.pdf (arxiv.org)

Nishida, M. Silvia., Weber, Silke.Anna. Theresa., Oliveira, Felipe Augusto K de. &Troll Juliana. Voz Humana. Nadi. Acesso em: 23/04/2024. Disponível em Nossa Missão (unesp.br).

Ponti, M., & Costa, G. (2017). Como funciona a Deep Learning (p. 63-93). ISBN978-85-7669-400-7. Acesso em 21/04/2024. Disponível em: Ponti_Costa_Como- funciona-o-Deep-Learning_2017.pdf (usp.br)

Sichman, Jaime. Inteligência Artificial e sociedade: Avanços e Riscos. Scielo Brasil, 35, p. (37-49),2021. Acesso em: 20/04/2024. Disponível em: SciELO - Brasil - Inteligência Artificial e sociedade: avanços e riscos Inteligência Artificial e sociedade:avanços e riscos.

Revista ARACÊ
ISSN: 2358-2472

Spencer, Michael K., tradução: Gabriela Leite (2019). DeepFake, a mais recente ameaça distópica. OutrasPalavras. Acessado em: 21/04/2024. Disponível em: DeepFake, a mais recente ameaça distópica - Outras Palavras

UOL Notícias. Deepfake: uso de inteligência artificial em eleições na Argentina enos Estados Unidos. Acesso em: 21/04/2024. Disponível em: https://noticias.uol.com.br/confere/ultimas-noticias/2024/03/03/deepfake-uso- inteligencia-artificial-eleicoes-argentina-estados-unidos.htm

Your Physicist, 4 tipos mais comuns de técnicas de análise espectral. Acessado em: 23/04/2024. Disponível em: 4 tipos mais comuns de técnicas de análise espectral (your-physicist.com)