Identificação de Deep Fakes

1st Amir Youssef dos Santos

Bacharelado em Inteligência Artificial
Universidade Federal de Goiás
Instituto de Informática (INF)
Goiânia, Brasil
amiryoussef@discente.ufg.br

2nd Bernardo Aires de Oliveira

Bacharelado em Inteligência Artificial
Universidade Federal de Goiás
Instituto de Informática (INF)
Goiânia, Brasil
bernardoaires@discente.ufg.br

3th Daniel Machado Pedrozo

Bacharelado em Inteligência Artificial
Universidade Federal de Goiás
Instituto de Informática (INF)
Goiânia, Brasil
danielpedrozo@discente.ufg.br

4th Julia Soares Dollis

Bacharelado em Inteligência Artificial
Universidade Federal de Goiás
Instituto de Informática (INF)
Goiânia, Brasil
juliadollis@discente.ufg.br

***Resumo* — Esse artigo apresenta uma abordagem para o processamento de áudio e a identificação de tons usando a combinação de diferentes técnicas de programação e fundamentos matemáticos em Python para que seja possível o desenvolvimento do projeto. O objetivo principal do artigo é discorrer sobre experimentos relacionados ao estudo de ondas, área considerada crucial para a eficácia das técnicas aplicadas a favor de processar áudios e identificar tons presentes. (*Abstract*)**

Palavras chaves — Fast Fourier Transform, classificação de tons, processamento de áudio, filtros, deep fakes. (*key words****)***

1. INTRODUÇÃO (*HEADING 1*)

Atualmente, com o avanço rápido e progressivo da inteligência artificial, tornou-se possível replicar, através de vários modelos, a voz e a imagem de várias pessoas, muitas vezes, pessoas públicas, cantores, atores, influenciadores digitais. Por mais que isso seja muito benéfico em várias ocasiões, como por exemplo a dublagem de línguas em filmes e séries, há um lado obscuro que essas tecnologias possibilitam. Pessoas más intencionadas acabam por criar vídeos e áudios onde pessoas reais aparentam falar coisas que não falaram de fato e até mesmo crimes, as quais muitas vezes podem ser extremamente danosas à reputação dessas. Esses vídeos e áudios são chamados de deep fakes, portanto, surge a necessidade de identificar se um vídeo é gerado por um modelo de IA ou se é de fato uma fala de alguém.

Propomos explorar as soluções existentes, examinando como diferentes abordagens e tecnologias, especialmente no processamento digital de sinais e imagens, são empregadas para detectar discrepâncias e incoerências que diferenciam áudios genuínos de falsificados. Além da revisão teórica, o estudo inclui uma aplicação prática, utilizando técnicas como a Fourier transform (FFT) e a Fast Fourier Transform (STFT) para análise espectral, com o objetivo de ilustrar como as sutilezas nas características do áudio podem ser efetivamente utilizadas para identificar deep fakes.

2. FUNDAMENTOS TEÓRICOS

A. Mecanismos e Técnicas:

Aprendizado de Máquina Supervisionado: Treinamento de modelos com conjuntos de dados etiquetados de áudios autênticos e falsificados (sintetizados artificialmente). No contexto supracitado, o modelo aprende a identificar características específicas que diferenciam áudios reais de sintéticos.

Regressão Logística: Método estatístico usado para prever a probabilidade de ocorrência de uma variável categórica. Este modelo é adequado para problemas de classificação binária (classificar áudios como autênticos ou deep fakes) e opera estimando a probabilidade de ocorrência de um evento, baseando-se nas características fornecidas durante o treinamento.

Fourier Transform: Conceito matemático utilizado para transformar um sinal do domínio do tempo para o domínio da frequência. Sendo crucial para analisar o espectro de frequência do áudio e identificar anomalias que podem ser indícios de manipulação, servindo como parâmetros para o treinamento do modelo.

Fast Fourier Transform: Extensão da Transformada de Fourier, sendo utilizada dividindo o sinal em pequenos segmentos de tempo e aplicando a FFT a cada segmento, o que é particularmente útil para sinais não estacionários (cuja frequência muda com o tempo), como é frequentemente o caso em áudios deep fake.

Métricas de Avaliação: Precisão, recall e F1-Score visando mensurar a qualidade da distinção do modelo entre áudios autênticos e falsificados.

B. Possível Solução do Problema:

Propomos desenvolver, para a detecção de deep fakes em áudios, um modelo de Regressão Logística que é treinado utilizando características cruciais extraídas por meio da Transformada Rápida de Fourier (FFT) e da Transformada de Fourier de Tempo Curto (STFT). Essas técnicas são fundamentais para analisar as propriedades espectrais dos áudios, permitindo que o modelo identifique sutilezas e anomalias que diferenciam áudios autênticos de sintéticos.

3. METODOLOGIA

A. Coleta e Preparação dos Dados:

Conjunto de Dados: Inicialmente, realizamos gravações e coletas de diversos conjuntos de áudios de voz humana. Ao fim, resultados em um dataset contendo áudios autênticos e sintéticos. Estes áudios foram etiquetados manualmente, a fim de possibilitar o treinamento supervisionado.

Os áudios foram sintetizados a partir do XTTS[1], modelo pré treinado “text to speech". O modelo, além da síntese, permite clonagem com boa qualidade usando áudios de

apenas 3 segundos como referência, tornando, assim, muito mais fácil e acessível a fabricação do Deep Fake.

O modelo também permite clonagem de voz em outros idiomas, além do da entrada, já que o modelo é pré-treinado em 16 idiomas diferentes: inglês (en), espanhol (es), francês (fr), alemão (de), italiano (it), português (pt), polonês (pl), turco (tr) , russo (ru), holandês (nl), tcheco (cs), árabe (ar), chinês (zh-cn), japonês (ja), húngaro (hu) e coreano (ko).

Outro ponto positivo é o uso da API, e com isso, a clonagem pode ser feita com 5 linhas de código ou por linha de comando, e sem necessidade de GPU.

Dessa forma, foi realizada a clonagem da voz em portugues de uma locutora feminina, produzindo áudios de, em média, 10 segundos. Em seguida, o áudio sintético foi agrupado com um aúdio original da mesma locutora com duração de 10 segundos também. Assim, foi criado um áudio de 20 segundos com 10 segundos iniciais de voz clonada e os 10 segundos finais de voz humana.

Além da locutora feminina, também usamos a voz de um locutor masculino para diversificar o DataSet.

Dessa forma, nosso conjunto de dados consistia em 3 áudios de 20 segundos, tendo 10 segundos de voz sintética e 10 segundos de voz sintética, sendo dois áudios de locutora feminina e um áudio com locutor masculino. Os áudios possuem 48000 de frequência de amostragem. Os áudios podem ser vistos na figura 1.

B. Pré-processamento:

A etapa inicial no pipeline de processamento de dados envolve a implementação de técnicas de pré-processamento robustas para assegurar a homogeneidade e a integridade dos dados de áudio. Este procedimento é crucial para minimizar discrepâncias e variabilidades que poderiam potencialmente enviesar os resultados da análise subsequente.

Foi realizada uma segmentação que permite a comparação sistemática e confiável entre os diferentes áudios. A uniformidade na duração das amostras é vital para garantir a consistência na análise espectral, particularmente quando aplicamos transformações como a FFT e a STFT. Ao segmentar os áudios em unidades temporais padronizadas, garantimos que cada segmento contribua igualmente para o perfil analítico do conjunto de dados, eliminando quaisquer distorções decorrentes de variações no comprimento das gravações.

C. Análise Espectral:

Aplicação da FFT: Implementada para transpor os sinais de áudio do domínio do tempo para o domínio da frequência. Esta transformação é fundamental para decompor o sinal de áudio em seus componentes de frequência constituintes, permitindo uma análise detalhada das características espectrais. Através da FFT, é possível identificar padrões únicos e anomalias no espectro de frequência que podem indicar manipulações ou alterações, características típicas de áudios deep fake. A análise espectral proporcionada pela FFT oferece insights críticos sobre a estrutura e a composição frequencial dos sinais, elementos essenciais para a fase subsequente de classificação.

Aplicação da STFT: Para complementar a análise realizada pela FFT, empregamos a STFT. Essa técnica é particularmente eficaz para sinais que exibem variações de frequência ao longo do tempo, como é frequentemente o caso em gravações de áudio manipuladas. A STFT envolve a divisão do sinal de áudio em segmentos temporais menores e a aplicação da Transformada de Fourier a cada um desses segmentos. Isso permite uma análise mais granular das variações temporais no espectro de frequência. Ao aplicar a STFT, conseguimos capturar as nuances e as flutuações temporais nas características espectrais, o que é crucial para identificar as sutilezas na fabricação de áudios deep fake. Os espectrogramas da voz real e do deepfake podem ser vistos nas figuras 1 e 2 respectivamente.

D. Extração de Características:

Características Espectrais: Foram realizadas análises significativas dos espectros de frequência, como a amplitude em determinadas bandas de frequência, a fim de obter informações valiosas para o entendimento completo do funcionamento do conjunto de dados criado anteriormente. A partir da extração e análise dessas características, foi possível definir e compreender como deveria ser a metodologia seguinte.

E. Treinamento do Modelo:

A fase de treinamento do modelo é um componente primordial do processo da solução proposta, onde aplicamos rigorosas técnicas de aprendizado de máquina, visando construir um modelo eficaz, denso e preciso.

O modelo escolhido foi a Regressão Logística, um algoritmo de classificação robusto e amplamente reconhecido no campo do aprendizado de máquina supervisionado. Esta escolha é motivada pela capacidade do modelo de efetuar classificações binárias precisas, o que é essencial para o nosso objetivo de diferenciar entre áudios autênticos e falsificados. A Regressão Logística opera estimando probabilidades através de uma função logística, uma abordagem particularmente eficaz quando lidamos com variáveis dependentes categóricas, como é o caso na distinção entre áudios reais e gerados por inteligência artificial (IA).

F. Avaliação do Modelo:

A avaliação do modelo constitui uma etapa crítica no processo de desenvolvimento, essencial para verificar a eficácia e a precisão do modelo de Regressão Logística na identificação de áudios sintéticos

Após a fase de treinamento, o modelo é submetido a um teste rigoroso utilizando um conjunto de dados independente, que não foi utilizado durante o treinamento. Este conjunto de dados de teste é cuidadosamente selecionado para representar a diversidade e a complexidade dos cenários reais que o modelo enfrentará. Essa separação entre dados de treinamento e teste é crucial para garantir a validação objetiva do modelo, evitando o risco de sobreajuste (*overfitting*), onde o modelo se ajusta demais aos dados de treinamento e falha em generalizar para novos dados.

A eficácia do modelo é avaliada com base em um conjunto de métricas estatísticas robustas e amplamente reconhecidas na área de aprendizado de máquina:

Precisão: Mede a proporção de identificações corretas de áudios deep fake em relação ao total de casos classificados como deep fake pelo modelo. Uma alta precisão indica que o modelo tem uma taxa baixa de falsos positivos.

Recall (Sensibilidade): Avalia a capacidade do modelo de identificar corretamente áudios deep fake dentre todos os casos reais de deep fakes. Um alto recall é indicativo de que o modelo detecta eficientemente a maioria dos áudios falsificados.

F1-Score: Combina precisão e recall em uma única métrica, fornecendo um equilíbrio entre as duas. O F1-Score é particularmente útil em situações onde é crucial manter um equilíbrio entre minimizar falsos positivos e maximizar a identificação correta de casos positivos.

Os resultados obtidos a partir destas métricas são cuidadosamente analisados para interpretar a performance do modelo. Uma alta pontuação em todas essas métricas indica um modelo bem-sucedido, capaz de identificar de forma

confiável áudios deep fake. Por outro lado, pontuações mais baixas podem indicar a necessidade de revisão e ajuste do modelo, seja no processo de extração de características, seja na configuração do próprio algoritmo de Regressão Logística.

G. Iteração e Otimização:

A fase de iteração começa com uma análise minuciosa dos resultados obtidos durante a avaliação do modelo. Esta análise envolve a investigação dos casos em que o modelo falhou em detectar corretamente um deep fake ou identificou incorretamente um áudio autêntico como falsificado. Compreender esses erros é crucial para identificar áreas de melhoria no modelo e na estratégia de extração de características.

Em segundo plano, a otimização dos parâmetros do modelo é uma parte integrante deste processo. Isso envolve ajustar variáveis como a taxa de aprendizado, o número de iterações, e outros hiperparâmetros que influenciam diretamente o desempenho do modelo. A otimização é guiada pelo objetivo de melhorar as métricas de precisão, recall e F1-Score, equilibrando a capacidade do modelo de detectar deep fakes com a minimização de falsos positivos.

Este processo de iteração e otimização é cíclico e contínuo, com cada ciclo buscando aprimorar o desempenho do modelo. Através de um processo iterativo de testar, avaliar e ajustar, o modelo é constantemente refinado para melhor responder às complexidades envolvidas na detecção de deep fakes.

4. RESULTADOS E CONCLUSÕES

Desse modo, observamos que fomos capazes de desenvolver um modelo altamente eficaz para abordar nossa questão de pesquisa! Isso pode ser claramente visto de acordo com nossa matriz de confusão. Ademais, através dela, conseguimos inferir que a acurácia, a precisão e o recall foram todos ‘1’, ou seja, 100%.

Contudo, é importante considerar que as excepcionais métricas alcançadas podem ser atribuídas à limitada variabilidade em nosso conjunto de dados. Especificamente, nossa análise foi fundamentada exclusivamente nos registros sonoros de 2 pessoas, além de utilizar áudios gerados por um único modelo, o XTTS. Esta condição sugere que a performance superior pode não necessariamente refletir uma robustez geral do modelo, mas sim uma especialização no contexto muito específico proporcionado pelo conjunto de dados em questão.

REFERÊNCIAS

[1]: Coqui AI. (n.d.). Coqui TTS: A deep learning toolkit for Text-to-Speech, battle-tested in research and production. GitHub. Acessado em 22 de Janeiro de 2024, de https://github.com/coqui-ai/TTS

[2]: Khanjani, Z., Watson, G. e Janeja, VP (2023). Deepfakes de áudio: uma pesquisa. Fronteiras em Big Data , 5 , 1001063.

[3]: Dewesoft. (n.d.). Short-time Fourier Transform. In Dewesoft X Manual. Acessado em 21 de Janeiro de 2024, de https://manual.dewesoft.com/x/setupmodule/modules/general/math/freqdomainanalysis/stft

[4]: 2023-2-INF0413-PDSI. (n.d.). Materiais de INF0413. GitHub. Acessado em 20 de Janeiro de 2024, de https://github.com/2023-2-INF0413-PDSI/materiais

[5]: NTi Audio. (n.d.). Fast Fourier Transformation (FFT) - Basics. Acessado em 20 de Janeiro de 2024, de https://www.nti-audio.com/en/support/know-how/fast-fourier-transform-fft

[6]:http://lef.mec.puc-rio.br/wp-content/uploads/2019/06/Tranformada-de-Fourier-de-Tempo-Curto.pdf

[7]:Almutairi, Z. e Elgibreen, H. (2022). Uma revisão dos métodos modernos de detecção de deepfake de áudio: desafios e direções futuras. Algoritmos , 15 (5), 155.

[8]Mcuba, M., Singh, A., Ikuesan, RA e Venter, H. (2023). O efeito dos métodos de aprendizagem profunda na detecção de áudio deepfake para investigação digital. Procedia Ciência da Computação , 219 , 211–219.

APÊNDICES

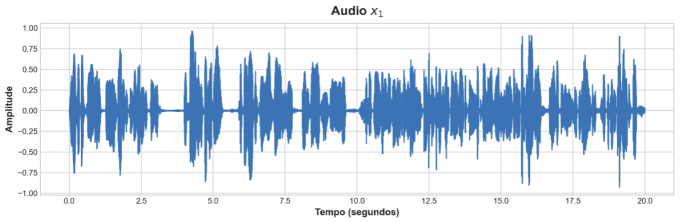

Figura 1

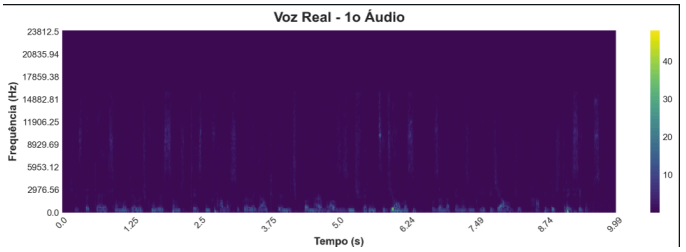

Figura 2

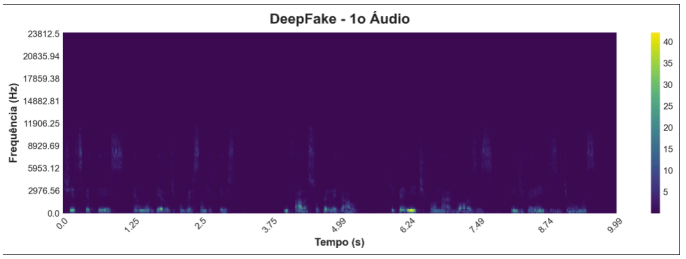

Figura 3

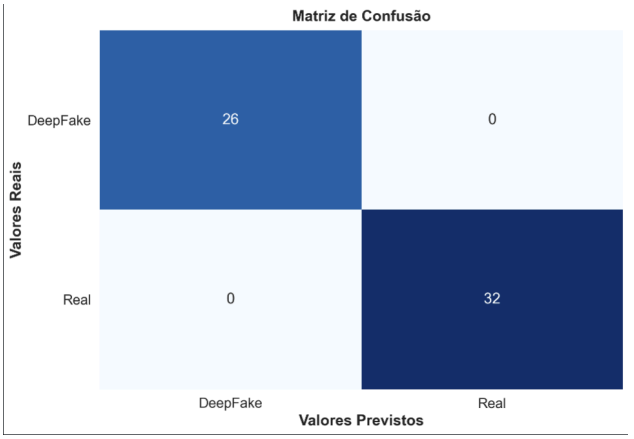

Figura 4