

# **Avanços na Clonagem de Voz: Uma Perspectiva do Sistema de Síntese de Fala TorToise**

**André Castro, Dayane Rodrigues, Lisandra Menezes e  
Michael Silva**  
Professor: Aldo André Dias Salazar

2023

# SUMÁRIO

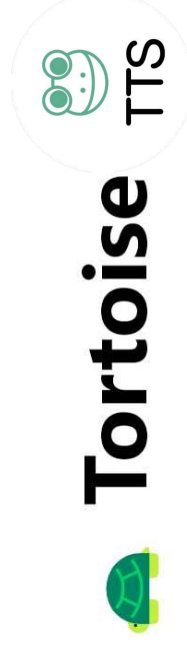
- ▶ Introdução
- ▶ Fundamentos Teóricos
- ▶ Metodologias
- ▶ Resultados
- ▶ Áudios Clonados
- ▶ Conclusão



# Introdução

# Introdução ao Desafio da Síntese de Fala

- **Objetivo:** Explorar Aplicações de Clonagem de Voz
- **TorToise:** O artigo de James Betker propõe o uso do sistema TorToise, utilizando técnicas de transformadores autoregressivos e Modelos Probabilísticos de Difusão com Desruído (DDPMs) para criar uma síntese de fala que consiga ser expressiva e multi-voz.



## Better speech synthesis through scaling

James Betker

### Abstract

In recent years, the field of image generation has been revolutionized by the application of autoregressive transformers and DDPMs. These approaches model the process of image generation as a step-wise probabilistic processes and leverage large amounts of compute and data to learn the image distribution.

This methodology of improving performance need not be confined to images. This paper describes a way to apply advances in the image generative domain to speech synthesis. The result is TorToise - an expressive, multi-voice text-to-speech system.

All model code and trained weights have been open-sourced at <https://github.com/neonbjb/tortoise-tts>.



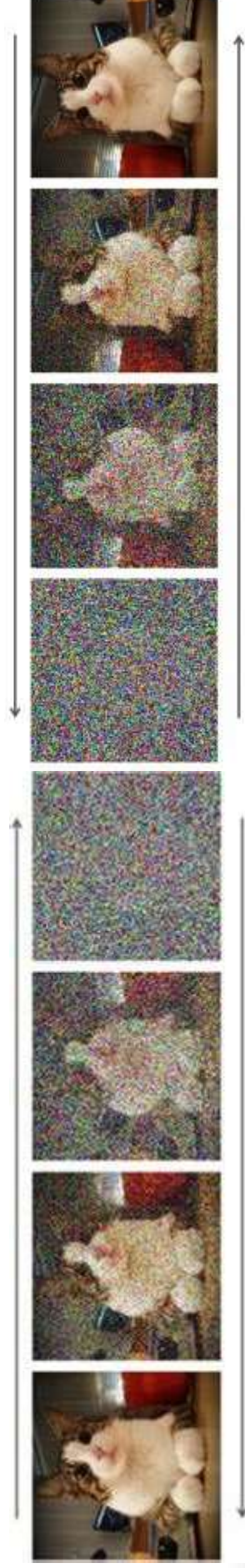


# Fundamentos Teóricos

---

# Fundamentos Teóricos da Metodologia TorToise

- Transformadores Autorregressivos: Estrutura poderosa para capturar a complexidade da fala, modelando dependências de longo alcance.
- DDPMs: Processo iterativo de difusão e *noise-free* que gera ondas sonoras realistas, partindo de ruído inicial e refinando até obter alta fidelidade.

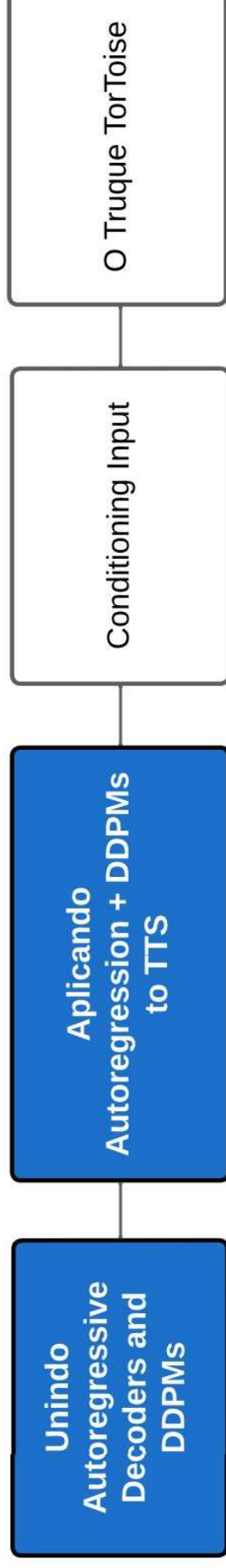




# Metodologia

---

# Metodologia







# Resultados

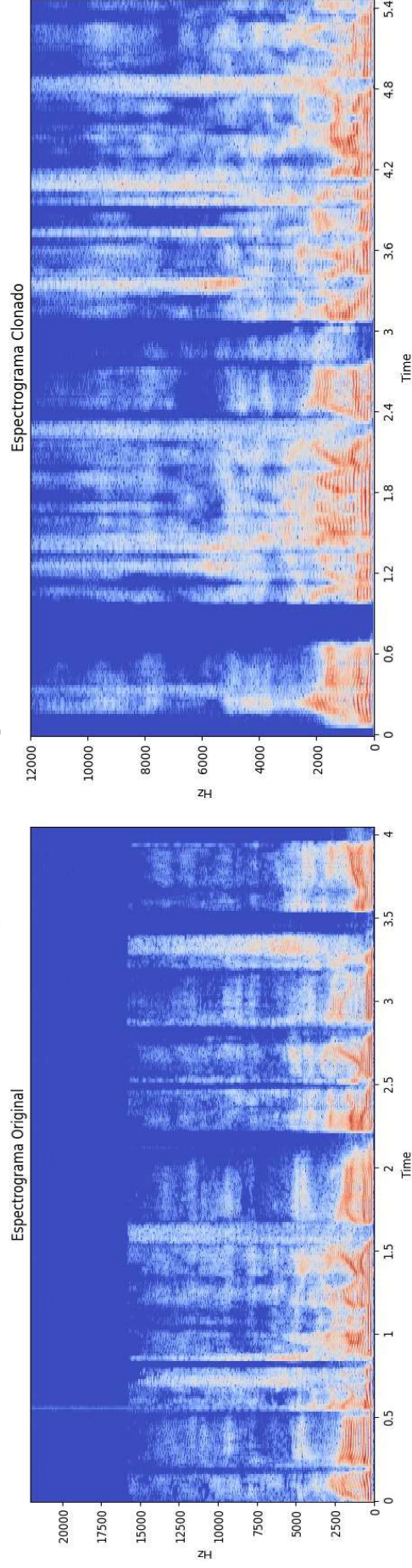
---

# Resultados

- Para esboçar nossos resultados utilizamos dos recursos apresentados em sala *espectrogramas*, *transformada de fourier* e uma abordagem complementar *Mel-frequency Cepstral Coefficients* (MFCCs);
- Ademais, apresentaremos três áudios um áudio original, a saber, um trecho extraído de uma entrevista da cantora norte-americana Miley Cyrus, um áudio com a voz clonada da artista contendo o mesmo conteúdo textual do áudio original, e um áudio da voz clonada com um conteúdo distinto.

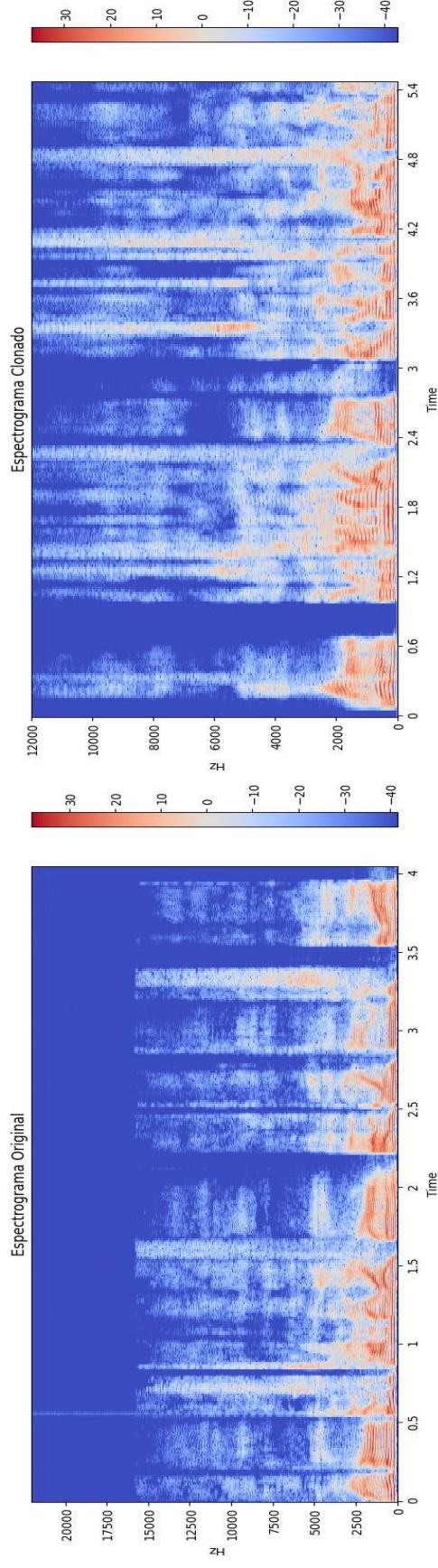


# Espectrogramas



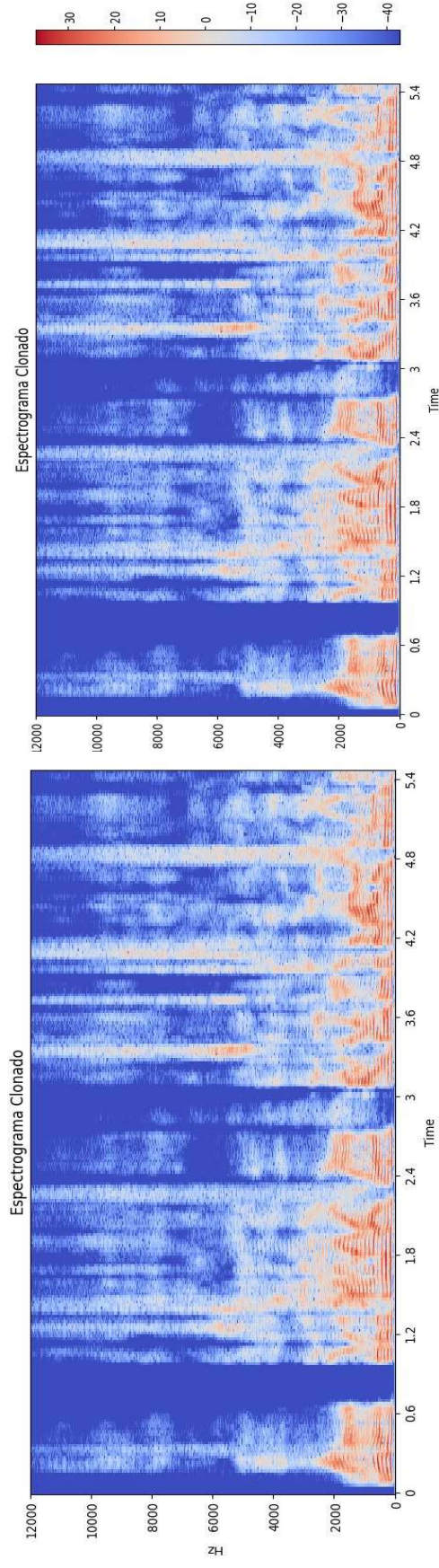
O espectrograma pode ser utilizado como uma representação visual do espectro de frequência do sinal ao longo do tempo. Na primeira dupla temos uma comparação entre o áudio original vs o áudio clonado com o mesmo conteúdo. Pode-se analisar, a partir dessa representação, características como entonação, timbre e qualidade sonora.





Na segunda dupla temos uma comparação entre o áudio original vs o áudio clonado com o mesmo conteúdo. Vemos determinados padrões de frequência entre os áudios clonados e o áudio original, que podem demonstrar similaridades entre as características da fala. Tais padrões na imagem, que se repetem, manifestam-se por meio da recorrência de cores e ondas do sinal.



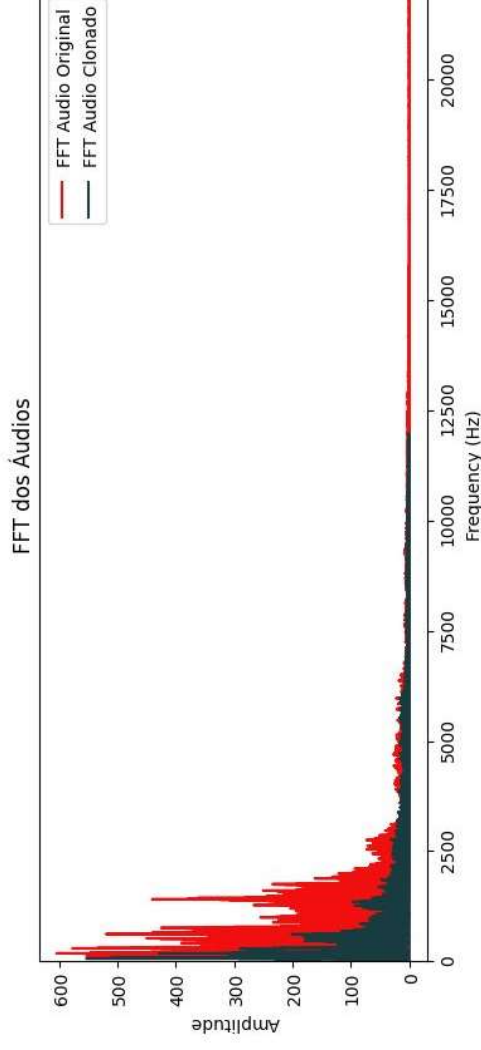


Por fim, uma comparação entre os áudios clonados.

Destaca-se que o processo de clonagem resulta em uma leve redução da frequência; no áudio original, a frequência máxima atinge aproximadamente 150.000 Hz, ao passo que o clone apresenta um limite de até 120.000 Hz.

# Transformada de Fourier

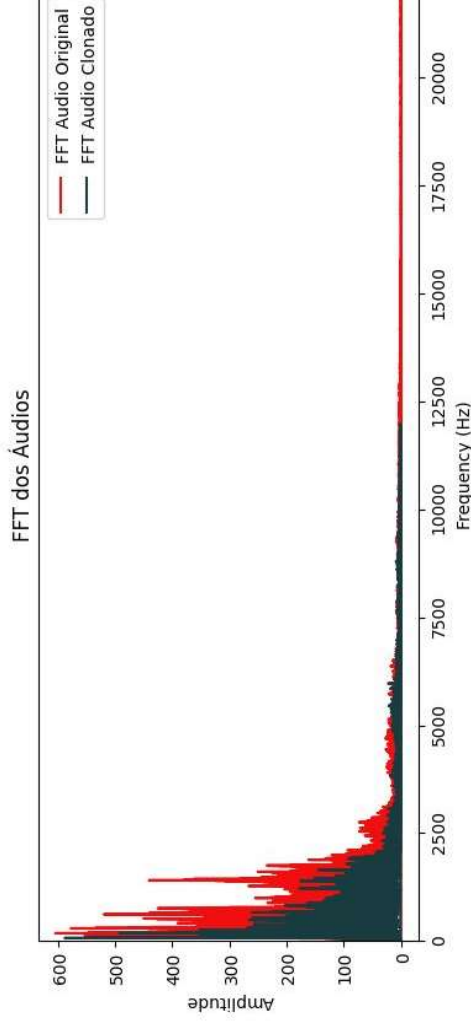
Outra abordagem é a representação do espectro de frequência, mediante o emprego da Transformada de Fourier. Assim como no espectrograma, o interesse reside na compreensão da similaridade entre a voz original e a clonada, contendo conteúdo textual similar ou distinto do áudio original.



Comparação áudio original (vermelho) e áudio clonado com conteúdo diverso (azul).

# Transformada de Fourier

Observamos uma aproximação e similaridade de padrões do áudio original com o clonado, porém, com magnitudes um pouco menores. Esse fenômeno sugere uma capacidade considerável do modelo em replicar as componentes fundamentais e harmônicas principais do sinal de voz original, sugerindo assim eficiência em capturar e recriar as características acústicas mais proeminentes e perceptíveis da voz.

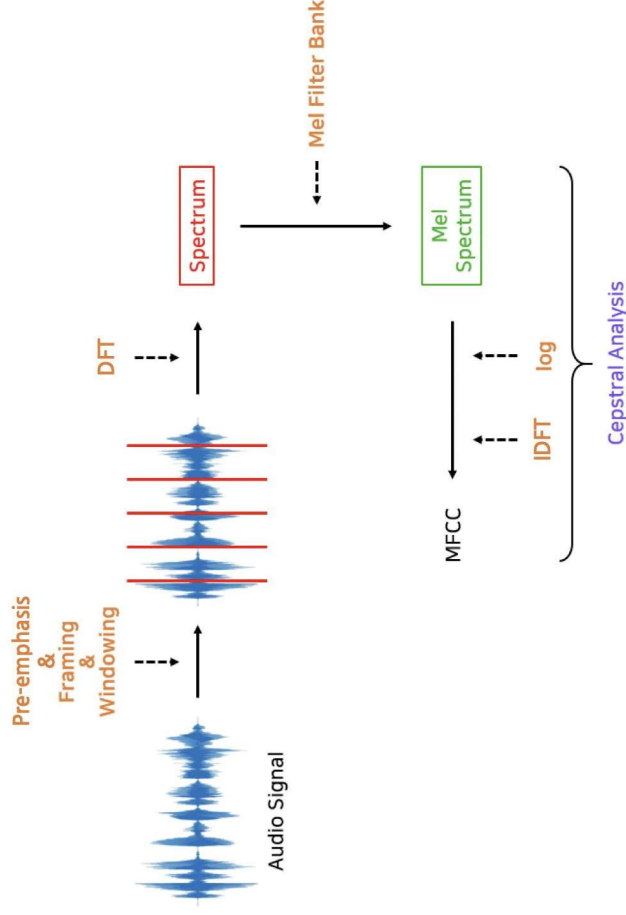


Comparação áudio original (vermelho) e áudio clonado com o mesmo conteúdo do original (azul).



# Coeficientes Cepstrais de Frequência Mel (MFCC)

- **Pré-ênfase e Janelamento:** Amplificação de altas frequências e divisão do sinal em frames.
- **Transformada de Fourier:** Conversão de cada frame para o domínio da frequência.
- **Filtro de Mel:** Aplicação de filtros que simulam a percepção auditiva humana.
- **Logaritmo da Energia:** Captura a percepção de intensidade sonora.
- **Transformada de Cosseno Discreta (DCT):** Decorrelação dos coeficientes e redução da dimensionalidade.







# Áudios clonados

# Forms para Participar do Desafio



acessa aqui:





Letícia



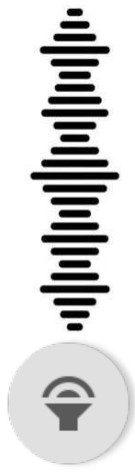
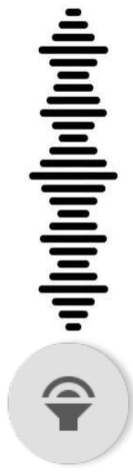
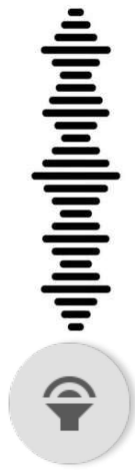
Edward



”

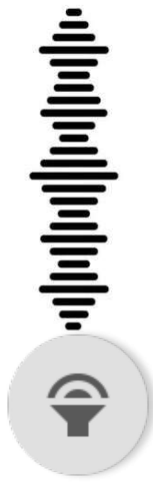
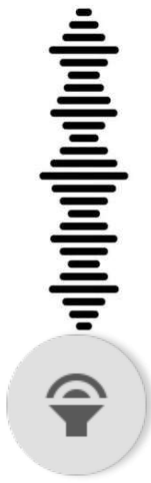


Desafios



”

—  
Áudio original - Miley Cyrus em entrevista



”



Áudios clonados



# | Conclusão

# Conclusão

- Análises demonstram que o modelo TorToise consegue imitar eficazmente as características fundamentais da voz original, apesar de variações nas magnitudes das frequências.
- A precisão na clonagem de voz é confirmada pelo valor MFCC de 0.105, refletindo a eficácia do sistema TorToise em capturar as nuances da voz original.
- O sucesso do TorToise na síntese de fala evidencia o avanço das técnicas de IA na clonagem de voz, apesar de limitações pontuais.



# Obrigado!

