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

André Cerqueira Castro
Universidade Federal de Goiás
Instituto de Informática (INF)
Goiânia, Brasil
andré.castro@discente.ufg.br

Dayane Rodrigues
Universidade Federal de Goiás
Instituto de Informática (INF)
Goiânia, Brasil
dayane.rodrigues@discente.ufg.br

Michael Vinícius Gomes da Silva
Universidade Federal de Goiás
Instituto de Informática (INF)
Goiânia, Brasil
michael.vinicius@discente.ufg.br

Lisandra Cristina de Moura Menezes
Universidade Federal de Goiás
Instituto de Informática (INF)
Goiânia, Brasil
lisandramoura@discente.ufg.br

***Resumo* — O Este relatório detalha o artigo "Better speech synthesis through scaling" de James Betker, que propõe uma metodologia inovadora para a síntese de fala expressiva e multi-voz através do sistema TorToise. Integrando transformadores autoregressivos e *Denoising Diffusion Probabilistic Models* (DDPMs). A reanálise do método TorToise sob a ótica do Processamento Digital de Sinais e Imagens visa explorar seu impacto e possibilidades para a clonagem de voz. Os resultados da implementação da metodologia demonstraram clonagens de voz com qualidade significativa.**

Palavras chave — Speech Synthesis, Voice Cloning, Autoregressive Transformers, Diffusion Models, Natural Language Processing.

I. INTRODUÇÃO

A reprodução precisa da voz humana representa um dos principais desafios no campo da síntese de fala. Betker [1] traz uma contribuição significativa, aplicando técnicas oriundas da geração de imagens para produzir fala sintética de alta qualidade. O sistema TorToise, fundamentado em transformadores autoregressivos e *Denoising Diffusion Probabilistic Models* (DDPMs), destaca-se pela sua expressividade e naturalidade, indicando novos caminhos para a pesquisa em clonagem de voz. Este relatório foca na análise da metodologia de Betker, discutindo sua aplicabilidade e resultados no contexto de Processamento Digital de Sinais e Imagens.

II. FUNDAMENTOS TEÓRICOS

Os transformadores autoregressivos, introduzidos por Vaswani et al. [2], representam um avanço no processamento de sequências, permitindo modelagens complexas de dependências de longo alcance em dados sequenciais. Eles são particularmente eficazes em tarefas de Processamento de Linguagem Natural (NLP) e têm sido adaptados para a síntese de fala, oferecendo uma estrutura poderosa para modelar a complexidade e variabilidade da fala humana.

DDPMs, conforme descrito por Ho et al. [3], são modelos generativos que aprendem a distribuição de dados através de um processo iterativo de difusão e *denoising*. Na síntese de fala, esses modelos facilitam a geração de ondas sonoras realistas, partindo de ruído inicial e refinando progressivamente até alcançar uma saída de alta fidelidade.

III. METODOLOGIA

Utilizando grandes volumes de dados de fala, o modelo aprende a gerar representações de espectrograma MEL a partir de texto. Esta etapa capitaliza na capacidade dos transformadores de modelar complexidades linguísticas e acústicas. Inicialmente, um decodificador autoregressivo, condicionado por entradas de texto, é responsável por prever tokens de fala que representam o conteúdo desejado. Esse processo se beneficia diretamente da habilidade dos modelos autoregressivos de capturar a sequencialidade e nuances entre texto e fala, transformando efetivamente texto em uma sequência preliminar de representações de fala.

Após a geração inicial dos espectrogramas, DDPMs são empregados para refinar essas representações, aumentando sua qualidade e realismo. Este processo iterativo de difusão e *denoising* é essencial para alcançar uma síntese de fala de alta fidelidade. Os DDPMs, operando em uma abordagem de modelagem no domínio contínuo, aprimoram as características dos espectrogramas ao simular e reverter processos de adição de ruído, detalhando e enriquecendo as nuances acústicas para uma representação mais precisa da fala.

A última etapa envolve a utilização de inversores MEL neurais para transformar os espectrogramas refinados em sinais de áudio. Essa conversão, realizada por vocoders avançados, é fundamental para produzir a saída final de fala sintética, com qualidade quase indistinguível da fala humana. Este estágio final assegura que os detalhes acústicos e características tonais capturados nos espectrogramas sejam fielmente convertidos em ondas sonoras, permitindo a geração de fala sintética que replica com precisão as entonações, ritmos e texturas da fala natural.

Adicionalmente, o processo é enriquecido pela introdução de uma Entrada de Condicionamento de Fala, que ajusta a geração para corresponder às características específicas de voz do falante-alvo, e pelo uso de um modelo contrastivo para avaliar e selecionar as melhores correspondências de fala geradas, garantindo assim uma saída de alta qualidade que é contextualmente alinhada com o texto de entrada. Essa abordagem integrada, combinando decodificação

autoregressiva, refinamento por DDPMs, e conversão por inversores MEL neurais, estabelece um novo padrão em síntese de fala, alcançando resultados que são notavelmente naturais e expressivos.

IV. RESULTADOS E CONCLUSÕES

Para avaliar os resultados da implementação do TorToise no contexto da matéria de Processamento Digital de Sinais e Imagens, consideram-se três análises que revelam características essenciais do estudo de sinais: espectrogramas, transformada de fourier e Mel-frequency Cepstral Coefficients (MFCCs). Foram utilizados três áudios para teste: 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.

O espectrograma pode ser utilizado como uma representação visual do espectro de frequência do sinal ao longo do tempo. Pode-se analisar, a partir dessa representação, características como entonação, timbre e qualidade sonora.

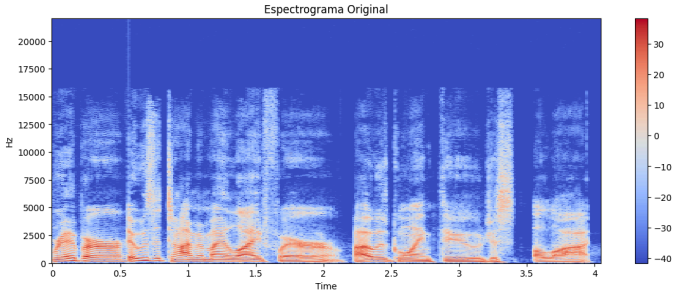

Fig 1: Espectrograma do áudio original.

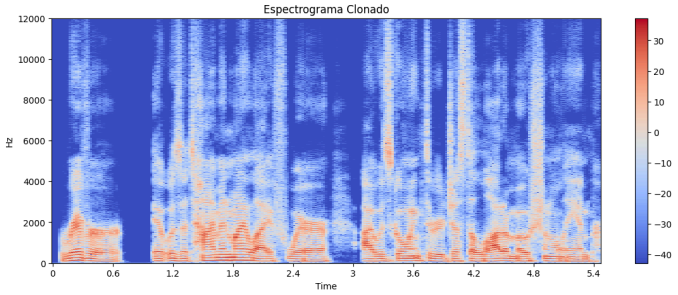

Fig 2: Espectrograma do áudio clonado com o mesmo conteúdo do áudio original.

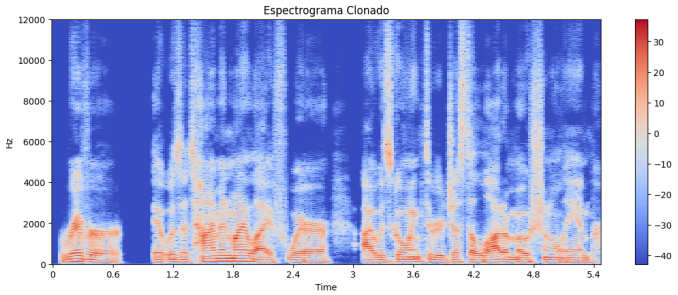

Fig 3: Espectrograma do áudio clonado com conteúdo diverso.

Para a avaliação de espectrogramas, não se torna imprescindível a comparação direta entre o áudio original (utilizado como entrada para a clonagem) e um áudio clonado que reproduza o mesmo conteúdo. No entanto, considera-se pertinente realizar tal comparação, além do áudio clonado com conteúdo díspar.

Observa-se visualmente 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. Ademais, 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.

Outra abordagem, similar à anterior, mas sem relação temporal, que se mostra viável para a análise mais aprofundada da composição da frequência de áudio é a representação do espectro de frequência, mediante o emprego da Transformada de Fourier. Da mesma forma que na análise precedente, o interesse reside na compreensão da similaridade entre a voz original e a clonada, contendo conteúdo textual similar ou distinto do áudio original.

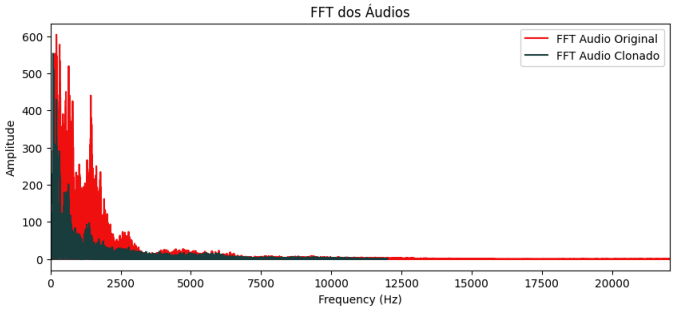

Fig 4: Comparação áudio original (vermelho) e áudio clonado com conteúdo diverso (azul).

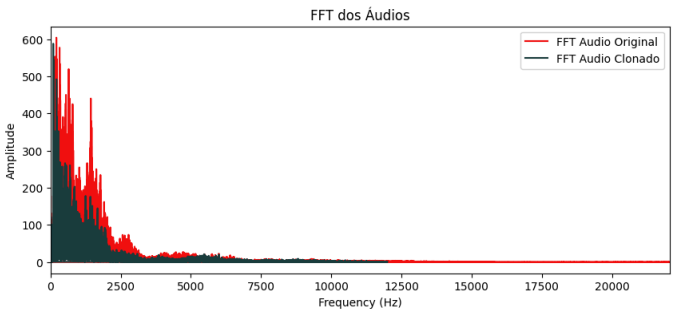

Fig 5: Comparação áudio original (vermelho) e áudio clonado com o mesmo conteúdo do original (azul).

As análises acima em relação à magnitude da transformada de Fourier revela que determinadas frequências apresentam uma aproximação e similaridade de padrões ao áudio original, porém ainda assim com uma diferença de magnitude entre áudio original, com magnitudes maiores, e o clonado, com 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.

Já ao observar apenas a Fig 4, que aborda o áudio clonado com conteúdo textual diferente, as frequências, apesar de possuírem padrões ainda bastante similares, demonstram diferenças mais acentuadas entre suas magnitudes. Isso indica que, embora o modelo demonstre competência em mimetizar os aspectos mais dominantes da voz, as frequências ainda têm magnitudes com diferenças.

Para complementar a exploração do modelo e seu desempenho em clonagem de voz, foram utilizados os Coeficientes Cepstrais de Frequência Mel (MFCCs), os quais foram empregados para comparar áudios original e clonado através da extração de características acústicas fundamentais de ambos os sinais. A distância entre os conjuntos de coeficientes é calculada utilizando métricas como a distância Euclidiana, que mede a raiz quadrada da soma dos quadrados das diferenças entre os elementos correspondentes dos vetores MFCC. Essa comparação quantifica a similaridade entre os áudios, com valores menores indicando maior semelhança acústica e, portanto, uma reprodução mais precisa das características perceptivas da fala, como timbre, textura, entonação, intensidade e ritmo de fala, entre outros.

Utilizados os mesmos áudios das análises de frequência, obteve-se o valor de MFCC de 0.105. Isso reflete uma distância muito pequena entre o áudio original e o clonado, indicando considerável fidelidade na reprodução das propriedades acústicas do original. A métrica nos indica eficácia do processo de clonagem de voz em capturar e replicar as nuances acústicas chave, resultando em uma voz sintetizada que captura bem as características da voz testada.

Em síntese, o sistema TorToise representa um marco na síntese de fala, evidenciando o potencial das técnicas de transformadores autoregressivos e DDPMs na clonagem de voz. O sucesso do TorToise reafirma o valor de abordagens generalistas e o poder de grandes conjuntos de dados e computação avançada no avanço da síntese de fala. Os resultados presentes neste relatório revelam que o método já permite clonagens de qualidade satisfatória, com limitações pontuais.

REFERÊNCIAS

[1] J. Betker, “Better speech synthesis through scaling”, Journal of Machine Learning Research, 2023.

[2] A. Vaswani, et al., “Attention is All You Need”, Advances in Neural Information Processing Systems, 2017.

[3] J. Ho, et al., “Denoising Diffusion Probabilistic Models”, Advances in Neural Information Processing Systems, 2020.

[4] A. van den Oord, et al., “WaveNet: A Generative Model for Raw Audio”, arXiv, 16 de setembro de 2016. Disponível em: https://arxiv.org/abs/1609.03499.

[5] C. Wei, et al., “Data-Efficient Text-to-Speech Synthesis with Contrastive Predictive Coding”, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2022.