CLONAGEM DA VOZ HUMANA POR SÍNTESE DE VOZ COM O USO DE INTELIGÊNCIA ARTIFICIAL

Alessander Alves Novaes
Pós-Graduação em Engenharia de Controle e Automação
Instituto Federal do Espírito Santo
Vitória-ES, Brasil
alessander@outlook.com

Luiz Alberto Pinto
Pós-Graduação em Engenharia de Controle e Automação
Instituto Federal do Espírito Santo
Vitória-ES, Brasil
luiz.pt@ifes.edu.br

***Resumo*—Esta pesquisa aborda a síntese de fala para reprodução de vozes em alta qualidade e tempo real no idioma português brasileiro. Nosso objetivo é atender à demanda do mercado publicitário por locuções, narrações e dublagens utilizando vozes geradas por modelos de Inteligência Artificial (IA). O modelo é treinado em dados de pares texto-fala e usa um mecanismo de atenção para alinhar o texto à fala gerada. Além disso, propomos uma estimativa automática dos parâmetros de um sintetizador por formantes, usando algoritmo genético (AG), para imitar vozes. Com base nos resultados, alcançamos uma alta similaridade entre as vozes sintéticas e originais, validada por uma medida de similaridade de coseno de 0,90473765. Isso evidencia a qualidade da síntese realizada e reforça nosso foco em produzir vozes sintéticas semelhantes às vozes-alvo.**

***Índice de Termos*—synthetic speech, Tacotron2 model, neural networks, genetic algorithm, speech synthesis**

***Abstract*—This research addresses the synthesis of speech for high-quality real-time voice reproduction in Brazilian Portuguese. Our aim is to cater to the advertising market's demand for voiceovers, narrations, and dubbing using voices generated by Artificial Intelligence (AI) models. The model is trained on text-speech pairs and employs an attention mechanism to align text with generated speech. Furthermore, we propose an automatic estimation of parameters for a formant-based synthesizer using a genetic algorithm (GA) to mimic voices. Based on our achieved results, we have successfully synthesized a male voice in Brazilian Portuguese, with a synthesized voice demonstrating a high degree of similarity to the original voice. A cosine similarity measure of 0.90473765 validates this, highlighting the quality of our synthesis approach and underscoring our commitment to producing synthetic voices that closely resemble target voices.**

***Índice de Termos*—synthetic speech, Tacotron2 model, neural networks, genetic algorithm, speech synthesis**

I. Introdução

A geração de fala é uma tarefa crucial em várias aplicações, incluindo assistentes virtuais e dublagens de filmes e séries. Com o avanço das técnicas de aprendizado de máquina, a síntese de fala de alta qualidade tornou-se possível com o modelo de síntese de fala baseado em redes neurais.

O objetivo principal deste trabalho é imitar uma voz alvo, tanto natural quanto sintética, através da estimativa dos valores dos parâmetros que compõem o arquivo de entrada de um sintetizador por formantes, utilizando Algoritmo Genético (AG). A partir de um arquivo de voz dado como entrada em um sistema Speech to Speech (STS), um modelo inverso gera a combinação de parâmetros de entrada, que é submetida ao sintetizador de voz por formantes, que por sua vez, produz uma voz sintética que imita a voz alvo (entrada).

A geração de fala é uma tarefa crucial em várias aplicações, incluindo assistentes virtuais e dublagens de filmes e séries [1] e [2]. A síntese de fala de alta qualidade, especialmente usando modelos baseados em redes neurais, tem permitido a criação de vozes artificiais cada vez mais realistas [3] [4].

Para a síntese da fala, nesse trabalho, foi utilizado o Tacotron 2 [5], um modelo de síntese de fala que utiliza o Espectrograma Mel [6] como entrada e gera fala de alta qualidade em tempo real. O modelo usa um mecanismo de atenção para alinhar o texto com a fala gerada e é capaz de gerar fala expressiva e natural em diferentes idiomas e estilos de voz. A Figura 1 ilustra esse objetivo.

Apesar de já existir um conjunto de pesquisas e trabalhos profissionais voltados a conversão de texto-fala, observou-se que existe pouca pesquisa utilizando o idioma português brasileiro, desta forma esta pesquisa é focada no desenvolvimento de sistema de conversão texto-fala do estado da arte, para o português brasileiro com um conjunto de dados fornecidos pela Rede Globo [7]. Isso permitirá uma ampla gama de aplicações em que a fala em português brasileiro é necessária.

O restante desse trabalho está organizado nas seguintes seções: Na Seção II, são apresentados os principais trabalho correlatos utilizados como base para o desenvolvimento da pesquisa. A Seção III destaca os principais conceitos aplicados na execução. A Seção IV descreve a metodologia que foi utilizado, apresentando cada uma de suas etapas constituíntes. Na Seção V, os resultados obtidos são apresentados e analisados de forma qualitativa. Por fim, na Seção VI são apresentadas as conclusões.

II. Trabalhos Correlatos

A síntese de fala tem desempenhado um papel essencial na interação humano-computador, permitindo que os sistemas gerem fala natural a partir de texto escrito. No entanto,

Figura 1. Sistema para imitação de voz.

para o idioma português brasileiro, a pesquisa nessa área é limitada, havendo uma lacuna no desenvolvimento de técnicas específicas de *Text to Speech* (TTS). Nesta artigo, revisamos o estado da arte em TTS para o idioma português brasileiro, focando nas abordagens utilizando Tacotron 2 [5] e WaveNet [8]. Para isso, analisamos três trabalhos de pesquisa relevantes [9], [10] e [11]. Além disso, realizamos uma comparação entre uma gravação de voz natural e uma reprodução feita por um sistema de Inteligência Artificial (I.A), explorando pesquisas não acadêmicas que utilizam o Tacotron 2 e o WaveNet para o idioma português brasileiro.

A síntese de fala em português brasileiro tem desafios específicos devido às características fonéticas e prosódicas únicas do idioma. Embora abordagens de TTS baseadas em aprendizado de máquina, como o Tacotron 2 e o WaveNet, tenham alcançado resultados promissores em outros idiomas, ainda não existem trabalhos direcionados especificamente para o português brasileiro. Portanto, é fundamental revisar o estado da arte em TTS para esse idioma e identificar oportunidades para futuras pesquisas. Além disso, este trabalho visa realizar uma comparação entre uma gravação de voz natural e uma reprodução feita por um sistema de I.A [12], com foco no idioma português brasileiro.

O trabalho de Ana Catarina Rosa Gonçalves [9] aborda a síntese de fala em português europeu utilizando técnicas de aprendizado profundo. Embora o foco seja no português europeu, seu trabalho fornece insights valiosos sobre a utilização do Tacotron 2 e do WaveNet para o processamento da língua portuguesa. O estudo destaca a importância de considerar as peculiaridades linguísticas e fonéticas do português ao projetar sistemas de TTS eficazes. No entanto, ele não aborda especificamente as características únicas do português brasileiro.

O autor em [10] aborda a detecção de fala sintética usando redes neurais profundas. Embora não seja diretamente relacionado à síntese de fala, o estudo destaca a relevância de técnicas de aprendizado de máquina para o processamento de fala artificial. Isso pode ser aplicado ao contexto do TTS, onde a detecção de fala sintética é essencial para garantir a qualidade e autenticidade da fala gerada.

O trabalho de Sisamaki Eirini [11] é focado na síntese de fala em grego e fornece insights valiosos sobre abordagens de TTS baseadas em redes neurais de ponta a ponta. A autora explora técnicas como o Tacotron 2 e destaca a importância de ajustar o modelo para se adequar às características específicas do idioma. No entanto, essa abordagem ainda não foi adaptada para o português brasileiro.

Além das pesquisas acadêmicas mencionadas acima, foram encontradas pesquisas não acadêmicas em outros repositórios, como GitHub [13], Kaggle [14] e outros [15], que utilizam o modelo Tacotron 2 e integração com o vocoder neural WaveNet. Essas pesquisas exploram a combinação de valores dos parâmetros que levam a uma voz sintética suficientemente parecida com uma voz alvo, levando em consideração a combinação fonética do povo brasileiro e seu regionalismo (prosódia). Embora não sejam estudos acadêmicos, eles contribuem para o conhecimento prático e aplicado no desenvolvimento de sistemas de TTS para o idioma português brasileiro.

Este trabalho avança em relação aos trabalhos correlatos de diversas maneiras. Em primeiro lugar, abordamos a síntese de fala em português brasileiro, preenchendo a lacuna de pesquisa específica para esse idioma. Nossa abordagem visa a otimização dos parâmetros de síntese, buscando vozes sintéticas de alta fidelidade e naturalidade. Enquanto os trabalhos anteriores se concentram em idiomas diferentes ou não abordam diretamente as características do português brasileiro, nosso trabalho se destaca por considerar as peculiaridades fonéticas e prosódicas únicas desse idioma.

III. Referencial Teórico

Nesta Seção são apresentados os principais conceitos aplicados no desenvolvimento do trabalho. Inicialmente é feita uma breve discussão sobre processamento de sinais de fala. Em seguida é apresentada a teoria de processamento de sinais que dá suporte a extração das frequências da escala mel dos sinais de fala. Uma breve descrição da síntese de fala para fala, conceito sobre o qual estão fundamentados os sistemas de clonagem de voz. Por último o Tacotron 2, modelos de síntese de fala utilizado neste trabalho, é descrito em todas as suas etapas.

A. Processamento do Sinal de Fala

Os sinais de voz são, essencialmente, não-estacionários, portanto, seus parâmetros estatísticos e sua forma de onda se alteram ao longo do tempo. Estas alterações são provocadas pelas modificações dos articuladores envolvidos no processo de geração do sinal de fala, [16].

As ferramentas utilizadas no processamento desses sinais requerem que os mesmos permaneçam invariantes no tempo.

Na produção da voz estão envolvidos diferentes órgãos, ossos e músculos e, devido à inércia destes articuladores, não é possível alterar as suas posições de forma abrupta, nem instantaneamente. Modificar o posicionamento dos diversos articuladores e consequentemente alterar a forma do trato vocal é, portanto, um processo contínuo e suave. Dessa forma, se um sinal de voz for dividido em segmentos de duração suficientemente curta (aproximadamente 20ms), estes segmentos de curta duração poderão ser considerados quase estacionários, pois, durante a sua duração, os articuladores movem-se pouco e lentamente para que as características acústicas do segmento possam ser consideradas invariantes no tempo [16].

Vários métodos foram propostos para a extração de descritores de sinais de voz, como por exemplo: o *Linear Predictive Coding* (LPC) [17], o *Perceptual Linear Prediction* (PLP) [18] e o *Mel-Frequency Cepstral Coefficients - MFCC* [19].

Devido a sua natureza linear, onde se assume um mesmo peso para todo o espectro da fala, o *LPC* torna-se uma opção menos atraente. O *PLP* e *MFCC* são baseados no conceito dos *filter banks* logaritmicamente espaçados, que se aproximam da forma da escuta humana, mostrando-se boas opções para a tarefa de reconhecimento automático de fala [20]. As *MFCCs* apresentam algumas vantagens sobre a abordagem *PLP*: reduz o volume de informação das frequência da fala em um pequeno número de *features*, aproxima a percepção de *loudness* ao do sistema de escuta humano e é um modelo de processamento de áudio bem simples, tendo baixo custo computacional para sua obtenção [21].

B. As Frequências da Escala Mel

Um sinal de voz é o resultado da convolução entre a sequência de excitação e a resposta ao impulso do sistema vocal. Existem ocasiões em que é conveniente separar as duas componentes para que seja possível manipular apenas uma das partes, mas este processo não é trivial. A análise cepstral [22], [23] foi desenvolvida para tornar mais simples a solução desse problema. De forma ideal, a análise cepstral representa uma transformação do sinal de voz de forma a evidenciar duas propriedades importantes:

1) As representações das componentes do sinal estarão separadas no *cepstrum*;
2) As representações das componentes de sinal no *cepstrum* vão corresponder a uma combinação linear.

Após o sinal estar representado no *cepstrum*, aplica-se um filtro linear para remover os trechos indesejados e selecionar algumas componentes específicas. As componentes que não foram eliminadas aplica-se uma transformação inversa. Este procedimento obedece o princípio da sobreposição, que no caso da convolução pode ser representado pela Equação 1,

$$H[x(n)] = H[x_1(n) * x_2(n)] = H[x_1(n)] * H[x_2(n)] \quad (1)$$

sendo "$H[.]$" um sistema homomórfico e o símbolo $(*)$ representa a operação de convolução.

Sistemas Homomórficos são aqueles que obedecem ao princípio da sobreposição para a convolução. O operador de *cepstrum* complexo $D_*[.]$, cuja representação está mostrada na Figura 2 desempenha um papel importante na teoria de sistemas homomórficos, que é baseado em uma generalização do princípio da superposição.

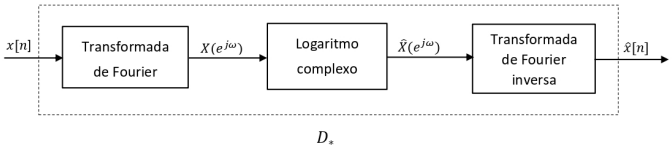

Figura 2. Representação do cálculo do cepstrum complexo, [24]

Na filtragem homomórfica de sinais convoluídos, o operador $D_*[.]$ é denominado sistema característico para convolução, pois tem a propriedade especial de transformar a convolução em adição [24].

Consideremos que,

$$x[n] = x_1[n] * x_2[n] \quad (2)$$

de modo que a transformada z correspondente é

$$X(z) = X_1(z).X_2(z) \quad (3)$$

Se o logaritmo complexo for calculado de acordo com a definição do *cepstrum* complexo, então,

$$\widehat{X}(z) = log[X(z)] = log[X_1(z)] + log[X_2(z)] = \widehat{X_1}(z) + \widehat{X_2}(z) \quad (4)$$

o que implica que o *cepstrum* complexo é

$$\widehat{x}(n) = D_*[x_1[n] * x_2[n]] = \widehat{x_1}(n) + \widehat{x_2}(n) \quad (5)$$

De acordo com [24], uma análise similar mostra que, se

$$\widehat{y}[n] = y_1[n] + y_2[n] \quad (6)$$

então segue que

$$D_*^{-1}[\widehat{y}_1[n] + \widehat{y}_2[n]] = \widehat{y}_1[n] * \widehat{y}_2[n] \quad (7)$$

Se os componentes cepstrais $\widehat{x}_1[n]$ e $\widehat{x}_2[n]$ ocuparem diferentes faixas de frequência, a filtragem linear pode ser aplicada ao *cepstrum* complexo para remover ou $\widehat{x}_1[n]$ ou $\widehat{x}_2[n]$. Ainda de acordo com [24], se esta etapa for seguida da transformação por meio do sistema inverso $D_*^{-1}[.]$, cuja constituição esta ilustrada na Figura 3, o componente correspondente será removido na saída.

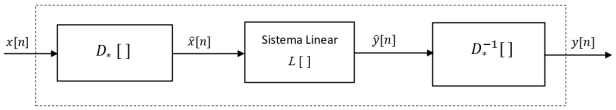

Figura 3. Sistema homomórfico em que entrada e saída são combinadas por convolução, [24]

O *MEL-Cepstrum* é uma variação do *cepstrum* normal que melhor se ajusta à percepção auditiva humana. A verdadeira

frequência de um som e a percepção que um humano tem dessa frequência não têm uma correspondência linear. A frequência percebida pelo ouvido humano, também conhecida como *pitch*, tem como unidade de medição o *MEL*. A Figura 4 representa a escala *MEL*, criada por Stevens e Volkman em 1940, e relaciona o *pitch* com a frequência real.

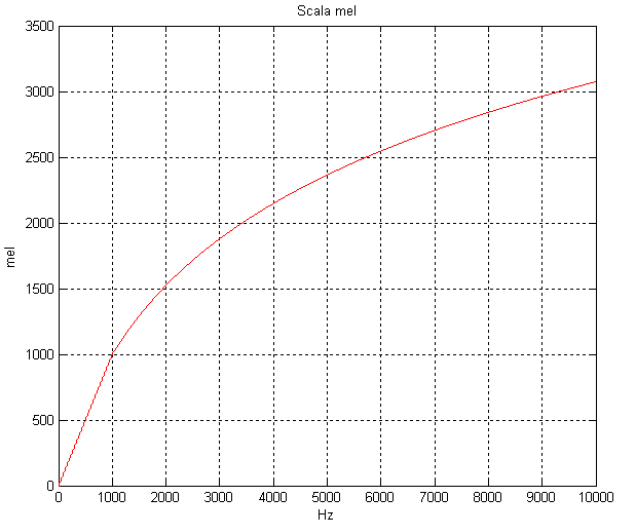

Figura 4. A Escala das Frequências Mel.

Como pode ser observado, a relação entre as duas é praticamente linear até os 1.000 Hz, se tornando logarítmica para frequências superiores a esse valor. Um fato relevante a ser considerado é que a percepção que se tem de uma determinada frequência é influenciada pela energia das bandas de frequências crítica em torno dessa, e que a largura das bandas bandas críticas varia com a frequência.

C. Speech-To-Speech synthesis

Com a evolução dos sistemas TTS baseados em aprendizado de máquina, a comunidade acadêmica começou a pesquisar formas de criar um modelo de síntese de fala para fala. A ideia é ter como input um enunciado falado pela pessoa A e, tendo como output um enunciado com o mesmo conteúdo (palavras), mas na voz da pessoa B. Embora existam usos legítimos para essa tecnologia, como a criação de vozes para aplicativos comerciais, ela pode ser usada para fins maliciosos, ou seja, personificar alguém.

No início de 2018, pesquisadores do Baidu Labs começaram a investigar formas de realizar a clonagem de voz [25]. Além disso, o objetivo deles era gerar uma voz do locutor usando apenas algumas amostras. A ideia geral é treinar um modelo de codificador-decodificador capaz de ouvir a voz de alguém (decodificar) e reproduzir as mesmas palavras, mas na voz de outra pessoa (codificação). Vários modelos de alto-falantes já foram explorados no passado (como visto no Deep Voice e outros projetos). No entanto, a chave deste projeto é um sistema de áudio para áudio capaz de clonar uma voz com apenas alguns exemplos (em contraste com trabalhos anteriores que exigiam mais de 2 horas de gravações para treinar um modelo).

Os pesquisadores usam uma arquitetura de codificador-decodificador semelhante à utilizada no projeto Deep Voice 3: uma rede neural profunda de convolução baseada em atenção. No entanto, em vez de texto-para-fala (como no projeto Deep Voice 3), eles usam um modelo de fala para fala, que tem áudio como entrada e saída. Os pesquisadores propõem duas abordagens para a clonagem de vozes com poucas amostras: adaptação do locutor e codificação do locutor. O primeiro é basicamente o ajuste fino de um modelo treinado de vários alto-falantes. Esse ajuste fino pode ser feito modificando a incorporação do alto-falante ou retreinando o modelo generativo com as poucas amostras fornecidas. A outra abordagem, codificação de alto-falante, consiste em treinar novamente o modelo de codificação do zero para inferir diretamente uma incorporação de alto-falante do áudio de clonagem, que é usado no modelo generativo.

Os autores comparam os resultados das duas abordagens e mostram que a adaptação do falante produz melhor similaridade (entre o áudio original e o áudio gerado) e melhor naturalidade da fala. No entanto, a incorporação do alto-falante requer menos tempo e usa menos memória durante o processo de inferência, o que significa que pode ser mais adequada em casos onde recursos como CPU e memória são um gargalo. Os autores também publicaram 15 áudios gerados pelo método de clonagem proposto, e os resultados mostram uma voz clonada muito semelhante à voz original.

D. Tacotron2

O Tacotron2 é um modelo de síntese de fala baseado em redes neurais profundas, que utiliza o espectrograma de mel como entrada e gera fala de alta qualidade em tempo real. O modelo é treinado em um conjunto de dados de pares de texto-fala e usa um mecanismo de atenção para alinhar o texto com a fala gerada. Conforme mostrado na Figura 5, o Tacotron 2 consiste em dois componentes principais. No primeiro componente, os espectrogramas de mel são obtidos a partir da sequência de entrada e fornecidos à rede de previsão de características de sequência para sequência. O segundo componente do sistema TTS inclui o WaveNet, que é responsável por gerar amostras de formas de onda no domínio do tempo. Depois que a rede de predição recorrente de sequência a sequência gera espectrogramas a partir do texto de entrada, esses espectrogramas são transformados em amostras de forma de onda no domínio do tempo usando o WaveNet.

O modelo Tacotron2 é composto por duas partes principais: um codificador e um decodificador. O codificador recebe o texto como entrada e produz um vetor de contexto que representa o conteúdo do texto. O decodificador usa esse vetor de contexto para gerar o espectrograma de mel, que é então sintetizado em fala usando um vocoder.

O uso de uma representação calculada a partir de formas de onda no domínio do tempo na arquitetura do Tacotron 2 requer o treinamento dos dois componentes separadamente. O espectrograma de frequência mel está relacionado ao espectrograma

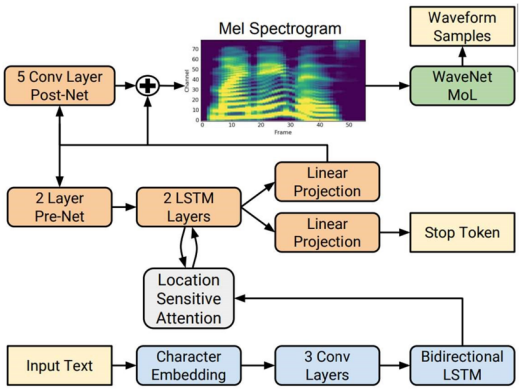

Figura 5. Modelo Tacotron 2 para geração de espectrogramas de mel a partir de texto.

de frequência linear, ou seja, a Short-time Fourier Transform (STFT). É obtido pela aplicação de uma transformação não linear ao eixo de frequência do STFT, inspirada nas respostas medidas do sistema auditivo humano. O uso de uma escala de frequência auditiva dessa maneira destaca detalhes em frequências mais baixas que são essenciais para a inteligibilidade da fala. No entanto, enquanto espectrogramas lineares descartam informações de fase, algoritmos como Griffin–Lim [26] pode prever essas informações descartadas. A estrutura Griffin-Lim usada no Tacotron 1 possibilita a transformação no domínio do tempo via STFT inversa. Os espectrogramas de mel usados no Tacotron 2 descartam mais informações, apresentando um problema desafiador de transformação inversa. No entanto, quando comparado aos recursos linguísticos e acústicos usados no WaveNet, o mel-espectrograma é uma representação acústica mais simples e de nível inferior dos sinais de áudio. Portanto, é possível produzir fala de alta qualidade a partir de espectrogramas de mel usando uma estrutura WaveNet.

O WaveNet, que é a base da arquitetura do Tacotron 2, representa uma rede neural convolucional auto-regressiva que prevê amostras de fala a partir de recursos linguísticos. Espectrogramas são usados em vez de recursos linguísticos como entrada na estrutura WaveNet. A principal desvantagem dos codificadores de som preparados com a estrutura especificada é que eles podem produzir apenas um exemplo de fala em uma passagem direta. WaveNet é um modelo autorregressivo que usa amostras anteriores para gerar cada nova amostra. Como resultado, o WaveNet precisa processar as amostras anteriores uma a uma para gerar a forma de onda, o que pode aumentar o tempo de processamento.

O Tacotron 2 foi usado para estimar espectrogramas de mel após realizar o pré-processamento em textos turcos. A predição de Espectrogramas Mel foi baseada na arquitetura Tacotron 2 na síntese natural de TTS no estudo preparado por Shen e seus colegas [27]. Espectrogramas Mel fornecem uma representação visual dos componentes de frequência da fala e como eles mudam ao longo do tempo. No entanto, extrair Espectrogramas Mel com precisão pode ser um desafio devido às nuances da fala, como ênfase, ritmo, tom e presença de componentes de alta ou baixa frequência. Espectrogramas Mel derivados incorretamente podem levar a problemas com fala sintetizada, como entonação incorreta, acentos distorcidos ou mistura de vozes. Portanto, a qualidade dos Espectrogramas Mel impacta diretamente na qualidade da fala sintetizada. Ao adquirir Espectrogramas Mel, o número de canais é um fator crucial na obtenção de uma visualização detalhada do conteúdo espectral do sinal de fala. Contagens de canal mais altas oferecem uma resolução de frequência mais alta, mas isso requer mais poder de processamento no processo de síntese e tempos de treinamento mais longos. Portanto, ao determinar o número de canais, os resultados do Tacotron 2 para diferentes idiomas foram considerados, e um número de canal do espectrograma de mel de 80 foi definido usando STFT, como no Tacotron 2 [28], [29] . O tamanho STFT foi convertido para a escala mel usando um banco de filtros mel de 80 canais variando de 125 Hz a 7,6 kHz e, em seguida, registrando a compressão da faixa dinâmica.

O codificador é composto por uma pilha de camadas convolucionais bidirecionais seguidas por uma camada GRU. Cada camada convolucional é seguida por uma camada de normalização de lote e uma função de ativação ReLU. O vetor de contexto é calculado como a concatenação dos estados finais da camada GRU. O decodificador é composto por uma camada GRU e uma camada de atenção. A camada GRU recebe como entrada um vetor de contexto e um vetor de entrada que é inicializado com um vetor de zeros. A camada de atenção é responsável por alinhar o texto com a fala gerada e é implementada usando um mecanismo de atenção baseado em conteúdo.

Para sintetizar fala usando o Tacotron2, é necessário fornecer um texto como entrada para o modelo. O texto é convertido em um vetor de contexto usando o codificador e, em seguida, o decodificador é usado para gerar o espectrograma de mel correspondente à fala gerada.

O espectrograma de mel é então sintetizado em fala usando um vocoder. O vocoder é um modelo de sinal de áudio que pode ser treinado para converter o espectrograma de mel

em forma de onda de áudio. O Tacotron2 utiliza o vocoder WaveNet, que é um modelo de rede neural generativa capaz de sintetizar áudio de alta qualidade.

IV. Metodologia

Nesta seção são descritas todas as etapas da metodologia que orientou a realização do trabalho. É feita uma breve descrição da base de dados utilizada para o desenvolvimento da pesquisa. A etapa de pré-processamento é apresentada. Por fim o desempenho da rede na etapa de treinamento da rede é descrito.

A. Base de Dados

A base de dados utilizada neste estudo consiste em um dataset de vozes brasileiras fornecido pela Rede Globo, respeitando os termos de uso para trabalhos acadêmicos [7]. Para realizar a divisão dos dados em treinamento, testes e validação, adotamos uma abordagem padrão amplamente utilizada na pesquisa de aprendizado de máquina.

O conjunto de dados foi dividido de forma estratificada, garantindo que as proporções de cada classe (no caso, frases de voz) fossem mantidas em cada subconjunto. Especificamente, 70% dos dados foram alocados para o conjunto de treinamento, 20% para o conjunto de testes e 10% para o conjunto de validação, conforme representado na tabela I.

Tabela I
Conjunto de dados.

Finalidade	Frases
Treinamento	3840
Teste	1100
Validação	550
Total	5490

- **Treinamento**: Contém 70% do conjunto de dados, utilizado para treinar os modelos de aprendizado de máquina.
- **Testes**: Contém 20% do conjunto de dados.
- **Validação**: Contém 10% do conjunto de dados.

Essa divisão estratificada ajuda a evitar um desequilíbrio significativo entre os subconjuntos, garantindo que o modelo seja treinado, testado e validado em dados que representem a distribuição real dos exemplos, permitindo que outros pesquisadores possam replicar e validar os resultados obtidos.

Os metadados estão classificados da seguinte forma na tabela II

Tabela II
Metadados.

Total Clips	**5490**
Total Words	178873
Total Duration	19:56:30
Mean Clip Duration	13.09
Max Clip Duration	31.05
Min Clip Duration	4.78
Distinct Words	23407

O conjunto de dados pré-processado e dividido, ele está pronto para ser normalizado de acordo com as especificações.

B. Normalização de Áudio

Os arquivos foram coletados de diversas fontes de dados distintas, e como parte do primeiro passo de pré-processamento, todos eles foram convertidos para um formato de arquivo unificado. Escolheu-se o formato WAV, amplamente empregado em aprendizado de máquina e processamento de áudio digital, para essa conversão.

Adicionalmente, procedeu-se à normalização do volume de todas as falas, eliminando assim a possibilidade de o volume atuar como um fator distintivo. Tanto as falas sintéticas quanto as reais passaram por esse processo a fim de atingir um nível de 0dB.

A amostragem de todos os arquivos de áudio foi fixada em uma taxa de 22kHz. Essa taxa de amostragem, levando em consideração a gama típica da fala humana entre 300Hz e 5000Hz, não acarretará perda significativa na qualidade do áudio. Além disso, arquivos com dois canais foram convertidos para um único canal por meio da técnica de mistura de canais, que consiste em combinar duas faixas de áudio em um único canal mono.

C. Definição dos Hiperparâmetros de Treinamento

Nesta seção, apresentamos os hiperparâmetros de treinamento utilizados no desenvolvimento do modelo:

- **Tamanho do Lote:** O tamanho do lote é definido como 1, o que pode ser ajustado de acordo com a disponibilidade de memória RAM.
- **Épocas de Treinamento:** O modelo é treinado ao longo de 20 épocas, um valor recomendado que permite a convergência dos pesos da rede.
- **Taxa de Aprendizado:** A taxa de aprendizado é inicializada em 3e-4 e decai ao longo do treinamento. O valor de decaimento (B_) é definido como 8000, e o valor de C_ é 0. Isso controla a diminuição gradual da taxa de aprendizado, que ajuda na estabilidade do treinamento e na obtenção de resultados mais precisos.
- **Taxa Mínima de Aprendizado:** A taxa mínima de aprendizado é definida como 1e-5, garantindo que a taxa de aprendizado não diminua para valores muito pequenos.
- **Outros Parâmetros:** Além dos hiperparâmetros de taxa de aprendizado, também definimos outras configurações, como comprimento do filtro, dropout de atenção e dropout do decodificador, que são definidos como 1024, 0.1 e 0.1, respectivamente. Também especificamos o ponto de início do decaimento da taxa de aprendizado (`decay_start`) como 15000.

D. Treinamento da I.A

Na segunda etapa do trabalho, foi iniciado o treinamento da inteligência artificial utilizando os arquivos de espectrograma mel convertidos no passo anterior, em conjunto com as transcrições de áudio para cada arquivo. O processo foi executado em 20 épocas com duração média de 02:30 hs cada época, ultrapassando 50 horas de treinamento (Figura 6).

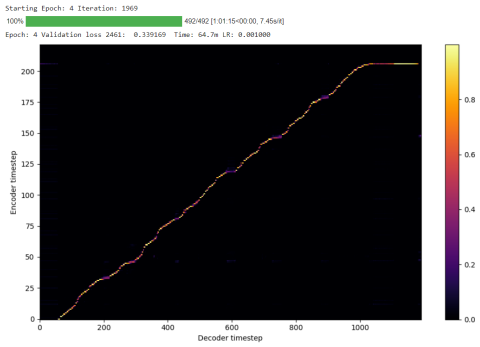

Figura 6. Treinamento da I.A.

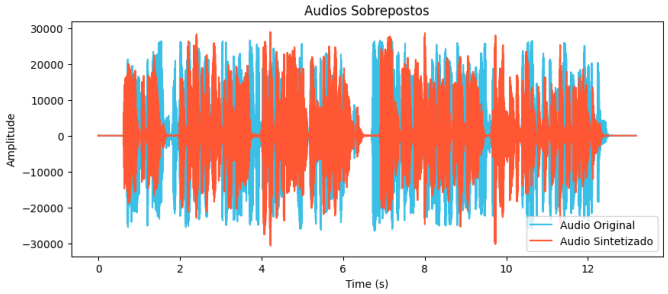

Figura 7. Forma de onda com a sobreposição dos arquivos apresentados.

E. Criação da síntese de voz

Após o treinamento do modelo, foi utilizado o código em *Python* para execução do processo TTS, executando a frase "Ainda segundo o ministro, o presidente está satisfeito com o ritmo da vacinação e que uma possível decisão de liberar a máscara não é para agora. As mãos esfregadas com angústia: gestos comuns entre os jovens que nem chegaram aos trinta anos.". O resultado obtido foi uma síntese de voz clara e natural, que foi comparada com a gravação de voz original. As diferenças entre as duas vozes foram mínimas, o que demonstra a eficácia do modelo de síntese de fala proposto.

V. Resultados

O objetivo principal deste trabalho foi sintetizar uma voz masculina em português brasileiro, utilizando um dataset composto por vozes humanas brasileiras masculinas. O experimento foi bem-sucedido, e os resultados obtidos demonstraram a viabilidade e eficácia do método de síntese utilizado.

Durante o desenvolvimento do trabalho, foi realizada uma análise aprofundada das características fonológicas da fala em português brasileiro, levando em consideração aspectos como entonação, ritmo e pronúncia. Essa pesquisa serviu como base para o treinamento do modelo de síntese de voz.

Um dos principais desafios encontrados durante o projeto foi a configuração do ambiente de treinamento no Google Colab, devido ao tamanho do dataset e à complexidade do treinamento. Foram necessários ajustes cuidadosos nos parâmetros e configurações para otimizar os recursos computacionais e garantir o melhor desempenho. O treinamento foi realizado por mais de 50 horas, e os resultados obtidos foram satisfatórios.

Após o treinamento, foram realizadas avaliações comparando a voz sintetizada com a voz original do dataset. Foram conduzidos testes de percepção de qualidade da voz sintetizada, nos quais foi obtido feedback de um grupo reduzido de ouvintes especializados. Os resultados mostraram um alto grau de similaridade entre a voz sintetizada e a voz original, evidenciando a eficácia do método de síntese utilizado.

A fim de ilustrar os resultados obtidos, foram geradas formas de onda (Figura 7) e espectrogramas (Figura 8) comparando o arquivo de áudio original e o arquivo sintético gerado pela inteligência artificial (IA). Essas representações visuais demonstram a semelhança entre os sinais de fala e corroboram os resultados das avaliações subjetivas de percepção.

Também é possível acessar os arquivos de áudio original e sintético pelos links a seguir:

Ouça o áudio original Ouça o áudio sintetizado

Também foram realizadas comparações dos espectrogramas utilizando a Similaridade de Coseno, seguindo as seguintes etapas:

1) Extração dos *Mel Spectrograms* de cada arquivo de áudio usando a biblioteca *Librosa* via *Python*.
2) Normalizamos os *Mel Spectrograms* para garantir que eles estejam na mesma escala.
3) Usamos a função `cosine_similarity` do `scikit-learn` para calcular a similaridade de cosseno entre os Mel Spectrograms normalizados.
4) Exibimos o valor da similaridade de cosseno na saída.

Na Tabela III, apresentamos os resultados da similaridade de cosseno.

Tabela III
Resultados de comparação

Descrição	**Valor**
Dimensões de mel_spectrogram1	(128, 568)
Dimensões de mel_spectrogram2	(128, 568)
Similaridade de Cosseno	0,90473765

O valor de similaridade de cosseno é igual a 0,90473765. Essa medida varia entre 0 e 1, onde 1 indica uma similaridade perfeita e 0 indica nenhuma similaridade. Portanto, um valor de 0,90473765 indica uma alta similaridade entre os dois Mel Spectrograms. Isso sugere que os arquivos não são idênticos, mas são muito semelhantes, devido à alta similaridade de cosseno.

Com base nos resultados obtidos até o momento, podemos concluir que o objetivo principal deste trabalho foi alcançado com sucesso. Foi possível sintetizar uma voz masculina em português brasileiro, utilizando um dataset de vozes humanas brasileiras masculinas. A voz sintetizada demonstrou um alto

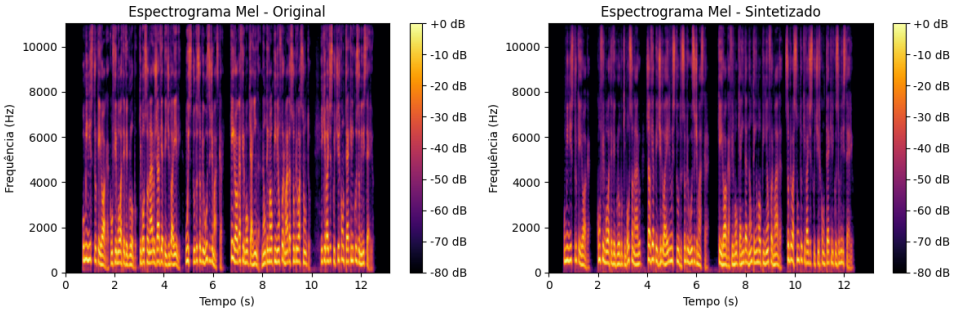

Figura 8. Comparativo entre os espectrogramas do arquivo de áudio original usado para treinamento e do arquivo com a voz sintetizada.

grau de similaridade com a voz original, evidenciando a qualidade da síntese realizada.

Para futuras pesquisas, recomenda-se explorar outras técnicas de síntese de fala a fim de aprimorar ainda mais a qualidade e expressividade das vozes sintetizadas. Além disso, a expansão do dataset utilizado, incorporando maior variedade de vozes e contextos linguísticos, pode contribuir para tornar as vozes sintéticas ainda mais realistas e versáteis.

Em resumo, este projeto representa um avanço significativo na tecnologia de síntese de fala em português brasileiro. Os resultados obtidos são valiosos para o desenvolvimento de aplicações e sistemas que dependem de vozes sintéticas, como assistentes virtuais, sistemas de leitura de textos e dispositivos de interação por voz.

Referências

[1] Taina Almeida Martins, Christiane Ratton Sanchez, Liriane Soares Araújo, "UM ESTUDO DA INTELIGÊNCIA ARTIFICIAL SOBRE IMPACTOS DAS ASSISTENTES VIRTUAIS NAS EMPRESAS," *Revista Interface Tecnológica*, vol. 19, no. 2, pp. 319–329, dez. 2022. Available: https://revista.fatectq.edu.br/interfacetecnologica/article/view/1543 doi: 10.31510/infa.v19i2.1543

[2] Pietro Sgarbosa and Gustavo Henrique Del Vechio. *INTELIGÊNCIA ARTIFICIAL E SUAS IMPLICAÇÕES: como os dispositivos inteligentes e assistentes virtuais influenciam o cotidiano das pessoas*. Revista Interface Tecnológica, volume 17, number 2, pages 193–205, dezembro de 2020. URL: https://revista.fatectq.edu.br/interfacetecnologica/article/view/936. doi: 10.31510/infa.v17i2.936.

[3] Jungil Kong, Jaehyeon Kim, Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17022–17033, 2020.

[4] Rongjie Huang, Chenye Cui, Feiyang Chen, Yi Ren, Jinglin Liu, Zhou Zhao, Baoxing Huai, Zhefeng Wang, "Singgan: Generative adversarial network for high-fidelity singing voice generation," *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 2525–2535, 2022.

[5] Shen, Jonathan, et al. *Tacotron 2: Generating human-like speech from text*. arXiv preprint arXiv:1712.05884 (2018).

[6] Davis, S., & Mermelstein, P. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), 357-366.

[7] Termos de Uso. https://www02.smt.ufrj.br/~gpa/terms_of_use.pdf

[8] Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu. *WaveNet: A generative model for raw audio*. arXiv preprint arXiv:1609.03499 (2016).

[9] Gonçalves, Ana Catarina Rosa. *Text-to-Speech Synthesis in European Portuguese using Deep Learning*. Disponível em: https://www.inesc-id.pt/publications/12757/pdf (novembro 2018).

[10] Reimao, Ricardo. *Synthetic Speech Detection using Deep Neural Networks*. Disponível em: https://core.ac.uk/download/pdf/240138805.pdf.

[11] Eirini, Sisamaki. *End-to-End Neural based Greek Text-to-Speech Synthesis*. Disponível em: https://www.csd.uoc.gr/~sspl/MSc/Sisamaki.pdf.

[12] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 4th edition, Pearson, 2022. ISBN: 978-0136680946.

[13] Casanova, Edresson. Disponível em: https://edresson.github.io/YourTTS. Acesso em: 24/08/2023.

[14] Lopes, Pedro. Disponível em: https://www.kaggle.com/datasets/pedrohlopes/tacotron2custom. Acesso em: 24/08/2023.

[15] Kobashikawa, Rodrigo. Disponível em: https://repositorio.ufsc.br/handle/123456789/228258.

[16] R. Goldberg and L. Riek, A Practical Handbook of Speech Coders, 1. Ed. CRC Press, New York, 2000, ISBN 9780849385254.

[17] D. O'Shaughnessy, "Linear predictive coding." IEEE Potentials, v. 7, 1988, pp. 29–32.

[18] H. Hermansky, Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Am., v. 57, n. 4, 1990, pp. 1738-–52.

[19] Z. K. Abdul and A. K. Al-Talabani, "Mel Frequency Cepstral Coefficient and its Applications: A Review,"in IEEE Access, vol. 10, pp. 122136-122158, 2022, doi: 10.1109/ACCESS.2022.3223444.

[20] N. Dave, "Feature extraction methods lpc, plp and mfcc in speech recognition." International journal for advance research in engineering and technology, v. 1, n. 6, 2013, pp. 1-–4.

[21] U. Shrawankar and V. M. Thakare, "Techniques for feature extraction in speech recognition system: A comparative study." arXiv preprint arXiv:1305.1145, 2013.

[22] B. P. Bogert and J. R. Healy and J. W Tukey, "The Quefrency Analysis of Time Series for Echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum, and Saphe Cracking," Proceedings of the Symposium on Time Series Analysis, 1963, pp. 209-243.

[23] A. Oppenheim and V. Alan Alan V and R. W. SCHAFER, "From frequency to quefrency: A history of the cepstrum." IEEE signal processing Magazine, IEEE, v. 21, n. 5, p. 95–106, 2004.

[24] A. V. Oppenheim and R. W. Schafer, "Processamento em Tempo Discreto de Sinais." Person Education, 2012. ISBN 978-85-8143-102-4.

[25] Sercan O. Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. Neural

[26] Perraudin, N.; Balazs, P.; Søndergaard, P.L. A fast griffin-lim algorithm. In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, 20–23 October 2013

[27] Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R. Natural TTS Synthesis by conditioning wavenet on mel spectrogram predictions. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Calgary, AB, Canada, 15–20 April 2018

[28] Win, Y.; Masada, T. Myanmar text-to-speech system based on tacotron-2. In Proceedings of the International Conference on Information and Communication Technology Convergence, Jeju, Republic of Korea, 21–23 October 2020.

[29] Wang, G.; Chen, M.; Chen, L. An end-to-end Chinese speech synthesis scheme based on Tacotron 2. J. East China Norm. Univ. 2019, 4, 111–119.