UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE ENGENHARIA ELÉTRICA E ELETRÔNICA
CURSO DE GRADUAÇÃO EM ENGENHARIA ELETRÔNICA

Rodrigo Kobashikawa Rosa

CONVERSÃO TEXTO-FALA PARA O PORTUGUÊS BRASILEIRO UTILIZANDO O MODELO TACOTRON-2 E O VOCODER GRIFFIN-LIM

Florianópolis
2021

Rodrigo Kobashikawa Rosa

CONVERSÃO TEXTO-FALA PARA O PORTUGUÊS BRASILEIRO UTILIZANDO O MODELO TACOTRON-2 E O VOCODER GRIFFIN-LIM

Trabalho Conclusão do Curso de Graduação em Engenharia Eletrônica do Centro Tecnológica da Universidade Federal de Santa Catarina como requisito para a obtenção do Título de Bacharel em Engenharia Eletrônica.
Orientador: Prof. Danilo Silva, Ph.D.

Florianópolis

2021

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Rosa, Rodrigo Kobashikawa
Conversão texto-fala para o português brasileiro utilizando o modelo Tacotron-2 e o vocoder Griffin-Lim / Rodrigo Kobashikawa Rosa ; orientador, Danilo Silva, 2021.
57 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Tecnológico, Graduação em Engenharia Eletrônica, Florianópolis, 2021.

Inclui referências.

1. Engenharia Eletrônica. 2. Treinamento de modelos do estado da arte de síntese de voz para o português brasileiro. 3. Tacotron 2. 4. Processamento de sinais de fala. I. Silva, Danilo. II. Universidade Federal de Santa Catarina. Graduação em Engenharia Eletrônica. III. Título.

Rodrigo Kobashikawa Rosa

CONVERSÃO TEXTO-FALA PARA O PORTUGUÊS BRASILEIRO UTILIZANDO O MODELO TACOTRON-2 E O VOCODER GRIFFIN-LIM

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de Bacharel em Engenharia Eletrônica e aprovado em sua forma final pelo Curso de Graduação em Engenharia Eletrônica

Florianópolis, 25 de Setembro de 2021.

Documento assinado digitalmente
Fernando Rangel de Sousa
Data: 27/09/2021 23:19:54-0300
CPF: 884.649.114-91
Verifique as assinaturas em https://v.ufsc.br

Prof. Fernando Rangel de Sousa, Dr.

Coordenador do Curso

Banca Examinadora:

Documento assinado digitalmente
Danilo Silva
Data: 27/09/2021 21:22:31-0300
CPF: 031.557.924-26
Verifique as assinaturas em https://v.ufsc.br

Prof. Danilo Silva, Ph.D.

Orientador

Universidade Federal de Santa Catarina

Documento assinado digitalmente
Marcio Holsbach Costa
Data: 27/09/2021 21:31:46-0300
CPF: 572.170.680-53
Verifique as assinaturas em https://v.ufsc.br

Prof. Márcio Holsbach Costa, Dr.

Universidade Federal de Santa Catarina

Documento assinado digitalmente
RANNIERY DA SILVA MAIA
Data: 27/09/2021 20:09:50-0300
CPF: 967.987.424-91
Verifique as assinaturas em https://v.ufsc.br

Dr. Ranniery Maia

Universidade Federal de Santa Catarina

Este trabalho é dedicado aos meus pais, amigos e colegas.

AGRADECIMENTOS

Sou grato à minha família, meu pai, mãe e irmã, pelo apoio que recebi durante a minha graduação, sempre me motivando a ir além, sem desistir, para assim conquistar meus objetivos e concluir o sonho de me formar em engenharia em uma ótima universidade como a UFSC.

Agradeço muito ao meu orientador Danilo Silva, por todos os ensinamentos passados que se estenderam muito além do desenvolvimento deste trabalho, sendo em momentos quase um psicólogo, disposto a ouvir as minhas frustrações e ajudar da melhor forma possível. Fico muito feliz de o ter tido como meu orientador.

Um agradecimento para todos os meus colegas de graduação que em vários momentos passamos por muitas dificuldades juntos, ajudando uns aos outros para chegar a este momento.

Agradeço a todos os professores que tive durante a graduação que contribuíram no meu desenvolvimento profissional, como também no meu desenvolvimento pessoal.

E por último agradeço a todas as pessoas com quem eu convivi durante este período pandêmico. Cada um ajudou de alguma forma, mesmo não percebendo, através de pequenas conversas, risadas e experiências compartilhadas.

“In a properly automated and educated world, then, machines may prove to be the true humanizing influence. It may be that machines will do the work that makes life possible and that human beings will do all the other things that make life pleasant and worthwhile.”

(ASIMOV, 1990)

RESUMO

A síntese de fala é uma área de pesquisa antiga, motivada pelo desejo humano de fazer as máquinas falarem e interagirem como humanos. Durante muito tempo, os resultados obtidos estavam muito longe da fala humana natural devido à complexidade do aparelho fonador humano. Porém, com o advento do aprendizado profundo, novas arquiteturas de redes neurais estão aparecendo e os modelos do estado da arte estão conseguindo sintetizar falas tão naturais quanto as de humanos reais, sendo quase imperceptível a diferença. Neste trabalho será apresentado o treinamento de um modelo do estado da arte com redes neurais, o Tacotron-2. Será utilizado um conjunto de dados de fala de código aberto do projeto Common Voice em português brasileiro. Foram avaliados os resultados do treinamento do modelo do zero e da aplicação de *transfer learning* a partir de um modelo pré-treinado em inglês. Os resultados mostraram que é possível treinar o modelo com recursos de dados limitados, a partir da avaliação da inteligibilidade dos modelos e da qualidade do áudio sintetizado.

Palavras-chave: síntese de fala. Redes Neurais. Tacotron 2. Griffin-Lim.

ABSTRACT

Speech synthesis is an old research field, motivated by the human desire of making machines talk and interact as humans. For a long time, the obtained results were very far from natural human speech due to the complexity of the human speech organs. However, with the advent of deep learning, new neural networks architectures have been appearing and the state of the art models are capable of synthesizing voices as natural as of real humans, with the difference being almost imperceptible. In this work it will be presented the training of a state-of-the-art neural network model, Tacotron-2. It will also use an open-source brazilian portuguese voice dataset from the Common Voice project. Results from training the model from scratch and by applying transfer learning of a pre-trained english model were evaluated. The results show that it is possible to train the model with limited data resources, from the evaluation of the models intelligibility and synthesized audio quality.

Keywords: Speech synthesis. Neural Networks. Tacotron-2. Griffin-Lim.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DFTF	Discrete Time Fourier Transform
DNN	Deep Neural Network
FFT	Fast Fourier Transform
GAN	Generative Adversarial Network
GLA	Griffin-Lim Algorithm
GPU	Graphics Processing Unit
GRU	Gated Recurrent Unit
HMM	Hidden Markov Model
ISTFT	Inverse Short Time Fourier Transform
LSTM	Long Short Term Memory
MFCC	Mel Frequency Cepstral Coefficients
MSE	Mean Squared Error
RNN	Recurrent Neural Network
Seq2Seq	Sequence to Sequence
STFT	Short Time Fourier Transform
SUE	Synthesis Unit Entry
TTS	Text to Speech

SUMÁRIO

1 INTRODUÇÃO

Há muito tempo engenheiros e cientistas produzem pesquisas na área de processamento de fala inspirados nas máquinas de ficção científica capazes de ouvir e falar. Em codificação de sinais e reconhecimento de fala, os sistemas foram evoluindo progressivamente, porém a área de geração de fala foi a que teve mais dificuldade para obter sucesso (TAYLOR, 2009).

As tentativas dos pesquisadores de imitarem os processos físicos da geração de fala usando modelos articulatórios do aparelho fonador humano, ou por modelos de síntese das propriedades temporais e espectrais variantes no tempo da fala eram complexas (KLATT, 1987). Isso resultou que por muito tempo a fala sintética gerada por essas máquinas não tinha a mesma naturalidade da fala humana (TAYLOR, 2009).

Os principais objetivos para a criação de um sistema capaz de sintetizar fala são construir um sistema que seja fácil de ser entendido e fazer isso utilizando uma voz parecida com a de um humano. Esses dois objetivos são descritos pela comunidade científica como inteligibilidade e naturalidade (RABINER, 2010). Nos últimos anos, tivemos grandes avanços nas tecnologias desenvolvidas e os sistemas atuais de síntese de fala conseguem gerar falas muito próximas às de humanos (SHEN et al., 2018).

A complexidade da comunicação humana trouxe muita dificuldade para esses sistemas devido à necessidade de etapas de pré-processamento para a extração dos parâmetros acústicos e linguísticos (TAYLOR, 2009). Até recentemente os modelos tradicionais mais utilizados em aplicações comerciais eram o concatenativo (SAGISAKA, 1988), baseado na concatenação de trechos de fala e o paramétrico estatístico (ZEN; TOKUDA; BLACK, 2009).

Com a popularização do aprendizado profundo para as mais diversas aplicações, surgiram também usos para a área de síntese de fala (PURWINS, 2019) . Em aplicações em que se tinha disponibilidade suficiente de dados, os modelos de redes neurais profundas muitas vezes tiveram performances superiores aos modelos tradicionais. Os modelos de *Text-to-Speech* (TTS) *end-to-end* com redes neurais, conseguem ser treinados diretamente com pares de áudio e texto sem precisar das etapas intermediárias de extração de atributos, pois o próprio modelo é capaz de fazer esse trabalho e também de facilitar o condicionamento do modelo para outros falantes, linguagens diferentes (ZHANG et al., 2019) ou até sentimentos diferentes na fala (SKERRY-RYAN et al., 2018).

A pesquisa em relação aos modelos estado da arte com redes neurais profundas no Brasil não possui a mesma disponibilidade de recursos de código aberto como os vários conjuntos de dados em linguagens como o inglês e mandarim e portanto existem poucos trabalhos sobre o assunto. Um deles é o trabalho de Casanova (2020), que propôs um conjunto novo conjunto de dados aberto para uso e fez vários experimentos com implementações do Tacotron 2 (SHEN et al., 2018) e DCTTS (TACHIBANA et al., 2018). Utilizando caracteres ou fonemas como entrada, algoritmo de redução de ruído e o vocoder RTISI-LA, tal trabalho conseguiu resultados comparáveis ao trabalho de Quintas (2020) que avaliou o Tacotron 2 para o português europeu.

O presente trabalho busca treinar o modelo em um conjunto de dados de código aberto adaptado, utilizando o modelo Tacotron 2 (SHEN et al., 2018) , com o vocoder Griffin-Lim (GRIFFIN; LIM, 1984) em vez do WaveNet (OORD et al., 2016) devido à limitação de recurso computacional disponibilizado para o treinamento, e avaliar os resultados obtidos para o português brasileiro.

No capítulo 2, será descrito o conhecimento teórico a respeito dos fundamentos da síntese de fala, abordando tópicos como a fonologia do trato vocal humano, processamento digital de sinais de fala, histórico dos modelos tradicionais de TTS e introdução aos modelos do estado da arte utilizando aprendizado profundo. No capítulo 3, é apresentada a metodologia para o treinamento do modelo Tacotron 2, desde a preparação do conjunto de dados, pré-processamento dos dados, treinamento do modelo e resultados do treinamento. No capítulo 4, é feita uma avaliação dos resultados do modelo utilizando um conjunto de validação. No último capítulo, são apresentadas as conclusões finais e sugestões para trabalhos futuros

1.1 OBJETIVOS

O objetivo geral e os objetivos específicos do presente trabalho são descritos a seguir.

1.1.1 Objetivo Geral

O objetivo deste trabalho é treinar e avaliar um conversor texto-fala para o português brasileiro baseado em modelos do estado da arte.

1.1.2 Objetivos Específicos

- Aprofundar, como também aprender novos conceitos relativos à área de processamento de sinais de áudio, em específico os de fala;
- Estudar os principais modelos utilizados na literatura para sistemas TTS;
- Treinar um sistema de conversão texto-fala, aplicar para o português brasileiro e avaliar o resultado.

2 FUNDAMENTOS DA SÍNTESE DE FALA

Desde o começo do desenvolvimento dos sistemas de síntese de fala, houve muitos avanços na qualidade da fala sintetizada. Durante esses longos anos de pesquisa muito se aprendeu em relação ao funcionamento da produção humana da fala, do processamento digital do sinal de fala e dos modelos para a síntese de fala a partir dos mais variados modelos. Neste capítulo será apresentada uma introdução sobre cada um desses assuntos.

2.1 FONÉTICA E FONOLOGIA

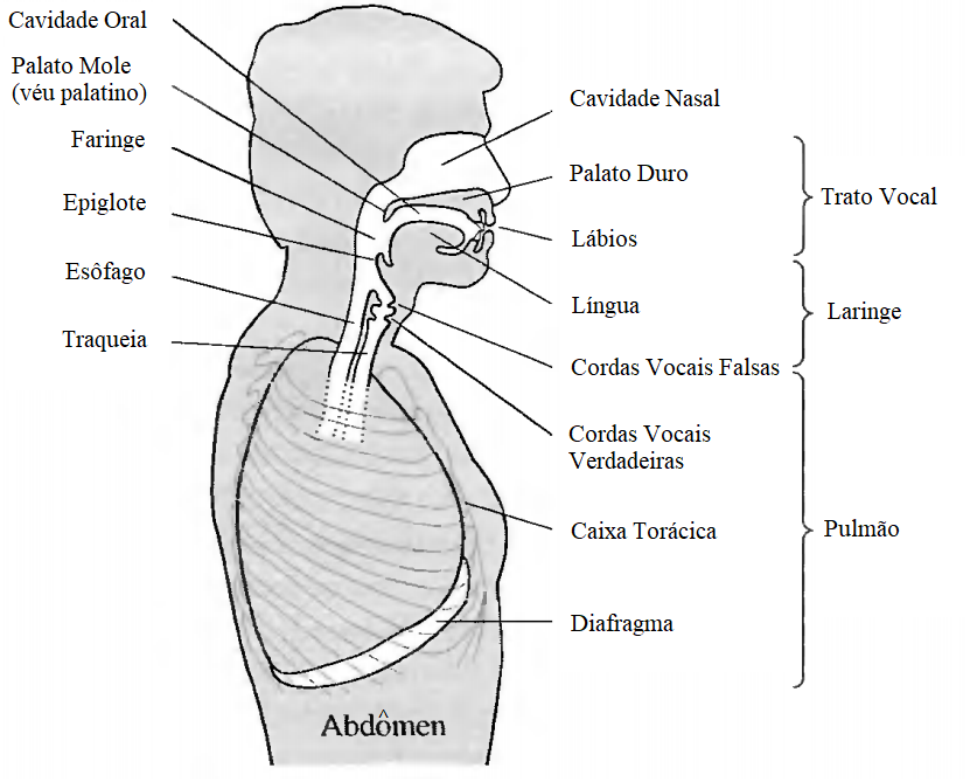
O conhecimento da forma que cada pessoa utiliza seu aparelho fonador não é completo de forma que seja possível saber exatamente como um som de fala é produzido. Todo o processo é coordenado pela utilização de articuladores anatômicos conhecidos como órgãos vocais e a fala final produzida é o resultado de variadas configurações dos órgãos vocais. Na Figura 1 podem ser vistos os órgãos responsáveis pela produção da fala. Devido à complexidade do processo de produção ser afetada por diversos fatores, por muito tempo os modelos não eram capazes de prever com um elevado grau de acurácia como seria a forma de onda produzida a partir de uma determinada pronúncia (TAYLOR, 2009).

A maioria dos sons são produzidos a partir da excitação gerada pelos pulmões, que faz com que o ar se mova a partir deles através dos órgãos vocais para os lábios e então para fora. As constrições desse fluxo de ar na laringe, trato vocal, boca e nariz são responsáveis pela geração do som (TAYLOR, 2009).

As cordas vocais são duas dobras de tecido na laringe que podem estar totalmente fechadas, estreitas ou abertas. O espaço entre as cordas vocais é chamado de glote. Quando as cordas vocais formam uma abertura estreita e uma corrente de ar passa por ela causam uma vibração, criando um som periódico (LOIZOU, 2013). A taxa de vibração das cordas vocais é chamada de frequência fundamental, denotada por F0. O termo pitch é usado pela taxa de vibração percebida pelo ouvinte e em geral é igual à F0 (TAYLOR, 2009).

Variando a tensão nas cordas vocais, o falante consegue mudar a F0 do som produzido. Homens típicos vibram as cordas vocais entre 80-250 vezes por segundo (80-250 Hz) e mulheres entre 120-400 Hz (TAYLOR, 2009).

Figura 1 - Visão em corte transversal da anatomia da produção de fala.



Fonte: Adaptado de Quatieri (2001).

As vogais são sons vozeados que além da frequência fundamental, possuem energia em outras frequências conhecidas como harmônicas, múltiplas da fundamental. O efeito das harmônicas dá ao som seu timbre.

Se a glote se abre um pouco mais, a vibração periódica para e um fluxo de ar não periódico e turbulento é criado, gerando um tipo diferente de som, não harmônico. É então possível usar uma combinação da língua, lábios e dentes para formar construções e gerar sons não glotais, chamados de não vozeados.

Entre as características do sinal de fala, aponta-se como uma das mais importantes a sua não estacionariedade devido o seu espectro de potência mudar com o tempo. Porém, em curtos intervalos de tempo de 10 a 30 ms, suas características espectrais podem ser consideradas estacionárias. No tempo, possui segmentos quase-periódicos, aleatórios e de silêncio e na amplitude apresenta segmentos de alta e baixa intensidade, com durações e características espectrais diferentes (LOIZOU, 2013).

2.2 PROCESSAMENTO DE SINAIS DE FALA

A fala é um sinal contínuo gerado a partir da pressão do ar exalado pelo pulmão passando pelas constrições na laringe e no trato vocal. Uma das possíveis formas de tratamento da fala é a digital, realizada a partir da discretização do sinal. A seguir serão apresentadas algumas das principais técnicas de processamento discreto de fala.

2.2.1 Transformada de Fourier de tempo curto

Uma forma consagrada para a análise dos sinais de fala é o espectro de magnitude, que pode ser realizada através da *Short Term Fourier Transform* (STFT).

Quando falamos, a glote e o trato vocal estão constantemente mudando e isso se torna problemático, pois a estimação do espectro através da transformada de Fourier requer a estacionariedade do sinal. Modelando a forma de onda como uma série de trechos de tempo curto de fala, pode-se considerar cada um deles como um sinal estacionário.

Um trecho de fala $x[n]$é obtido fazendo a multiplicação da janela $w[n]$ pelo sinal de fala completo $s[n]$:

$$x[n] = w[n]s[n] \tag{1}$$

As três janelas mais comuns são a retangular, de hanning e de hamming definidas da seguinte forma:

$$retangular \quad w[n] = \begin{cases} 1 & 0 \le n \le L-1 \\ 0 & outros \end{cases} \tag{2}$$

$$hanning \quad w[n] = \begin{cases} 0.5 - 0.5cos(\frac{2\pi n}{L}) & 0 \le n \le L-1 \\ 0 & outros \end{cases} \tag{3}$$

$$hamming \quad w[n] = \begin{cases} 0.54 - 0.46cos(\frac{2\pi n}{L}) & 0 \le n \le L-1 \\ 0 & outros \end{cases} \tag{4}$$

Temos então que a STFT é dada pela equação (5), em que $x(m)$é o sinal de entrada e $w(m)$ é a janela de análise. É uma função de duas variáveis, sendo n o índice de tempo discreto e ω a variável de frequência contínua.

$$X(n,\omega) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)e^{-j\omega n} \quad (5)$$

A forma discreta da STFT é obtida a partir da amostragem da frequência ω para N frequência uniformemente espaçadas, de forma que $\omega_k = 2\pi k/N$, $k = 0, 1,..., N-1$.

$$X(n,k) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)e^{-j\frac{2\pi}{N}kn} \quad (6)$$

A STFT pode ser interpretada como a *Discrete Time Fourier Transform* (DTFT) da sequência $x(m)w(n-m)$quando assumido que n é fixo. E portanto, a STFT tem as mesmas propriedades que a DTFT. Outra possível interpretação da STFT pode ser a visualização dela como a saída de uma operação de filtragem. Onde que $w(n)$ atua como um filtro de resposta ao impulso aplicado ao sinal $x(n)$ deslocado no domínio da frequência por ω_k (LOIZOU, 2013).

Na análise do sinal de fala através da STFT, existe um problema chamado de compromisso entre tempo-frequência. Pela definição da DFT, para se obter uma alta resolução na frequência é necessário um grande número de amostras da forma de onda. Mas conforme o tamanho do trecho aumenta, o comportamento não estacionário passa a ser relevante. E ao tentar usar uma janela pequena, se houver poucas amostras, a estimativa do espectro se torna ruim (TAYLOR, 2009).

Além do tamanho do quadro, deve-se considerar o quão frequente deve-se calcular o espectro. O deslocamento do quadro é definido como a distância de dois quadros consecutivos. Para diferentes aplicações, diferentes deslocamentos são desejados. Normalmente usa-se como base o período em que o trato vocal pode ser considerado invariante no tempo para a escolha do tamanho do trecho e deslocamento. Valores típicos são 25ms para o tamanho do quadro e 10ms para o deslocamento. Esses intervalos de tempo são expressos em números de amostras, a partir da multiplicação da taxa de amostragem pelo período e escolhidos usando potências de dois devido a utilização de FFT nas implementações computacionais (TAYLOR, 2009).

Uma forma usual de visualização de sinais não-estacionários é o espectrograma. O espectrograma é um gráfico tridimensional em que a abscissa denota o tempo, a ordenada denota a frequência e a cor do gráfico representa a potência espectral da fala.

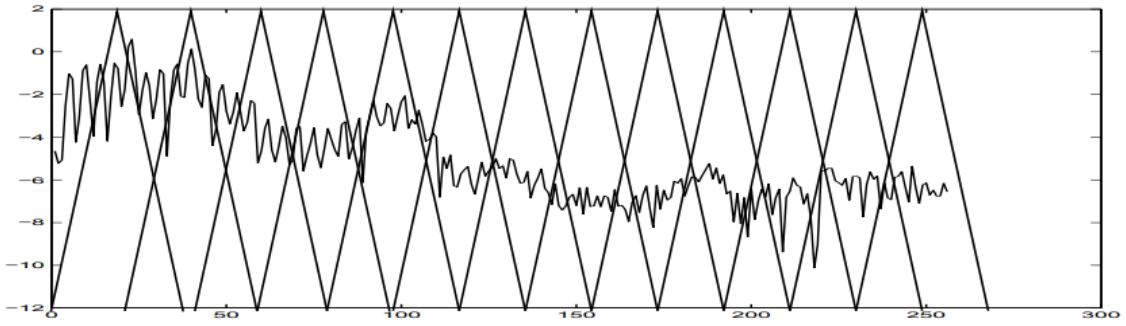
2.2.2 Espectrograma mel

É conhecido que a sensibilidade humana para a frequência não é linear, conforme o aumento da frequência fica mais difícil a percepção da altura (característica subjetiva do som relacionada à frequência fundamental). Estudos sobre o comportamento psicoacústico de percepção da altura, levou à elaboração de escalas auditivas em uma nova escala de frequências com maior associação à sensibilidade humana.

Segundo Taylor (2009), a escala mel foi criada a partir de experimentos com senóides em que os indivíduos do teste tentavam dividir faixas de frequências em seções espaçadas em distâncias iguais baseadas nos seus julgamentos. Empiricamente ficou definido que um tom de 1000 Hz corresponde a um *pitch* de 1000 mels e o seu mapeamento da escala de frequência linear para a escala mel ficou da seguinte forma:

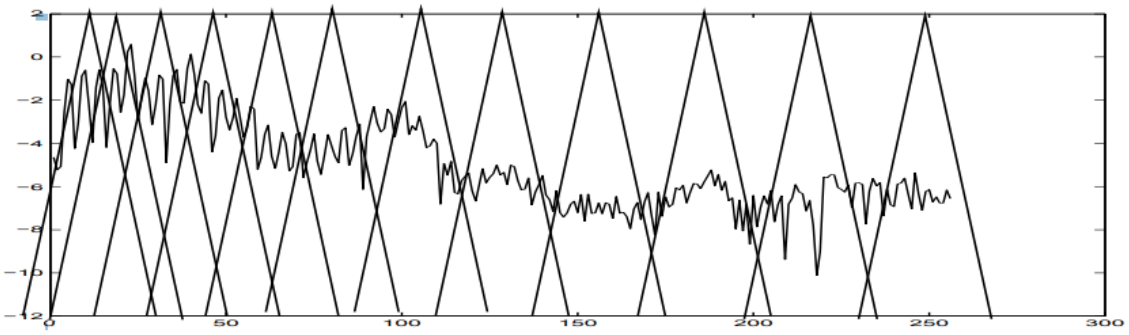
$$m = 2595\, log10(1 + f/700) \qquad (7)$$

Figura 2 - Análise de um banco de filtros igualmente espaçado sobre o espectro da fala.



Fonte: Taylor (2009).

Figura 3 - Análise de um banco de filtros espaçado de acordo com a escala mel.



Fonte: Taylor (2009).

Com o conhecimento sobre a percepção da altura, uma representação do sistema auditivo humano pode ser simulada a partir de um banco de filtros passa-banda, cuja largura de banda aumenta de acordo com o aumento da frequência central (RABINER, 2010).

Na Figura 2 pode ser visto o espectro de um sinal de fala e os bancos de filtros triangulares espaçados igualmente na frequência, enquanto que na Figura 3 espaçados de acordo com a escala mel, representando a percepção do ouvido humano.

Os coeficientes de frequência mel cepstrais (MFCCs) foram por muito tempo um dos atributos de representação acústica dominantes em análise de áudio. Os MFCCs são a projeção dos espectros de magnitude em um conjunto reduzido de bandas de frequência mel e comprimidos com uma transformada de cosseno discreto (DCT). Para os modelos baseados em aprendizado profundo, esse último passo é desnecessário por remover informação e destruir relações espaciais. O nome da transformação sem esse último passo é chamado de espectro log-mel (PURWINS, 2019).

2.3 MODELOS TRADICIONAIS

A síntese de fala é um problema que há décadas vem sendo estudado e diversas técnicas diferentes foram utilizadas ao passar do tempo. As técnicas tradicionais que se mantiveram por muito tempo como o estado da arte e dominaram os sistemas comerciais de TTS foram a síntese concatenativa e a paramétrica.

Nos anos 70, a ideia de utilizar a concatenação de unidades básicas de fala, difonos representando pares de fonemas, foi identificada como uma forma prática. Porém, mais de uma década depois, apesar de ter possibilitado uma elevação de inteligibilidade, o resultado

ainda não era muito natural, portanto não sendo utilizado em aplicações reais (TAYLOR, 2009).

No final da década de 80, Yoshinori Sagisaka no Japão trouxe um grande avanço ao utilizar vários *tokens* de difonos, extraídos com cuidado de bancos de dados de fala. Tendo disponíveis milhares de *tokens* de cada possível difono da linguagem, resta apenas o trabalho de concatenar a sequência correta para gerar a fala humana natural. O método utilizado foi o de seleção unitária, em que cada fonema é segmentado em um banco de dados e um índice da unidade de fala é criado baseado nos parâmetros acústicos como a F0, duração, posição da sílaba e fonemas vizinhos. Um dicionário de *synthesis unit entries* (SUE) foi proposto para reduzir o tempo de busca do fonema usando um formato ordenado por tamanho do fonema e em estrutura de árvore (SAGISAKA, 1988).

A síntese concatenativa necessitava de bancos de dados de fala enormes contendo cada fonema, sílaba, palavra e frases como unidades para cobrir todas as possíveis combinações das unidades de fala. É uma técnica com alto grau de inteligibilidade, porém a concatenação entre fonemas resulta na dificuldade de se obter transições suaves entre fonemas, comprometendo também a expressividade da fala sintetizada e a sua naturalidade (TAN et al., 2021).

A síntese paramétrica se refere ao método que utiliza tecnologias de processamento digital de sinal para sintetizar fala. Nesse método o processo vocal humano é simulado a partir de uma fonte de sinal periódico representando as vibrações das cordas vocais dos sons vozeados e do ruído branco para indicar os sons não vozeados, que então passa por um filtro digital variante no tempo que caracteriza as propriedades ressonantes do canal. Ajustando os parâmetros desse filtro é possível sintetizar vários tipos de fala. Métodos típicos são a síntese da fala baseada no modelo oculto de Markov (HMM) e a síntese da fala baseada em redes neurais profundas (NING, 2019).

A sínteses paramétrica estatística propôs resolver alguns dos problemas da síntese concatenativa dividindo o processo em três módulos: um de análise textual, outro para a predição dos parâmetros acústicos utilizando um modelo estatístico e por fim o módulo do vocoder. O módulo de análise textual processa o texto, normalizando-o, fazendo a conversão grafema para fonema e a segmentação das palavras, para então extrair os atributos linguísticos. Os modelos acústicos como os de HMM e de redes neurais profundas são treinados com pares dos atributos linguísticos e acústicos (F0, espectro ou cepstro, duração) e

os vocoders sintetizam a fala baseados nesses atributos acústicos preditos (TAN et al., 2021). Os três módulos podem ser vistos na Figura 4 abaixo.

Figura 4 - Fluxo de um sistema TTS tradicional.

Fonte: Elaborado pelo autor.

A síntese paramétrica estatística é usualmente dividida em duas fases: a de treinamento e a de síntese. Na fase de treinamento, os atributos acústicos como a F0 e os parâmetros espectrais são extraídos do corpus para que seja treinado o modelo acústico estatístico junto dos atributos linguísticos resultantes do módulo de processamento linguístico. Na fase de síntese os atributos acústicos são preditos usando o modelo acústico treinado com a orientação dos atributos linguísticos (NING, 2019).

As vantagens da síntese paramétrica estatística estão na naturalidade do áudio, a flexibilidade para modificar parâmetros e controlar a fala e o baixo custo de dados, por necessitar de muito menos do que a síntese concatenativa. Porém, também existem desvantagens em relação à concatenativa: uma delas seria a inteligibilidade inferior devido a artefatos de áudio, soando abafado ou com ruídos; outra seria a produção de uma fala contínua robótica e facilmente diferenciável com a de um humano (TAN et al., 2021).

2.4 MODELOS DE REDES NEURAIS PROFUNDAS

O aprendizado profundo é uma área de pesquisa emergente nos últimos anos, a aplicação de novas arquiteturas de modelos de redes neurais profundas teve um grande valor na área de processamento de fala. Em aplicações de tradução (SUTSKEVER, 2014), reconhecimento de fala (GRAVES; MOHAMED; HINTON, 2013), síntese de fala (WANG et al., 2017) e outras soluções de NLP. O uso das redes neurais profundas resultou em um grande avanço em relação ao que era o estado da arte anteriormente.

As redes neurais profundas são a aplicação das redes neurais artificiais usando múltiplas camadas para aumentar o poder de aprendizagem de representação. São estruturas inspiradas no cérebro humano, possuindo vários neurônios (unidades de processamento da informação) organizados em camadas que vão ajustando os pesos das conexões entre cada

neurônio durante o treinamento através do algoritmo de *backpropagation* (RUMELHART; HINTON; WILLIAMS, 1986).

Existem alguns tipos diferentes de redes neurais desenvolvidos para diferentes aplicações. Os principais seriam as redes neurais *feedforward*, as convolucionais e as recorrentes. Para o processamento de dados sequenciais, as redes neurais recorrentes são as mais usualmente utilizadas, pois elas estendem a ideia das redes *feedforward* ao incluir conexões de realimentação que funcionam como memória ao processar sequências (GOODFELLOW et al., 2017).

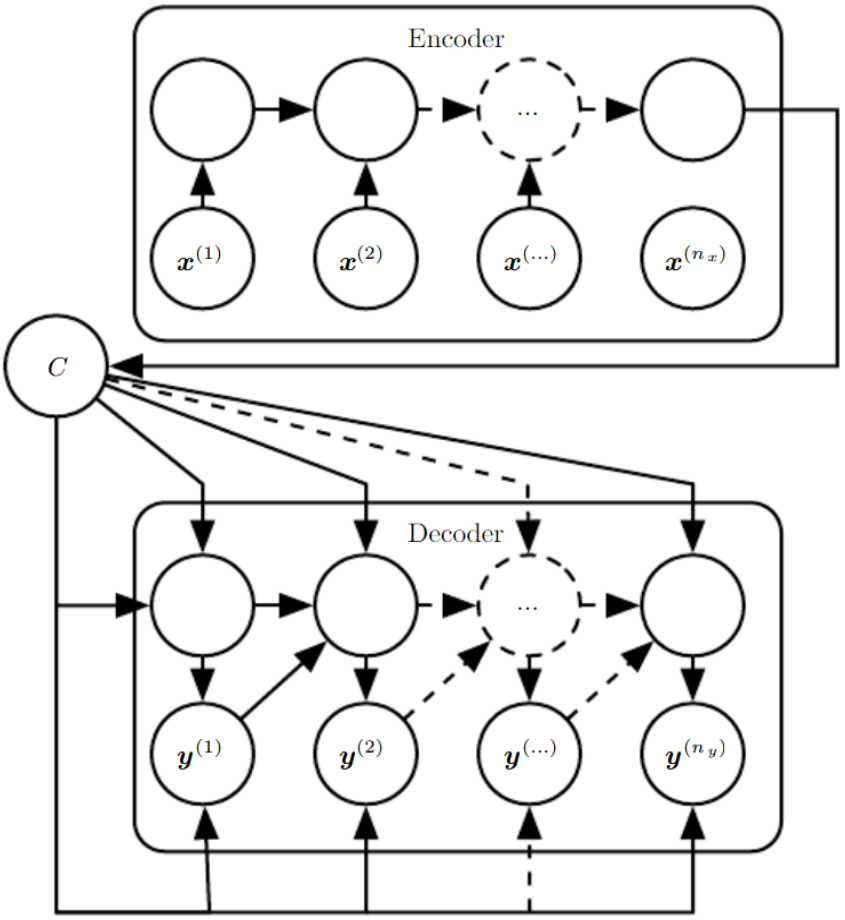
Nas próximas seções serão apresentadas algumas arquiteturas de redes neurais profundas que permitiram o salto no desenvolvimento da síntese de fala tradicional para os modelos *end-to-end* que são o estado da arte atual.

2.4.1 Modelos sequence-to-sequence

Com uma rede neural recorrente (RNN) é possível mapear uma sequência de entrada em um vetor de tamanho fixo, mapear um vetor de tamanho fixo em uma sequência, ou até mesmo uma sequência de entrada para outra sequência, desde que sejam do mesmo tamanho. Porém, várias aplicações necessitam mapear sequências de entrada para sequências de saída de tamanhos diferentes, como a tradução, reconhecimento de fala e sistemas TTS.

Nos artigos de Cho et al. (2014) e Sutskever et al. (2014) foram introduzidas a arquitetura sequence-to-sequence (Seq2Seq). O funcionamento desta arquitetura consiste de um codificador e um decodificador, podendo ser visto na Figura 5, em que as entradas do codificador são representadas pelos círculos com $x^{(n)}$ e as saídas do decodificador pelos círculos com $y^{(n)}$. O codificador processa cada item da sequência de entrada, compilando a informação em um vetor de contexto para então ser utilizado pelo decodificador para produzir a sequência de saída item por item. Tanto o codificador quanto o decodificador são RNNs e o tamanho do vetor de contexto tem o mesmo tamanho do número de unidades ocultas na RNN do codificador.

Figura 5 - Exemplo de uma arquitetura sequence-to-sequence.



Fonte: Goodfellow (2017)

Um problema da utilização da representação em um único vetor compilado de contexto é que torna-se difícil a captura de toda informação da sequência quando se torna muito longa, pelo fato do vetor de informação ser de tamanho fixo e muitas vezes menor do que a sequência de entrada. Nessa situação é necessário que aumente o tamanho da RNN, para que a perda de informação não ocorra. Devido a esse problema identificado, Bahdanau et al. (2014) propôs o mecanismo de atenção.

2.4.2 Modelo de atenção

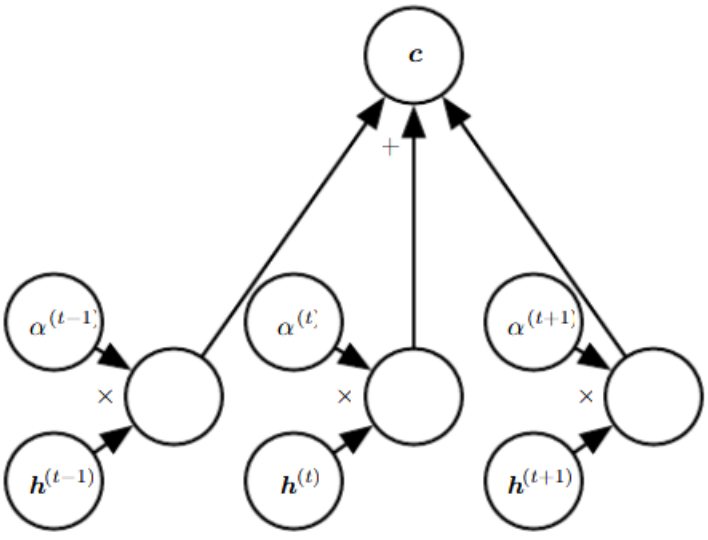
Para resolver o problema de captura de informações de sequência longas, uma abordagem mais eficiente de gerar a sequência de saída focando em diferentes partes da sequência de entrada foi proposta por meio do mecanismo de atenção por Bahdanau et al. (2014) e Luong et al. (2015).

Através do modelo de atenção, o codificador consegue passar muito mais informação para o decodificador, passando todos os *hidden states* de cada unidade da RNN do codificador em vez de apenas a última, que após ser utilizada como entrada da próxima unidade da RNN, contém toda a informação passada compilada e chamada de vetor de contexto.

O mecanismo de atenção atua no decodificador que passa por esse passo adicional de geração do novo vetor de contexto antes de produzir a saída. Em cada passo do decodificador são feitos as seguintes etapas:

1. Dentro do conjunto de *hidden states* do codificador. Cada *hidden state*, denotado por $h(t)$ na Figura 6, está mais correlacionado a um certo item da sequência de entrada.
2. Calculada uma nota para cada *hidden state,* aplica-se uma função *softmax* que resultará no peso de cada vetor de *hidden state*, denotado por $\alpha(t)$ na Figura 6.
3. É feita uma soma ponderada, multiplicando cada *hidden state* pelo seu peso calculado no passo anterior.
4. O novo vetor de contexto do passo do decodificador, denotado por **c** na Figura 6, será o resultado dessa soma ponderada.

Figura 6 - Mecanismo de atenção.



Fonte: Goodfellow (2017)

Sumarizando, dado o conjunto de *hidden states* ($h^{(t)}$) gerados pelo codificador, o mecanismo de atenção funciona dando um peso para cada *hidden state* e calculando a soma ponderada entre eles para gerar o vetor de contexto específico daquele passo do decodificador.

2.4.3 Modelos do estado da arte

A síntese de fala TTS é uma área de pesquisa com muitas ideias inovadoras e que nos últimos anos tem se desenvolvido muito, através de diversas arquiteturas de modelos. Em Ning et al. (2019) e Tan et al. (2021) é possível ter uma ideia da evolução dos trabalhos, sendo apresentada uma revisão dos modelos de aprendizado profundo.

Os primeiros trabalhos implementando DNNs tiveram como objetivo a predição de parâmetros acústicos em sistemas paramétricos estatísticos (ZEN; SENIOR; SCHUSTER, 2013), substituindo os modelos baseados em HMM. Com o avanço enorme que modelos de redes neurais tiveram a partir de 2009, surgiram as redes Seq2Seq que foram introduzidas inicialmente em diversas tarefas, como tradução e reconhecimento de fala, onde obtiveram resultados superiores aos modelos do estado da arte. Como a síntese de fala é o problema inverso do reconhecimento de fala, a técnica foi explorada em diversas arquiteturas novas dando luz à uma nova era de sistemas TTS *end-to-end,* obtendo resultados com níveis de naturalidade da fala sintetizada altos o suficiente, tornando-os difíceis de serem distinguidos da fala humana.

Os principais modelos do estado da arte, Deep Voice 3 (PING et al., 2017) , Tacotron (WANG et al., 2017) e Tacotron 2 (SHEN et al., 2018) mudaram a forma como é feito TTS atualmente e tiveram uma grande atenção por parte dos pesquisadores, levando a várias modificações e adições de funcionalidades para esses modelos. Nas próximas subseções serão apresentados o funcionamento desses três modelos.

Vale mencionar que desde 2018 a pesquisa na área está bem aquecida e novos modelos vêm sendo desenvolvidos como o Transformer TTS (LI et al., 2019) e o Flowtron (VALLE et al., 2021). Além disso, modelos que não são auto regressivos como o Fast Speech (REN et al., 2019), Fast Speech 2 (REN et al., 2021), Glow TTS (KIM et al., 2019) e o Flow TTS (MIAO et al., 2020) estão tornando o treinamento e a inferência mais rápidos e viáveis de serem utilizados com menos recursos computacionais.

Com o surgimento dos novos modelos, pode ser feito um agrupamento das arquiteturas da seguinte forma:

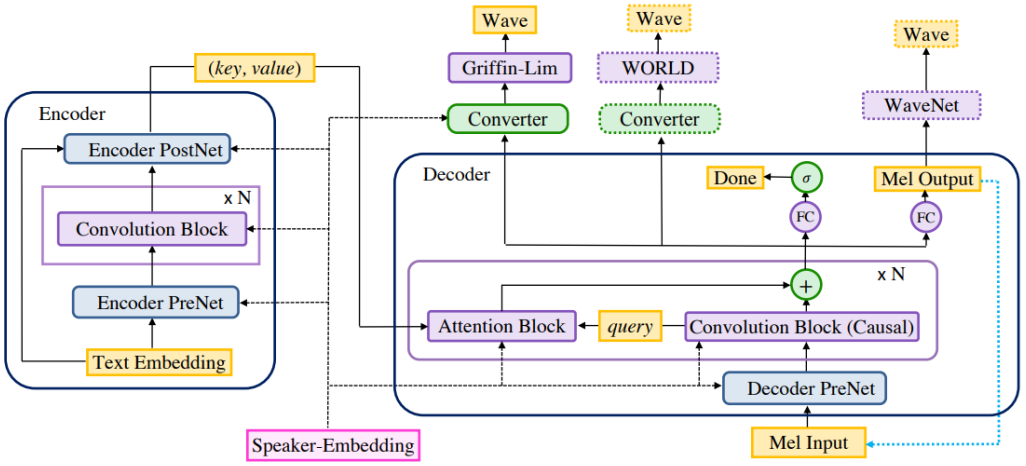
1) Baseados em redes recorrentes, como por exemplo os modelos Tacotron;
2) Baseados em redes convolucionais, como por exemplo os modelos Deep Voice;
3) Baseados em *Transformers*, como o Transformer TTS e os modelos Fast Speech;

4) E outros modelos acústicos (ex: baseados em *Generative Adversarial Network* (GAN) (GOODFELLOW et al., 2014), Flow)

2.4.3.1 *Deep Voice 3*

No artigo de Ping et al. (2017) é descrito o Deep Voice 3 como um modelo de conversão texto-fala neural baseado em uma arquitetura convolucional da Baidu para geração de espectrogramas permitindo a otimização do modelo por conseguir paralelizar suas operações, algo que não é possível com modelos de redes recorrentes. Foi utilizado um modelo de atenção no decodificador e uma outra característica é a possibilidade de se utilizar os vocoders Griffin-Lim, WORLD e WaveNet para a síntese final da fala. Uma visualização dessa arquitetura pode ser vista na Figura 7.

Figura 7 - Arquitetura do modelo do Deep Voice 3.



Fonte: Ping et al. (2017)

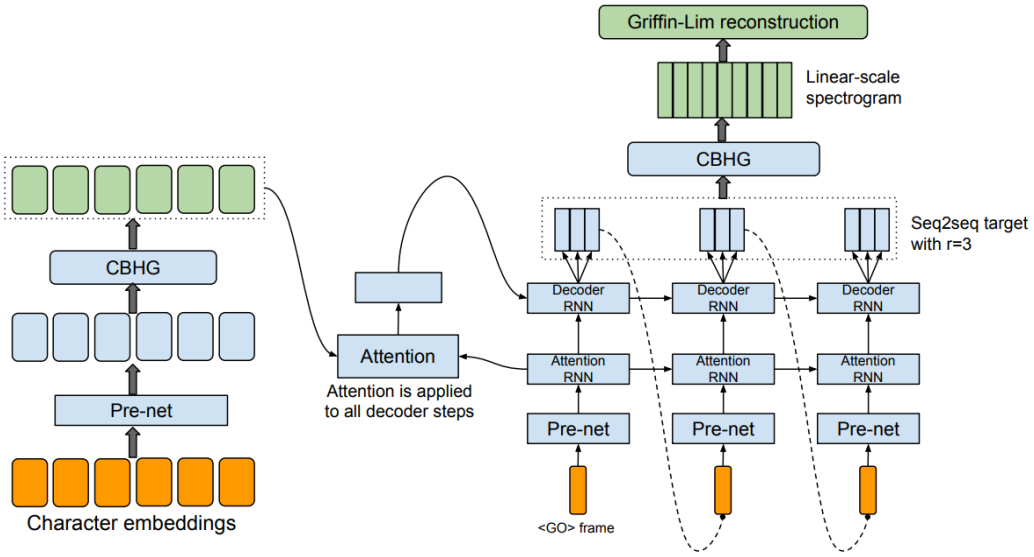
O modelo converte diversos parâmetros textuais como caracteres, fonemas e ênfases para diferentes parâmetros do vocoder como espectrogramas na banda mel, espectrogramas na escala linear e magnitude log, frequência fundamental F0, envelope espectral e parâmetros de aperiodicidade para servirem de entradas para o modelo de síntese do sinal de fala que pode ser adaptado para diferentes sintetizadores como o WaveNet, Griffin-Lim ou WORLD.

Outros melhoramentos do modelo foram no seu mecanismo de atenção que utiliza atenção monotônica durante o treinamento, resolvendo alguns problemas como a repetição, saltos de alguns trechos de texto ou enganos de pronúncia (PING et al., 2017).

2.4.3.2 *Tacotron*

Wang et al. (2017) apresenta o Tacotron como um modelo de conversão texto-fala da Google que sintetiza fala diretamente a partir de pares de áudio com suas respectivas transcrições. Sua arquitetura, vista na Figura 8, é baseada nos modelos Seq2Seq que incluem um codificador, um decodificador com modelo de atenção e uma rede de pós-processamento. O modelo tenta prever espectrogramas mel na saída antes de ser projetado para um espectrograma linear é reconstruído para uma forma de onda utilizando o vocoder de fase Griffin-Lim. São utilizados módulos CBHG, que consistem de filtros convolucionais 1-D, *highway networks* e Gated Recurrent Unit (GRU) bidirecional para a extração das representações das sequências, sendo originalmente criados para modelos de tradução neurais.

Figura 8 - Arquitetura do modelo do Tacotron.



Fonte: Wang et al. (2017)

codificador extrai representações sequenciais do texto, a partir de sequências de caracteres na entrada, representados por um vetor *one-hot* que passa por uma camada de *embedding* e resulta em um vetor contínuo. Esse vetor passa pela pré-rede, que aplica transformações não lineares nos *embeddings,* afunilando as informações e ajudando na convergência e generalização do modelo. A saída dessa pré-rede passa pelo módulo CBHG que se transforma na representação final do codificador utilizada pelo módulo de atenção.

No decodificador é utilizado um modelo de atenção tangente hiperbólica (tanh) *content-based* que está conectado a cada passo do decodificador e é concatenada na saída do vetor de contexto do codificador com a saída das células GRU de atenção. O decodificador é formado por uma pilha de GRUs com conexões residuais verticais para aumentar a convergência. A saída passa por uma camada totalmente conectada para prever os espectrogramas e, da mesma forma que no codificador, passa por uma pré-rede sendo então realimentada para o próximo passo do decodificador.

Wang et al. (2017) propôs um truque utilizado para tentar prever múltiplos quadros de saída por passo do decodificador, acelerando o tempo de convergência com alinhamentos mais rápidos e estáveis do modelo de atenção. A ideia por trás seria que quadros vizinhos são correlacionados e cada caractere geralmente corresponde a vários quadros do espectrograma.

Na saída tenta-se prever um espectrograma em escala mel de 80 bandas que após passar pela rede de pós processamento e ser transformado para um espectrograma linear, é processado pelo algoritmo de Griffin-Lim para sintetizar o sinal de fala (WANG et al., 2017).

2.4.3.3 *Tacotron-2*

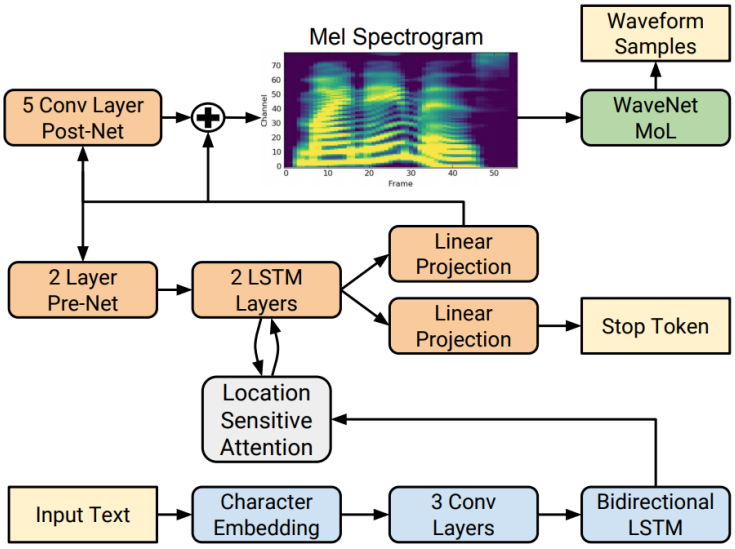
Pesquisadores da Google, Shen et al. (2018), introduziram no artigo "*Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions*" a arquitetura do Tacotron 2 que propôs uma abordagem totalmente neural da combinação das melhores características do Tacotron com o Wavenet. Obtendo resultados que são difíceis de distinguir da fala de um humano real.

A arquitetura do modelo é composta por 2 componentes principais. Uma rede de previsão de atributos na forma de espectrogramas mel, utilizando uma rede *sequence-to-sequence* recorrente com atenção. E uma versão modificada do WaveNet que em vez dos atributos linguísticos derivados dos textos e o log da frequência fundamental como feature acústica, a versão modificada utiliza quadros de espectrograma mel como entrada do modelo. Essa representação intermediária entre os dois componentes do modelo é mais fácil de ser computada do que as amostras em forma de onda e também mais fácil para treinar os modelos utilizando perda de erro quadrático.

Na Figura 9 pode ser vista a arquitetura do Tacotron 2. Sua rede é composta por um codificador e um decodificador com atenção, da mesma forma que no Tacotron, porém

algumas mudanças tornaram seus blocos mais simples. No codificador, os módulos CBHG não são mais utilizados na arquitetura e foram substituídos por 3 camadas de redes convolucionais e uma camada LSTM bidirecional. As camadas convolucionais funcionam para modelar o contexto de longo termo de forma parecida com N-gramas, um modelo de linguagem comum em tarefas de processamento natural de linguagem que auxilia na estimação da probabilidade de uma palavra baseada em N palavras próximas na sequência (JURAFSKY, D.; MARTIN, J. H, 2009).

Figura 9 - Arquitetura do Tacotron 2.



Fonte: Shen et al. (2018)

Outra mudança foi a utilização de células LSTM em vez das GRU para as redes recorrentes do modelo *sequence-to-sequence*. Houve também alterações no modelo de atenção utilizado, agora com um modelo de atenção *location-sensitive* (CHOROWSKI et al., 2015), em que o mecanismo de atenção aditiva é estendido para utilizar pesos de atenção de passos anteriores cumulativamente como atributos adicionais.

O decodificador é uma rede recorrente auto regressiva que prevê espectrogramas mel um quadro por vez. As predições do passo anterior passam por uma pré-rede com 2 camadas densas totalmente conectadas, funcionando como um gargalo de informações e sendo essencial para o treinamento do módulo de atenção. A saída é concatenada com o vetor de contexto da atenção, passando por 2 camadas de LSTM unidirecionais, diferentemente da camada de GRU do Tacotron. A concatenação da saída do LSTM com o vetor de contexto de atenção é projetada através de uma transformação linear para prever o quadro de

espectrograma mel alvo. Esse resultado passa por uma rede de pós processamento de cinco camadas convolucionais, com a finalidade de prever o resíduo, que é adicionado com a predição do espectrograma para tentar melhorar a reconstrução geral. A predição final do espectrograma mel é utilizada como entrada para o treinamento do modelo WaveNet para gerar os sinais de forma de onda da fala (SHEN et al., 2018).

2.5 VOCODERS

Os vocoders são um sistema de processamento de sinal feito para sintetizar a forma de onda da fala a partir de atributos representativos. Parâmetros clássicos de vocoders são motivados pelos modelos de produção de fala, usando atributos como: frequência fundamental, envelope espectral, entre outros.

Caso as representações de atributos sejam espectrogramas de magnitude, vocoders de fase como o Griffin-Lim são amplamente utilizados. Esses tipos de vocoders conseguem reconstruir a fase. Com a aplicação de uma *Inverse Short Time Fourier Transform* (ISTFT) é possível recuperar o sinal no domínio do tempo.

Os vocoders neurais tentam prever as formas de onda da fala a partir de amostras de sinais passados, podendo ser controlado pelas representações de atributos padrões usando modelos autorregressivos para prever a distribuição de probabilidade das amostras de forma de onda atuais baseadas nas antigas como no caso do WaveNet.

Assim como os modelos acústicos, muitos trabalhos vêm sendo publicados com arquiteturas inovadoras para os vocoders permitindo o treinamento com menos recursos computacionais e mantendo a qualidade de áudio. Os vocoders podem ser divididos em alguns grupos:

1) Autorregressivos (ex: WaveNet (Oord et al., 2016));

2) Baseados em Flow (ex: WaveGlow (PRENGER et al., 2019));

3) Baseados em GANs (MelGAN (KUMAR et al., 2019)).

Nas subseções seguintes serão apresentados os vocoders Griffin-Lim e WaveNet pelo fato do Tacotron 2 utilizar o WaveNet e neste trabalho ter sido utilizado o Griffin-Lim.

2.5.1 Griffin-Lim

O algoritmo Griffin-Lim (GLA), nomeado a partir de seus autores Griffin, Lim, em 1984, no artigo “Signal Estimation from Modified Short-Time Fourier Transform”, resolveu de forma computacionalmente simples a estimação do sinal, a partir da magnitude de sua STFT, sem a informação da fase. Portanto, sendo capaz de sintetizar sinais de fala em tempo real.

O algoritmo iterativo é obtido a partir da minimização do erro quadrático médio entre a magnitude da STFT do sinal estimado e da magnitude da STFT modificada. É demonstrado teoricamente que o algoritmo converge e diminui o erro quadrático médio em cada iteração, porém não é garantido que irá convergir para o mínimo global devido à inicialização aleatória da informação da fase.

Os passos para implementar o GLA são os seguintes:

1) Uma matriz com a magnitude da STFT é inicializada com valores aleatórios para a sua fase.
2) É feita a estimação do sinal através da ISTFT dessa STFT modificada com informação de fase aleatória.
3) Calcula-se a STFT desse novo sinal estimado que agora contém um pouco da informação da fase.
4) Com essa nova STFT, muda-se a magnitude da STFT pela magnitude original, dessa forma mudando apenas a informação da estimação da fase.
5) Cálculo do MSE entre a magnitude da STFT original com a magnitude da STFT modificada
6) Volta ao passo 2.

Após algumas iterações do GLA, o sinal estimado consegue recuperar a informação da fase de forma rápida e simples, porém a qualidade da sua reconstrução é baixa, se comparada a outros algoritmos mais modernos como o WaveNet.

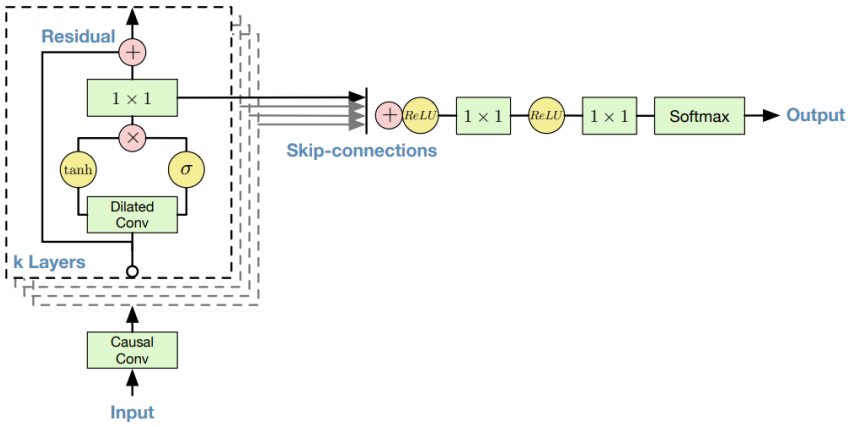
2.5.2 WaveNet

Oord et al. (2016) apresentou o WaveNet como sendo um modelo totalmente probabilístico e autorregressivo do grupo DeepMind da Google, baseado no PixelCNN e

capaz de produzir áudio muito similar à fala humana. Na figura 10, pode ser vista a arquitetura do modelo.

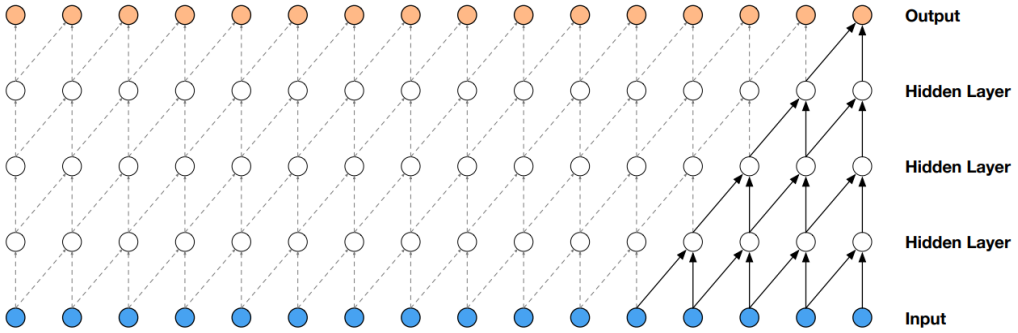
Nesse modelo generativo, cada amostra é condicionada na amostra anterior e cada probabilidade condicional é modelada por uma pilha de camadas convolucionais causais. Uma visualização do funcionamento pode ser visto na Figura 11.

Figura 10 - Arquitetura do modelo WaveNet.



Fonte: Oord et al. (2016)

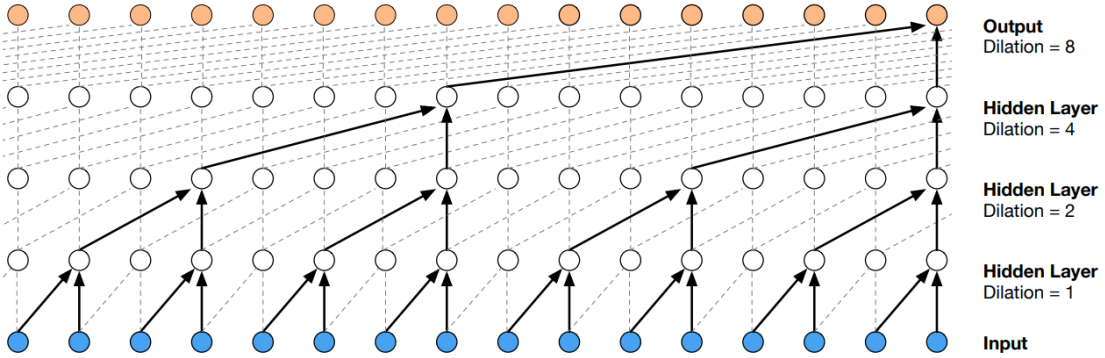
Figura 11 - Visualização de uma pilha de camadas convolucionais causais.



Fonte: Oord et al. (2016)

O uso de convoluções causais na arquitetura garante que o modelo não viole o ordenamento dos dados, sendo que cada amostra é realimentada na rede para prever a próxima. Para evitar o uso de muitas camadas para aumentar o campo receptivo, são utilizadas convoluções dilatadas, exemplificadas na Figura 12 (OORD et al., 2016).

Figura 12 - Visualização de uma pilha de camadas convolucionais causais dilatadas.

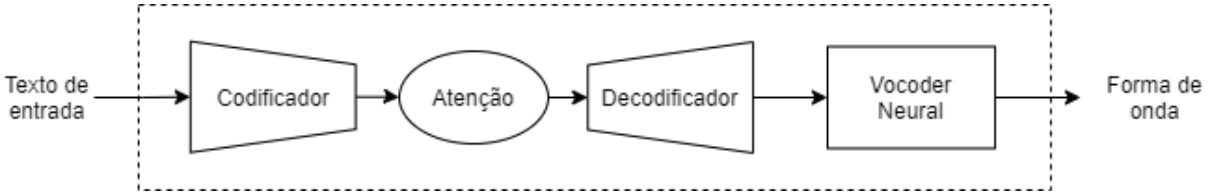


Fonte: Oord et al. (2016)

2.6 COMPARAÇÃO ENTRE O MODELO TRADICIONAL E O *END-TO-END*

Um sistema TTS típico consiste de um componente de processamento linguístico, um modelo acústico e um vocoder, como foi visto na Figura 4. Esses componentes são treinados independentemente e a elaboração de cada um é um processo trabalhoso que depende do conhecimento extenso dessa área de conhecimento, sendo que os erros de cada componente podem se propagar dentro da *pipeline*.

Figura 13 - Fluxo de um sistema TTS *end-to-end* neural estilo Tacotron-2.



Fonte: Elaborado pelo autor.

Para contornar esses problemas, os sistemas TTS *end-to-end* combinam a geração desses componentes em apenas um modelo. Um exemplo de um sistema TTS *end-to-end* pode ser visto na Figura 13. Fazendo isso, diminui-se a necessidade de entender como processar cada etapa e torna-se possível o treinamento usando apenas pares de fala e transcrição como entrada, com o mínimo de trabalho humano para anotação, alinhamento de fonema, etc. Outra vantagem está na facilidade de condicionar o modelo *end-to-end* para vozes novas, linguagens diferentes ou até mesmo para adicionar sentimentos diferentes na fala. A adaptação para novos dados também se torna mais fácil em relação aos modelos concatenativos, podendo ser treinados modelos com uma quantidade muito maior de dados, incluindo dados ruidosos e expressivos (WANG et al., 2017).

3 DESENVOLVIMENTO

Para a realização do experimento foi utilizado o modelo de síntese de espectrogramas mel Tacotron-2 sem o vocoder WaveNet modificado devido ao recurso necessário para treinar o vocoder, utilizando o vocoder Griffin-Lim. Muitas implementações de código livre desse modelo já existem e o propósito deste trabalho é avaliar a síntese de fala do modelo para o português brasileiro. Para tanto, foram utilizados o modelo do Rayhane-Mamah[1] e o TensorFlowTTS[2] do TensorSpeech.

Dois modelos com estratégias diferentes foram adotados para o treinamento e desenvolvimento do projeto. Primeiramente, avaliou-se a possibilidade de treinar um modelo do zero por 70 mil passos diretamente com o conjunto de dados em português. Como existem modelos pré-treinados para o inglês e com os pesos já definidos, tentou-se utilizar o conjunto de dados em português e treinar por mais 30 mil passos esses modelos pré-treinados. Esse tipo de experimento se chama *fine tuning,* uma das abordagens do *transfer learning* para adaptação de novos dados.

3.1 CONJUNTO DE DADOS

Uma das etapas fundamentais para qualquer projeto de aprendizado de máquina é a consolidação de um conjunto de dados extenso e formatado que consiga descrever o fenômeno a ser descrito pelo modelo da melhor forma possível.

Para sistemas de conversão texto-fala é necessário um vasto corpus que contenha textos naturais representativos da linguagem a ser sintetizada que, idealmente, sejam balanceados tanto foneticamente quanto prosodicamente, com qualidade de estúdio para a fala e livre de ruído. A construção desse corpus não é um trabalho trivial e existe uma área dentro da linguística específica para a coleta e análise de corpus que é a linguística de corpus, uma síntese do seu histórico e problemática pode ser vista no trabalho de Sardinha (2000).

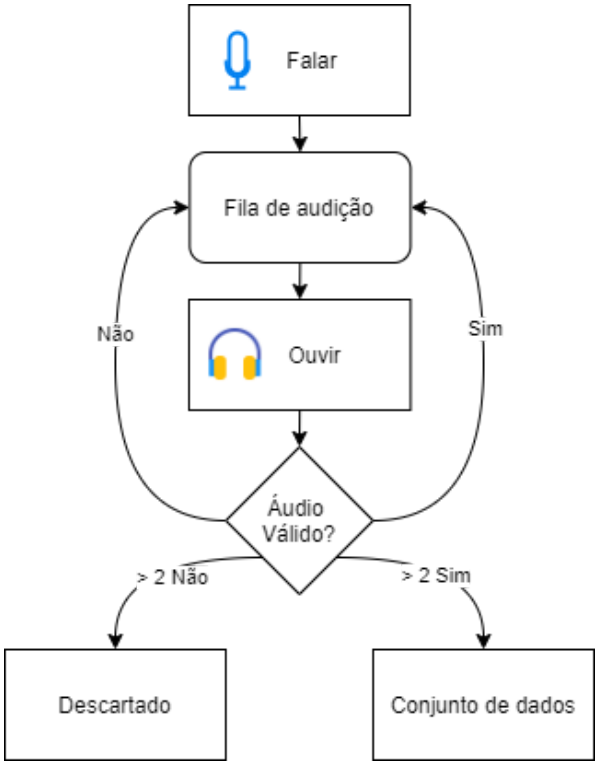
Diferentemente de linguagens como o inglês que possui inúmeros conjuntos de dados específicos para o desenvolvimento de sistemas TTS, o mesmo não é visto para o português brasileiro. Porém, a empresa Mozilla desenvolveu o Common Voice (ARDILA, R.

[1] Repositório do Github: https://github.com/Rayhane-mamah/Tacotron-2
[2] Repositório do Github: https://github.com/TensorSpeech/TensorFlowTTS

et al., 2020), um imenso corpus de fala multi-idioma colaborativo e *open source*. Com o objetivo de abrir e descentralizar as tecnologias de voz, o projeto Common Voice disponibiliza em seu site conjunto de dados para 60 idiomas com 7.335 horas validadas, sendo que para o português brasileiro são 50 horas de 1.120 falantes diferentes.

O método de geração desses conjuntos de dados consiste no fluxo da Figura 14. Os colaboradores do projeto acessam o site e possuem duas opções de contribuição: falar ou ouvir. Ao escolher a opção de falar, aparecem frases coletadas do idioma escolhido e um botão para iniciar a gravação da sua voz. Todo arquivo de voz gerado vai para uma fila de audição que fica disponível para as pessoas que desejem contribuir ouvindo e validando a fala em relação ao que estava escrito. O processo de validação por um terceiro é feito por várias pessoas diferentes e se mais de duas pessoas rejeitarem o áudio, ele será descartado. Se duas ou mais pessoas validarem o áudio, ele entra para o conjunto de dados da língua específica.

Figura 14 - Fluxo de validação de áudios para o conjunto de dados do Common Voice.

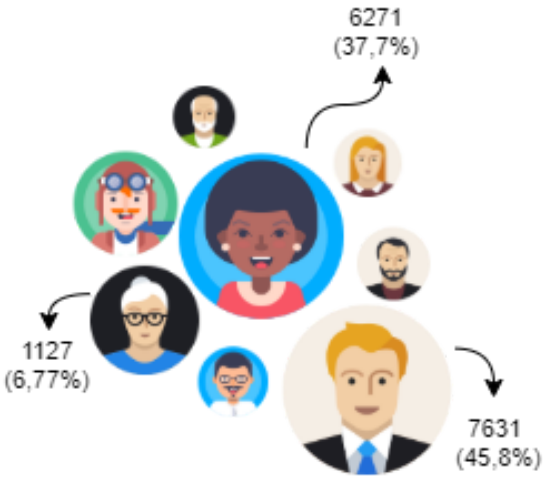


Fonte: Adaptado de Common Voice (2021).

A abordagem do modelo TTS a ser utilizado não consegue lidar com falantes diferentes e portanto foi necessário verificar se existia algum falante dentro do conjunto que possuísse uma quantidade significativa de áudios. Analisando os dados disponíveis, existe um

arquivo estruturado de validação de cada amostra do conjunto de dados. Dentro desse arquivo cada linha representa uma amostra contendo informações do id único anonimizado da pessoa que gravou a voz, nome do arquivo mp3, transcrição do áudio, votos positivos e negativos feitos por colaboradores da página que ouviram os áudios com fins de validar e informações de idade e gênero.

Figura 15 - Proporção de falantes no conjunto de dados.



Fonte: Elaborado pelo autor.

Foi feita uma listagem dos identificadores únicos com maiores números de áudios gravados, sendo que o número um da lista possui 7631 arquivos gravados dentro do conjunto de dados, resultando em cerca de 6 horas de áudio. Conforme a Figura 15, ele compõe cerca de 45,8% do conjunto de dados e os metadados disponíveis descrevem que são de um homem com cerca de 30 anos.

3.2 PRÉ-PROCESSAMENTO DO CONJUNTO DE DADOS

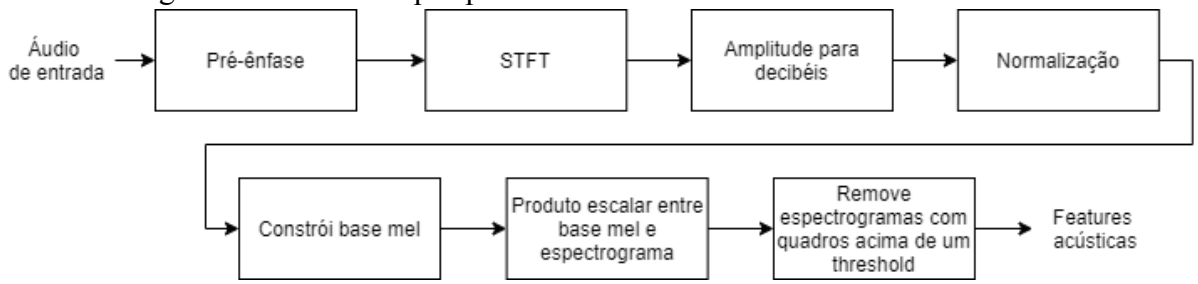
Com os dados do falante selecionados, foi feita uma análise desse subconjunto de dados, verificando os áudios das frases que possuíam um ou mais votos negativos durante sua validação pelos ouvintes. Em algumas frases o falante alterou uma ou mais palavras e portanto foi necessário corrigir as transcrições quando o sentido da frase continuava o mesmo e excluir quando se perdia.

Foi necessário deixar os arquivos de áudio em um padrão que os scripts do modelo *open-source* pudessem utilizá-los. Para adequá-los, a biblioteca de leitura dos arquivos de

áudio em Python precisa que estejam em formato wav e portanto foi feito um código que utilizando uma extensão em Python do ffmpeg conseguisse transformar cada arquivo mp3 para wav. Em relação aos dados de texto, foi criado um novo arquivo metadata.csv em que para cada transcrição do áudio houvesse o nome do arquivo de referência.

Antes de começar o treinamento, é necessária a extração dos atributos relevantes para que o modelo consiga encontrar correlações dos dados com as saídas. Um dos atributos acústicos mais utilizados nas últimas décadas são os espectrogramas mel, que são a representação do sinal de fala na frequência utilizando a escala mel. Os passos tomados para a transformação do sinal de áudio para o espectrograma mel são: pré-ênfase do sinal para amplificar as altas frequências e balancear o espectro de frequência, calcular a STFT de uma pequena janela pré-definida, aplicar a escala mel no espectrograma resultante e transformar a magnitude em escala log (decibel). Como o tamanho do espectrograma afeta na memória necessária para armazenar em GPU, para reduzir o processamento, são removidos os espectrogramas com quadros acima de um valor ajustado de acordo com o poder de processamento computacional. Todo esse processo é visto na Figura 16.

Figura 16 - Passos de pré-processamento dos dados acústicos de entrada.



Fonte: Elaborado pelo autor.

Para alimentar as camadas de *word embeddings*[3] do modelo, é feito antes um pré-processamento das transcrições de áudio, transformando os símbolos (letras) em sequências de índices numéricos. Para fazer isso, é definido o alfabeto do idioma e criado um grande dicionário no qual cada índice representa uma letra. Após isso, a frase é convertida em uma sequência de índices. No Quadro 1, temos uma síntese dos resultados do pré-processamento.

[3] *Word embedding* é o nome coletivo de um conjunto de modelagens linguísticas e técnicas de aprendizados de atributos da área de processamento natural de linguagem no qual as palavras de um vocabulário são mapeadas em vetores de números reais (ALMEIDA; GERALDO, 2019).

Quadro 1 - Descrição do conjunto de dados.

Frases	7631
Quadros mel totais	1771815
Amostras de áudio totais	487249125
Horas de áudio	~6h 8min
Comprimento máximo de caracteres	106
Comprimento máximo de quadros mel	645
Comprimento máximo de *timesteps* de áudio	177375
Tempo máximo de um áudio	~8s

Fonte: Elaborado pelo autor.

3.3 TREINAMENTO DO MODELO DO ZERO

Tendo feito o pré-processamento dos dados, foi realizado o treinamento utilizando uma das principais implementações do Tacotron-2 de código aberto disponível, o Tacotron-2 do Rayhane Mama sem alterações no hiperparâmetros do modelo, mantendo *batch size* de 32 e sem fator de redução. Foi utilizada a ferramenta Google Colab que disponibiliza um ambiente de notebook python na nuvem com GPUs do Google Cloud (K80, P4, T4, P100).

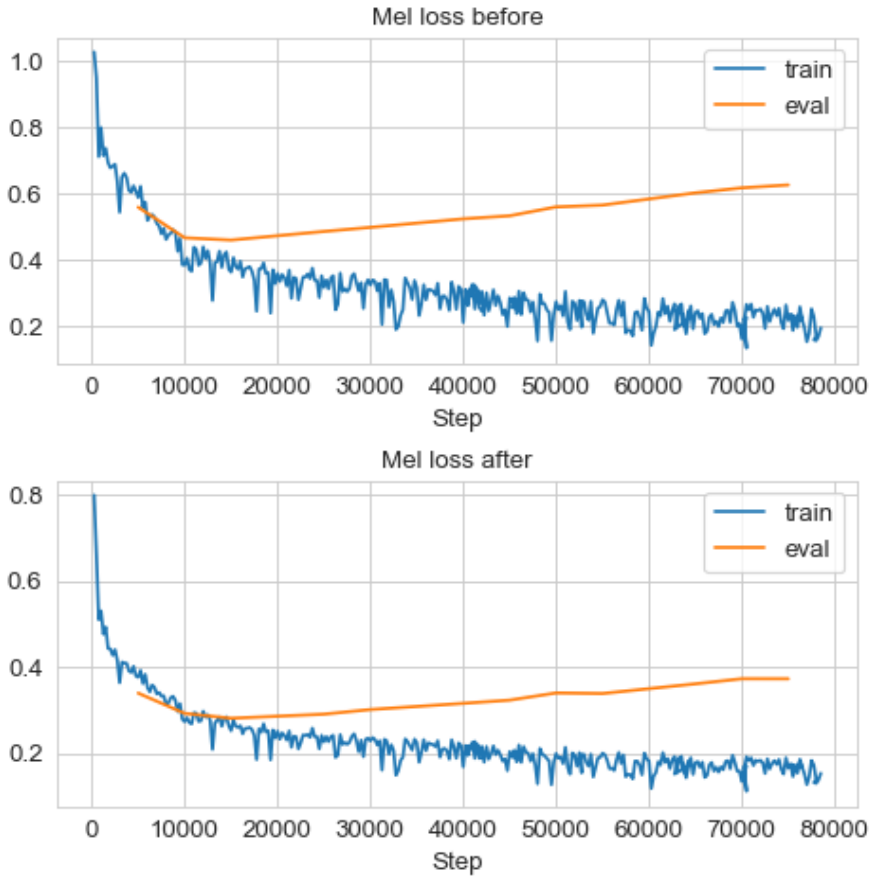
Como o Google Colab é um recurso gratuito, existe um limite de 12 horas contínuas antes que a sessão do notebook termine. Como o modelo do Tacotron-2 exige horas e até dias para conseguir treinar o suficiente para a convergência do modelo, foi utilizado uma integração com o Google Drive que salva *checkpoints* que podem ser recuperados mesmo se a sessão for terminada, é possível retomar o treinamento em outro momento.

3.3.1 Métricas do modelo treinado do zero

A soma dos erros quadráticos médios de antes e depois da rede de pós processamento convolucional do modelo é utilizada como métrica para estabelecer a convergência do modelo durante a etapa de treinamento. Na Figura 17 pode-se observar os resultados obtidos tanto para o conjunto de treinamento quanto para o de avaliação. A influência da rede posterior é determinada pela diferença dos valores entre o gráfico superior e o inferior. Observa-se diminuição da métrica do erro quadrático médio pela metade após passar por ela.

Nota-se que o erro quadrático diminui constantemente ao longo do treinamento, mas na avaliação começa a subir a partir dos 10.000 passos, sinalizando um *overfitting* do modelo. O erro quadrático médio é uma boa métrica para auxiliar o modelo para a convergência durante o treinamento, porém não é a única forma para avaliar a qualidade da fala sintetizada. É importante que durante o treinamento possa também ser realizada a comparação do espectrograma original com o resultante da predição. Adicionalmente, o gráfico de alinhamento do modelo de atenção também é muito importante para entender como o modelo está progredindo.

Figura 17 - Curvas de perda no treinamento e validação antes e depois da rede de pós-processamento da implementação do Rayhane Mamah.

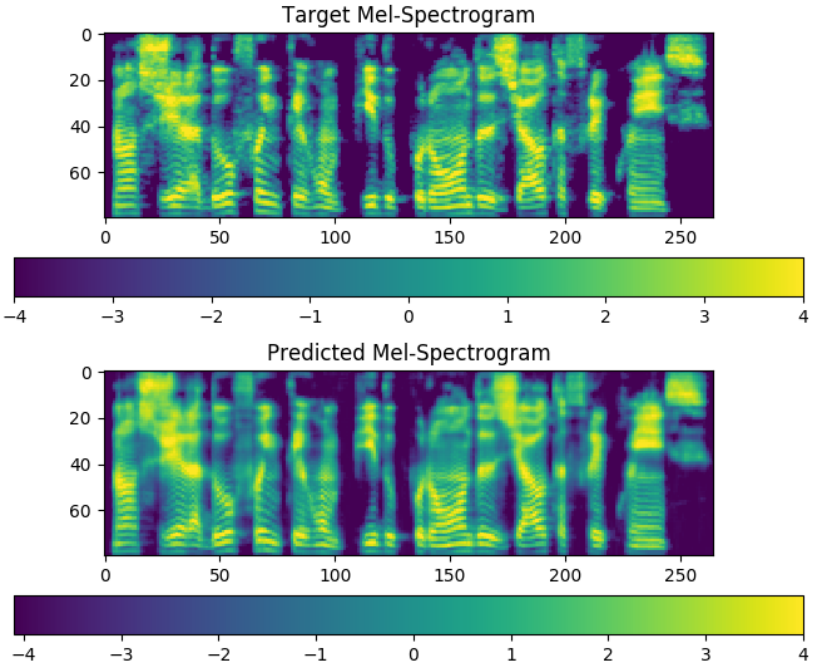


Fonte: Elaborado pelo autor.

Na Figura 18 temos a comparação dos espectrogramas e ao olhar com detalhe nos gráficos das predições (antes e depois), notam-se menos detalhes e uma aparência mais suave, resultado da otimização a partir da função de perda do erro quadrático médio (SHEN et al., 2018).

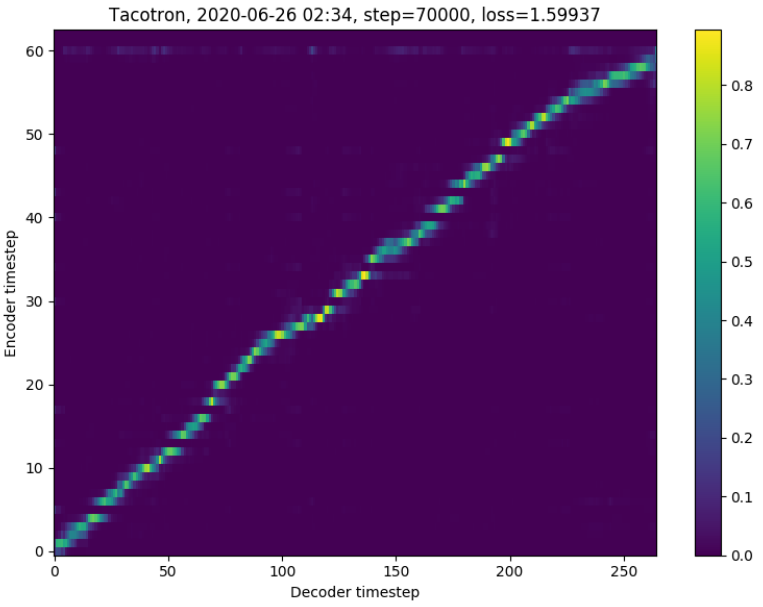
O gráfico da Figura 19 mostra o alinhamento entre os caracteres e trechos de espectrograma mel, o modelo de atenção tenta fazer com que o decodificador “preste atenção” nos vetores de representação corretos para aquele caractere e não se perca em frases mais longas. Um indicativo de um bom gráfico de alinhamento seria um com uma linha diagonal.

Figura 18 - Comparativo do espectrograma mel sintetizado em relação ao real.



Fonte: Elaborado pelo autor.

Figura 19 - Gráfico de alinhamento do modelo de atenção do modelo treinado do zero.



Fonte: Elaborado pelo autor.

3.4 *FINETUNE* A PARTIR DE UM MODELO TREINADO EM INGLÊS

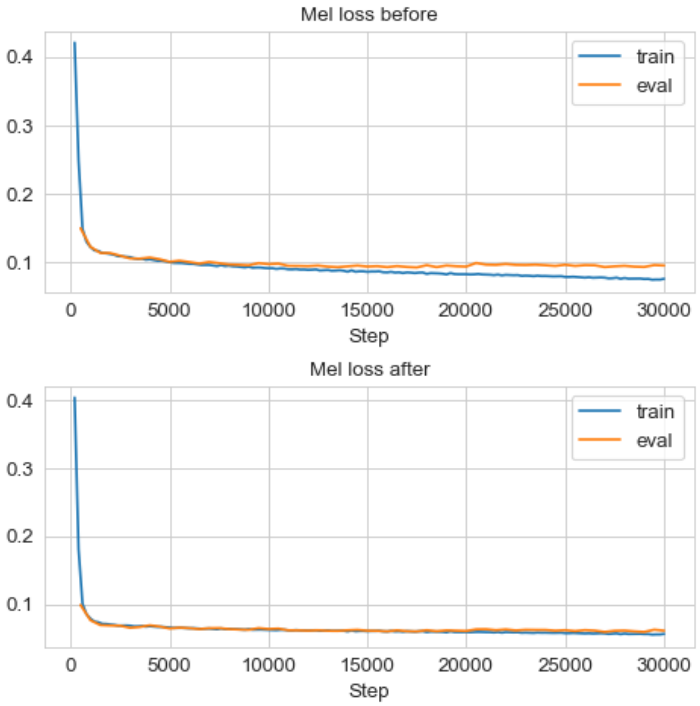
A área de síntese de fala está muito mais desenvolvida em outros idiomas, como o inglês, possuindo muitos conjuntos de dados com qualidade superior ao que temos disponível em português. Com isso em mente, seria interessante a possibilidade de utilizar um recurso chamado *Transfer Learning* para usar como base um modelo treinado em inglês que tenha bons resultados e com ele treinar mais um pouco com o conjunto de dados menor em português. Foi utilizado um modelo pré-treinado com o conjunto de dados LJSpeech em 65 mil passos com o TensorflowTTS. Os hiperparâmetros do modelo não foram alterados, por exceção do fator de redução que foi colocado em 2 e o tamanho da camada de embedding que foi alterada conforme o tamanho do alfabeto português brasileiro que foi utilizado para treiná-lo.

3.4.1 Métricas do modelo *finetuned*

Com o treinamento do modelo *finetuned* e avaliando as curvas de perda, pode se observar na Figura 20 que a métrica de erro quadrático médio obteve valores próximos de 0,05 enquanto que no modelo treinado do zero estava próximo de 0,6. Além disso, analisando-se a curva característica do *overfitting* verifica-se que os resultados de treinamento e validação se encontram muito mais controlados. Porém, como visto anteriormente, o erro quadrático médio é uma métrica adequada para a convergência do modelo, mas não deve ser utilizada de forma isolada. A síntese de fala é um problema muito complicado de se avaliar com métricas objetivas e ainda mais por apenas uma. Portanto deve ser analisado, adicionalmente, os espectrogramas , os gráficos de alinhamento e o áudio gerado.

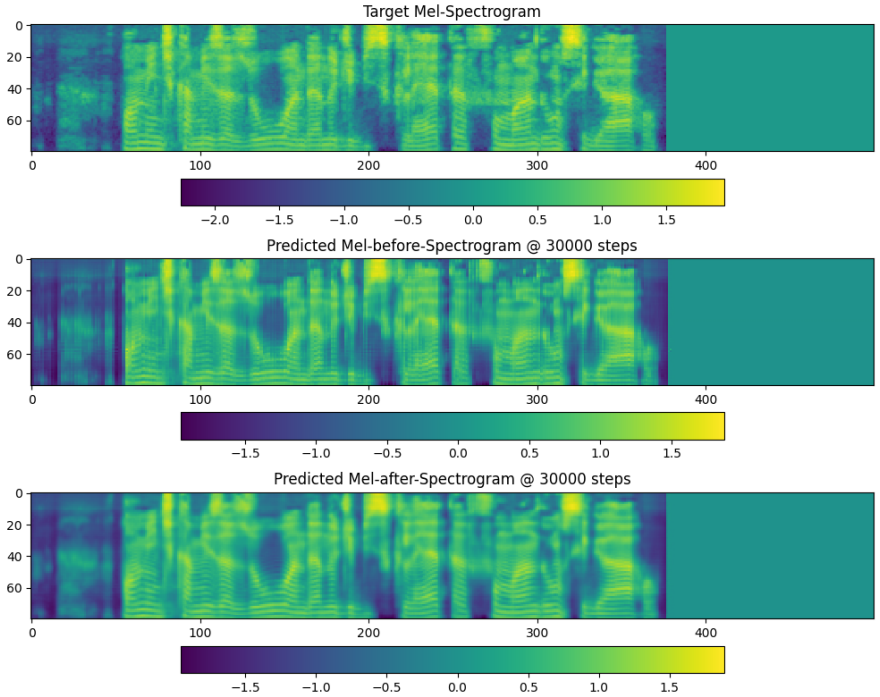
Na implementação do Tacotron-2 do TensorflowTTS, foi utilizado tamanho fixo de caracteres e de quadros mel durante o treinamento do modelo, adicionando *padding* (preenchimento por zero) ao final para preencher o restante no decodificador e codificador do modelo. Isso apenas ocorre durante o treinamento, pois verificou-se que na implementação do modelo utilizando o tamanho fixo resulta um aumento de 2x na velocidade de treinamento por permitir paralelizar as operações. Durante a inferência são utilizados tamanhos dinâmicos, treinados a partir da *flag* de *token* de parada que vem da projeção linear após cada passo do decodificador, como visto na arquitetura do Tacotron-2.

Figura 20 - Curvas de perda no treinamento e validação antes e depois da rede de pós-processamento do TensorflowTTS.



Fonte: Elaborado pelo autor.

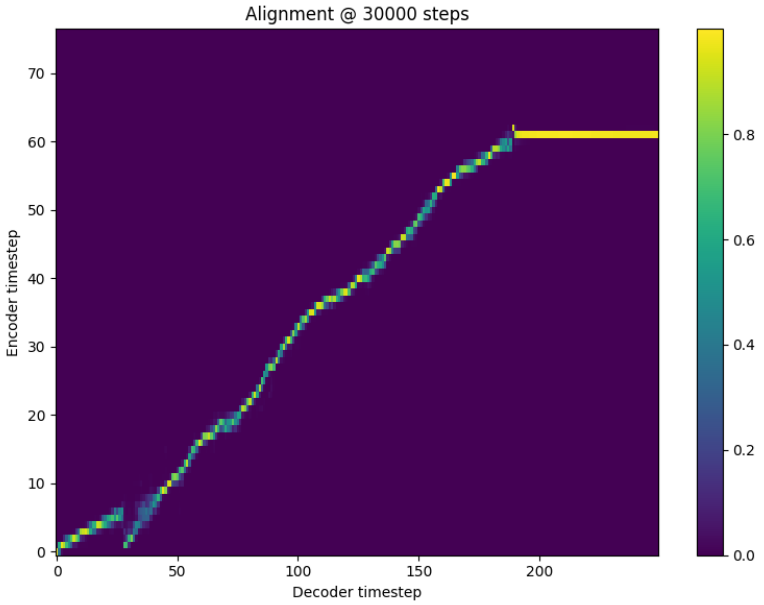
Figura 21 - Espectrogramas mel de avaliação do treinamento do modelo *finetuned*.



Fonte: Elaborado pelo autor.

Na Figura 22 temos o alinhamento entre a entrada e saída com a linha diagonal característica, porém com um desalinhamento no início da frase. Assim como o espectrograma mel da Figura 21, foi utilizado um tamanho fixo para o treinamento que resultou no *padding* que pode ser visto na reta horizontal amarela no final da imagem.

Figura 22 - Alinhamento de avaliação do treinamento do modelo *finetuned*.



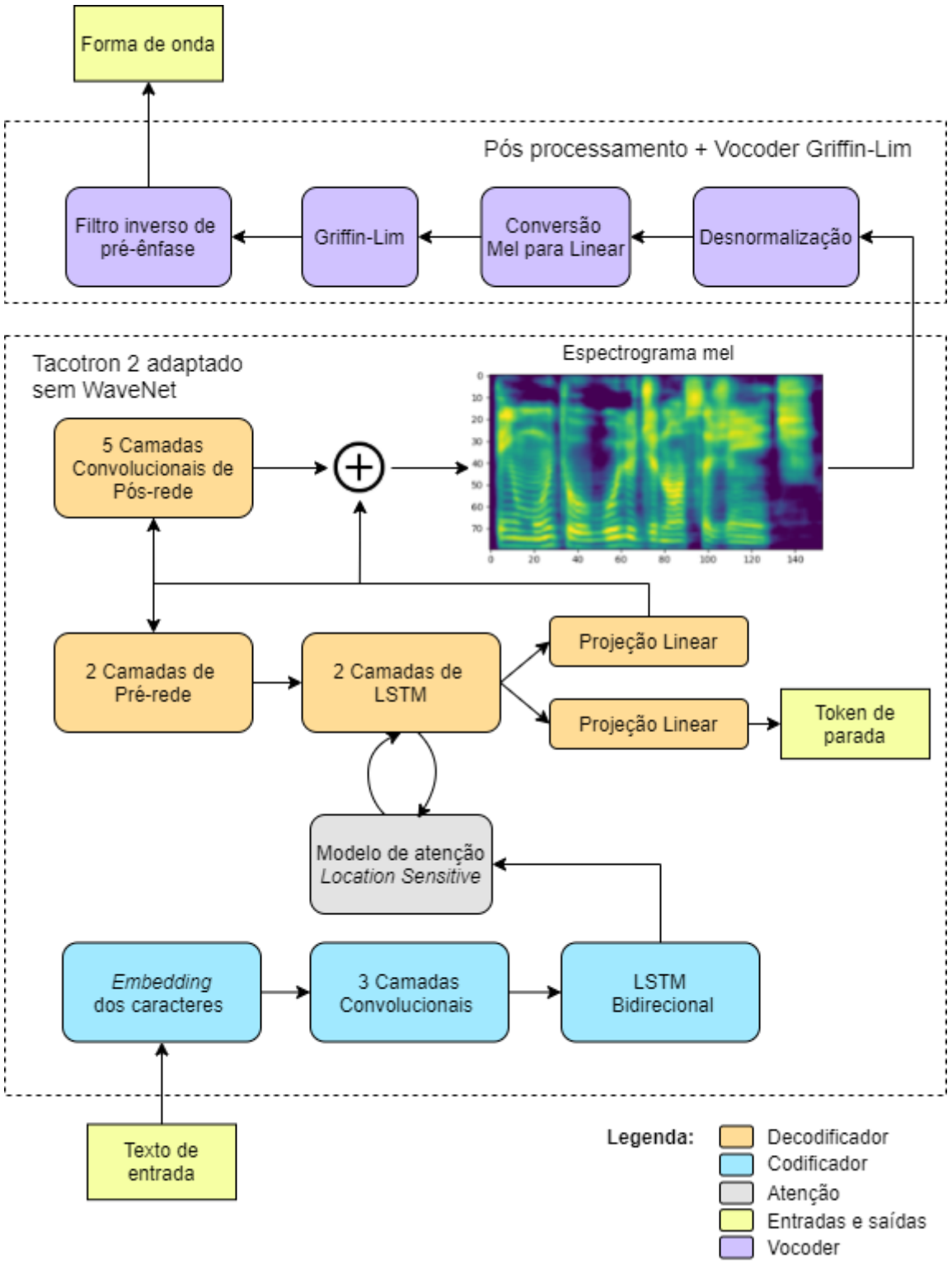
Fonte: Elaborado pelo autor.

3.5 INFERÊNCIA

Para a inferência da forma de onda do sinal de fala, foi feita uma adaptação na arquitetura do Tacotron 2 removendo a etapa do treinamento do vocoder neural WaveNet e utilizando o vocoder de fase Griffin-Lim que é computacionalmente mais leve, porém com qualidade do áudio sintetizado pior.

Com a remoção do WaveNet da arquitetura do Tacotron 2, foi preciso utilizar o espectrograma mel da saída do decodificador e passar por um pós-processamento para desfazer as etapas descritas na seção 3.2. Primeiro é feita a desnormalização do espectrograma mel, em seguida a conversão para um espectrograma linear seguida da estimação da fase com o vocoder Griffin-Lim e então realizar a inversão do filtro de pré-ênfase, assim resultando na forma de onda de saída esperada. A arquitetura final do modelo utilizado neste trabalho ficou de acordo com a apresentada na Figura 23.

Figura 23 - Arquitetura adaptada do Tacotron 2 utilizada.



Fonte: Adaptado de Shen et al. (2018)

4 AVALIAÇÃO DOS MODELOS

Quando se fala na avaliação de desempenho de sistemas com sinais de fala existem métricas objetivas e subjetivas para quantificação em termos de qualidade e inteligibilidade. Esses dois atributos, entre muitos outros do sinal de fala, são diferentes e não são equivalentes.

No livro *Speech Enhancement Theory and Practice* de Loizou (2013), a qualidade da fala é descrita como sendo altamente subjetiva em natureza e difícil de se avaliar com alta confiança pelo fato de vários ouvintes terem opiniões diferentes de como metrificar a qualidade, causando alta variabilidade dos resultados. Os testes subjetivos de análise de qualidade são demorados e precisam de um conjunto de avaliadores treinados. O teste mais utilizado na avaliação subjetiva de qualidade é o *Mean Opinion Score* (MOS).

A inteligibilidade, diferentemente da qualidade, é descrita como não sendo subjetiva e pode ser mensurada pedindo ao ouvinte para contar ou identificar as palavras e fonemas. Loizou (2013) descreve que para desenvolver testes de inteligibilidade confiáveis é necessário que atendam as seguintes considerações:

- Possuem uma adequada representação dos principais fonemas da lista de testes. Idealmente, a lista deve apresentar frequência relativa de fonemas refletindo a distribuição de fonemas normalmente utilizados em fala normal;
- Dificuldade das listas de teste iguais, no caso de testes extensivos usando o mesmo material. É preferível que o teste tenha várias listas com 10 frases ou 50 palavras monossilábicas;
- Controle da informação contextual. Palavras presentes em frases são mais inteligíveis do que frases apresentadas isoladas devido ao fato que o ouvinte usa o contexto para identificar as palavras numa frase. Por isso é necessário controlar a quantidade de informação dentro de uma frase para que sejam igualmente inteligíveis.

Em um trabalho correlato de Casanova (2020), o autor apresenta um novo conjunto de falas para ser utilizado para a síntese de fala em português. Em sua composição, foram utilizadas frases foneticamente balanceadas geradas na tese de mestrado em linguística de Seara (1994). Com o objetivo de ser aplicado em estudos de processamento de fala, Seara fez um estudo estatístico dos fonemas do português falado em Florianópolis e construiu 20 listas

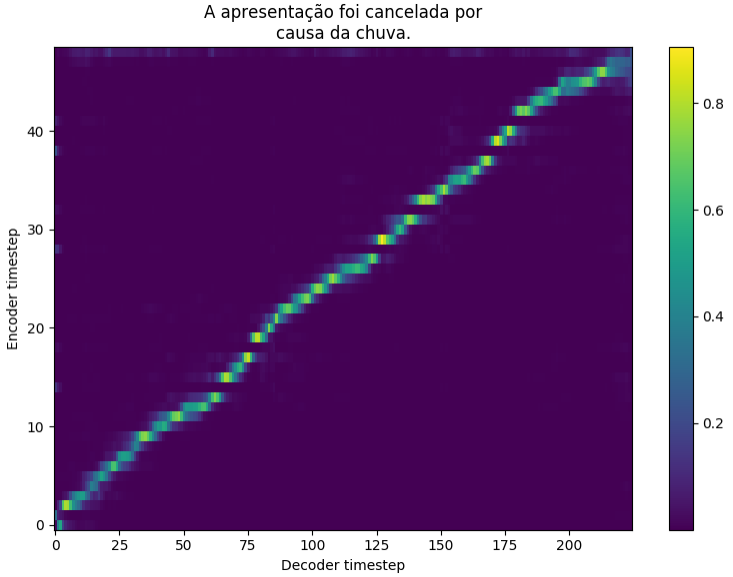
de 10 frases foneticamente balanceadas. As frases presentes compreendem os três pontos descritos anteriormente para serem utilizados para testes de inteligibilidade.

A metodologia de avaliação da inteligibilidade das falas sintetizadas pelos modelos se baseou na síntese de espectrogramas mel das 200 frases extraídas da dissertação de Seara (1994) que estavam fora do treinamento, análise dos gráficos de alinhamento do modelo de atenção, do espectrograma mel e dos áudios gerados usando o vocoder de fase Griffin-Lim. As frases sintetizadas, espectrogramas, gráficos de alinhamento, modelos treinados e a planilha com os resultados da avaliação podem ser encontradas no repositório do github[4].

4.1 MODELO TREINADO DO ZERO

Com a síntese das 200 frases pelo modelo do Tacotron 2 treinado do zero, foi feita uma análise das figuras de alinhamento do modelo de atenção entre os caracteres de entrada e quadros de espectrogramas mel na saída. Ao fazer a inspeção visual de cada uma das 200 figuras, foi possível observar presente a diagonal característica em quase todas as figuras, com poucas exceções. Isso mostra que os alinhamentos encontrados durante as avaliações no treinamento continuam acontecendo durante a inferência de novas frases nunca vistas antes. Um exemplo do alinhamento de atenção pode ser visto na Figura 24.

Figura 24 - Alinhamento de uma das frases sintetizadas do modelo treinado do zero.

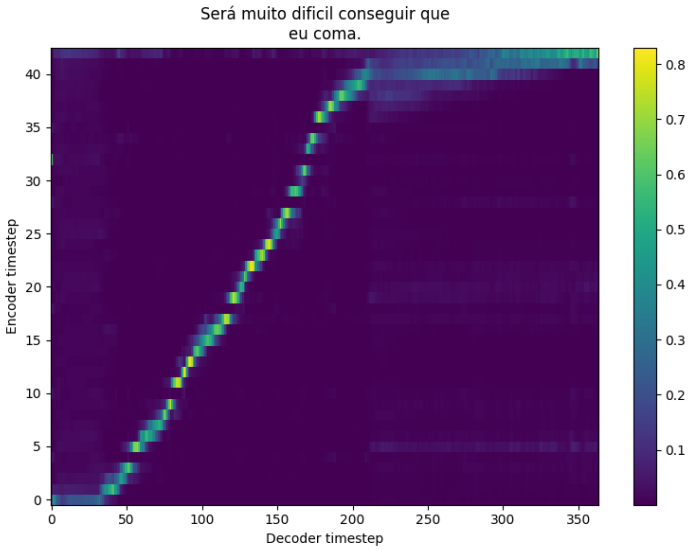


Fonte: Elaborado pelo autor.

[4] Repositório: https://github.com/kobarion/tacotron2-GL-brazillian-portuguese

Porém dentre as figuras de alinhamento das frases sintetizadas, foram observados casos em que o decodificador falha em detectar o *token* de parada causando com que a diagonal relacionando a entrada com a saída acaba continue tentando prever um trecho de silêncio ao final da síntese. Isso pode ser observado na Figura 25, a partir do *decoder timestep* acima de 200 em que há uma reta horizontal perto do *encoder timestep* 40.

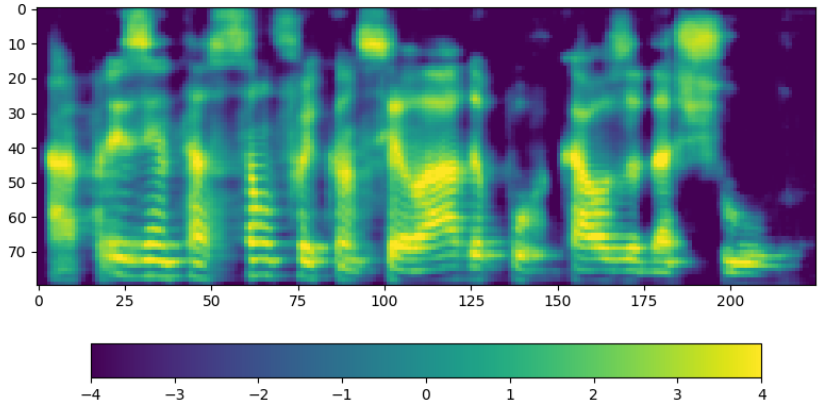
Figura 25 - Alinhamento de uma das frases sintetizadas com trechos de silêncio.



Fonte: Elaborado pelo autor.

Analisando agora a saída do modelo na Figura 26 para a mesma frase da Figura 24, observa-se que o espectrograma mel apresenta a condição suavizada do espectrograma gerado durante o treinamento, tendo menos detalhes ainda. Porém ainda assim é possível identificar facilmente as formantes bem definidas do sinal de fala no espectrograma.

Figura 26 - Espectrograma mel da frase sintetizada do modelo treinado do zero.



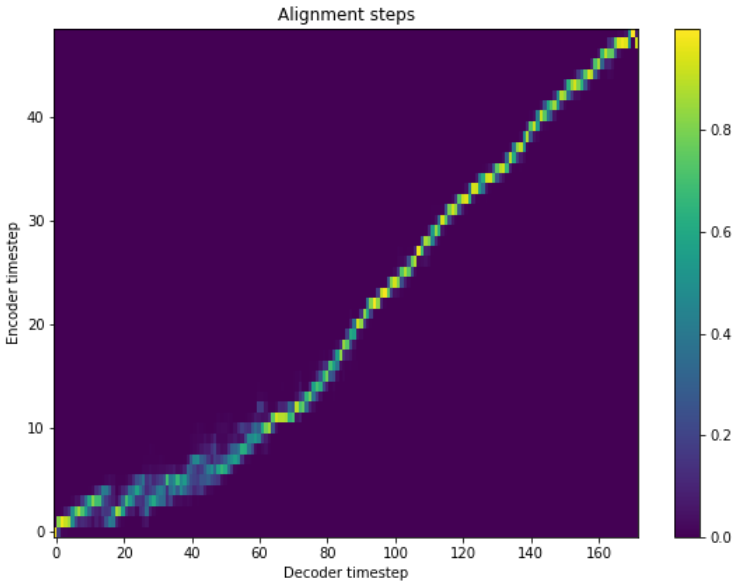
Fonte: Elaborado pelo autor.

4.2 MODELO *FINETUNED*

Fazendo a mesma análise dos alinhamentos das frases sintetizadas do modelo *finetuned* observa-se que apesar da diagonal característica aparecer em muitas figuras e não ocorrer o problema do modelo anterior de preencher o final do áudio com um trecho de silêncio, existe a ocorrência de alguns desalinhamentos, principalmente no início da frase.

Na Figura 27 temos o alinhamento do modelo de atenção da mesma frase utilizada no exemplo para o modelo treinado do zero. Na figura pode-se observar que entre os instantes 10 a 60 d0 *decoder timestep*, em vários passos no tempo da saída estão alinhados com as mesmas entradas. Ouvindo o áudio sintetizado, essa condição do alinhamento causou que a fala "gaguejasse" por um breve momento, alongando o som do primeiro "A" na frase "A apresentação foi cancelada por causa da chuva".

Figura 27 - Alinhamento de uma das frases sintetizadas do modelo *finetuned.*



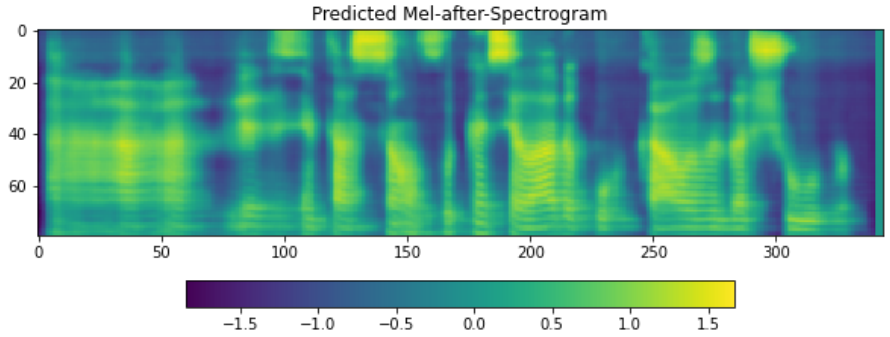
Fonte: Elaborado pelo autor.

Foram encontradas outras figuras de alinhamento em que é possível perceber que o modelo não foi capaz de realizar a inferência do áudio da frase, algo que não ocorreu nos áudios sintetizados pelo modelo treinado do zero.

Durante a análise dos espectrogramas, pôde ser percebido que o detalhamento dos espectrogramas é pior do que os resultados do modelo treinado do zero, tendo um efeito

borrado no espectro. Na Figura 28, pode ser visto esse efeito e também o alongamento do fonema “A” no começo do espectrograma.

Figura 28 - Espectrograma mel da frase sintetizada do modelo *finetuned*.



Fonte: Elaborado pelo autor.

4.3 AVALIAÇÃO DA INTELIGIBILIDADE

Conforme o processo descrito na seção 3.5, as 200 frases foram sintetizadas utilizando o vocoder Griffin-Lim e salvas como arquivos de áudio no formato wav e para facilitar o processo de avaliação. Para a avaliação da inteligibilidade, foi feita uma planilha com uma coluna identificando as frases sintetizadas e suas transcrições. O processo de testes consistiu na abertura de cada arquivo de áudio, a escuta de cada frase e identificação de fonemas das palavras pronunciadas incorretamente, anotando na planilha a forma como foi pronunciado. Além disso, foi verificado se durante a síntese os modelos pularam palavras, registrando quais e quantas palavras apresentaram erros em cada frase. Na Tabela 1 foi feito o resumo das análises obtidas.

Tabela 1 - Resultados obtidos na avaliação auditiva de cada frase sintetizada.

Total de frases	**Total de palavras**	**Modelo**	**Palavras puladas**	**% de palavras puladas**	**Erros de pronúncia**	**% de erros de pronúncia**
200	1349	Treinado do zero	12	0,89%	51	3,78%
		Finetuned	251	18,60%	70	5,19%

Fonte: Elaborado pelo autor.

Das 200 frases sintetizadas, foram encontradas 1349 palavras. Observou-se novamente que o modelo treinado do zero foi superior em ambas as métricas avaliadas. Foram 0,89% das palavras que acabaram sendo puladas e 3,78% das palavras em que houve erro de pronúncia. Em comparação, o modelo utilizando *transfer learning* do inglês teve 18,60% das palavras puladas ou sem conseguir sintetizar a frase e 5,91% do total de palavras sintetizadas tiveram erros de pronúncia. Em relação aos artefatos de áudio, na forma de ruído que acabaram introduzidos durante a síntese, foram identificados artefatos em 31 das frases do modelo treinado do zero enquanto que para o *finetuned* foram encontrados 78.

Com essas métricas, verifica-se que os resultados do modelo treinado do zero obteve os melhores resultados durante a síntese das frases, tendo menos palavras puladas, erros na pronúncia e introdução de artefatos. Os erros de palavras sendo puladas representam um problema maior na inteligibilidade da frase do que a pronúncia incorreta. Sendo que com uma taxa de 0,89% de palavras puladas e 3,78% de palavras com erros de pronúncia, o modelo não acarreta perda significativa de inteligibilidade. Um exemplo de aplicação seria na área de entretenimento, como robôs falantes, notificações de aplicativos ou narração de vídeos.

Vale mencionar que o resultado acima difere do obtido no trabalho de Casanova (2020), o qual obteve um melhor desempenho a partir do *transfer learning* do inglês. Essa diferença pode ter sido ocasionada pelo fato de que o presente trabalho utilizou duas implementações diferentes para a síntese com e sem *finetuning*, além do fato de que o trabalho de Casanova difere deste em diversos aspectos estruturais, como a implementação do Tacotron 2 utilizada, vocoder, conjunto de dados e uso de fonemas como entrada.

5 CONCLUSÃO

Neste trabalho foi realizado um estudo sobre a utilização de um conjunto de dados de fala para o treinamento de dois formatos de treinamento diferentes do Tacotron 2 para sistemas de conversão texto-fala.

Os resultados demonstraram que é possível treinar o Tacotron 2 para o português com poucas mudanças no modelo original. O treinamento a partir do modelo do zero trouxe os melhores resultados em relação aos dois experimentos, obtendo uma fala com qualidade, inteligibilidade e prosódia superiores, além de ter poucos erros de síntese. O modelo teve dificuldade em prever o *token* de parada da frase resultando em trechos de silêncio no final das frases e como também na pronúncia de algumas palavras.

Um dos testes realizados foi em relação ao *fine tuning* de um modelo pré-treinado em inglês para o português, com o objetivo de aproveitar o trabalho consolidado com outro conjunto de dados e diminuir recursos de treinamento e extensão do conjunto de dados. O alinhamento foi obtido muito rápido e o modelo convergiu rapidamente como esperado obtendo uma fala sintetizada inteligível, porém com muito mais artefatos de som, piorando muito a qualidade obtida.

A qualidade da fala sintetizada para ambos os modelos poderia ter tido uma melhoria significativa caso o treinamento de outros vocoders como o WaveNet pudessem ter sido feitos. A principal dificuldade encontrada foi em relação à quantidade de memória da GPU que esses modelos necessitam para fazer as operações de suas redes, além disso, o WaveNet é um modelo autorregressivo, o que dificulta a aceleração a partir da paralelização.

Em futuros trabalhos, seria interessante treinar conjuntamente um vocoder neural como o WaveNet ou outras implementações como o Fast WaveNet, WaveRNN, Parallel WaveGAN e o MelGAN. Outra opção seria atualizar o conjunto de dados do Common Voice, que em sua versão mais recente possui quase o dobro de horas disponíveis e treinar um modelo multi-idioma. Os avanços na área estão mais acelerados do que nunca e ainda existe muito a ser explorado e aprimorado para o português brasileiro.

REFERÊNCIAS

ALMEIDA, F.; GERALDO X. Word Embeddings: A Survey. **ArXiv**, abs/1901.09069. 2019.

ASIMOV, I.; MACQUARRIE, R. **Robot visions**. Tradução. [s.l.] Roc, 1990.

ARDILA, R. *et al*. Common Voice: A Massively-Multilingual Speech Corpus. Disponível em: <https://arxiv.org/abs/1912.06670>

BAHDANAU, D.; CHO, K.; BENGIO, Y. Neural Machine Translation by Jointly Learning to Align and Translate. **CoRR**, abs/1409.0473. 2014

CASANOVA E. *et al.* TTS-Portuguese Corpus: a corpus for speech synthesis in Brazilian Portuguese. **arXiv preprint arXiv:2005.05144**, 2020.

CHO, K. et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, 2014.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. Traducao . [s.l.] The MIT Press, 2017.

GOODFELLOW, I. *et al*. Generative adversarial nets. **In Proceedings of the 27th International Conference on Neural Information Processing Systems** - Volume 2 (NIPS'14). MIT Press, Cambridge, MA, USA, 2672–2680. 2014.

GRAVES, A.; MOHAMED, A.-R.; HINTON, G. Speech recognition with deep recurrent neural networks. **2013 IEEE International Conference on Acoustics, Speech and Signal Processing**, 2013.

GRIFFIN D.; LIM J.. Signal estimation from modified short time Fourier transform. **IEEE Trans. Acoust., Speech, Signal Process.**, vol. 32, no. 2, pp. 236–243, Abril, 1984.

JURAFSKY, D.; MARTIN, J. H. **Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition**. Traducao . [s.l.] Pearson Prentice Hall, 2009.

KIM, J. *et al*. Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search. **ArXiv**, abs/2005.11129. 2020.

KLATT D. Review of text-to-speech conversion for English. **The Journal of the Acoustical Society of America**, vol. 82 3, pp. 737-739, 1987.

KUMAR, K. *et al*. MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis. **NeurIPS**. 2019.

LI, N. *et al.* Close to Human Quality TTS with Transformer. **ArXiv**, abs/1809.08895. 2018.

LOIZOU, P. C. **Speech enhancement: theory and practice**. CRC Press, 2013.

LUONG, T.; PHAM, H.; MANNING, C. D. Effective Approaches to Attention-based Neural Machine Translation. **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, 2015.

MIAO, C. *et al*. Flow-TTS: A Non-Autoregressive Network for Text to Speech Based on Flow. **ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, 2020, pp. 7209-7213

NING, Y. *et al*. A Review of Deep Learning Based Speech Synthesis. **Applied Sciences**, v. 9, n. 19, p. 4050, 2019.

OORD, Aaron van den *et al*. WaveNet: A Generative Model for Raw Audio. **CoRR**, arxiv:1609.03499, 2016.

PING, W. *et al.* Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning. **ICLR**. 2017.

PRENGER, R. *et al*. Waveglow: A Flow-based Generative Network for Speech Synthesis. **ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**: 3617-3621. 2019.

PURWINS H. *et al*. Deep Learning for Audio Signal Processing. **IEEE Journal of Selected Topics in Signal Processing**, vol. 13, no. 2, pp. 206-219, Maio, 2019.

QUATIERI, T.. **Discrete-Time Speech Signal Processing: Principles and Practice**. Prentice Hall Press, 1st ed., Upper Saddle River, NJ, USA, 2001.

QUINTAS, S.; TRANCOSO, I. Evaluation of Deep Learning Approaches to Text-to-Speech Systems for European Portuguese. **Lecture Notes in Computer Science Computational Processing of the Portuguese Language**, p. 34–42, 2020.

RABINER, L. R.; SCHAFER, R. W. **Introduction to digital speech processing**. Prentice-Hall, 2010.

REN, Y. *et al*. FastSpeech: Fast, Robust and Controllable Text to Speech. **NeurIPS** (2019).

REN, Y. *et al*. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. **ArXiv** abs/2006.04558. 2021.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **Nature**, v. 323, n. 6088, p. 533–536, 1986.

SAGISAKA Y. Speech synthesis by rule using an optimal selection of nonuniform synthesis units. **Proceedings of the International Conference on Acoustics, Speech and Signal Processing**, pp. 679–682, 1988.

SARDINHA, Tony Berber. Lingüística de Corpus: histórico e problemática. **DELTA**, São Paulo , v. 16, n. 2, p. 323-367, 2000.

SEARA, I.. **Estudo Estatístico dos Fonemas do Português Brasileiro Falado na Capital de Santa Catarina para elaboração de Frases Foneticamente Balanceadas**. Dissertação de Mestrado, Universidade Federal de Santa Catarina, 1994.

SHEN, J. *et al*. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. **2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, Calgary, AB, 2018, pp. 4779-4783.

SKERRY-RYAN, R. J. *et al.* Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron. **CoRR**, abs/1803.09047. 2018.

SUTSKEVER, I.; VINYALS, O.; LE, Q.V. Sequence to Sequence Learning with Neural Networks. **NIPS**. 2014.

TACHIBANA, H. *et al*. Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention. **2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2018)**: 4784-4788.

TAN, X. *et al*. A Survey on Neural Speech Synthesis. **ArViX**, abs/2106.15561. 2021.

TAYLOR, P.. **Text-to-Speech Synthesis**. Cambridge: Cambridge University Press. 2009.

VALLE, R. *et al*. Flowtron: an Autoregressive Flow-based Generative Network for Text-to-Speech Synthesis. **ArXiv,** abs/2005.05957. 2021.

WANG, Y. *et al*. Tacotron: Towards End-to-End Speech Synthesis. **Proc. Interspeech 2017**, 4006-4010. 2017.

ZEN, H.; SENIOR, A.; SCHUSTER, M. Statistical parametric speech synthesis using deep neural networks. **2013 IEEE International Conference on Acoustics, Speech and Signal Processing**, 2013.

ZEN, H.; TOKUDA, K.; BLACK, A. W. Statistical parametric speech synthesis. **Speech Communication**, v. 51, n. 11, p. 1039–1064, 2009.

ZHANG, Y. *et al.* Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning. **Interspeech 2019**, 2019.