Estratégias de Sintetização de Voz utilizando Deep Learning para Leitura Humanizada de Textos

Aluno: ??? Orientador: ???
??? ???

Abstract

Text-to-speech (TTS) is an extremely important tool in the pursuit of accessibility. It benefits not only people with visual impairments but also those with difficulties in reading comprehension. Recent advancements in deep learning enable the synthesis of voice with quality very close to recordings made by humans. In this study, we aim to replicate state-of-the-art techniques used to synthesize voices within the context of Brazilian Portuguese.

Keywords: Deep learning, Acessibility

Resumo

A síntese de fala a partir de texto é uma ferramenta crucial para promover a acessibilidade. Não apenas beneficia pessoas com deficiência visual [(WAI)(2024)], mas também aquelas com dificuldades na compreensão de textos [Wood et al.(2017)]. Avanços recentes no campo do deep learning têm possibilitado a geração de voz sintética com qualidade comparável à de gravações humanas [Tan et al.(2022)]. Neste estudo visamos replicar técnicas de síntese de voz para o contexto do português brasileiro com vistas a uma posterior realização de estudo comparativo entre técnicas.

Palavras-chave: Deep learning, Acessibilidade

1 Introdução

Avanços recentes na área de deep learning têm possibilitado o desenvolvimento de sistemas de síntese de voz humana de maneira cada vez mais precisa e natural. Tan et. al. por exemplo, produziram o NaturalSpeech, um modelo capaz de replicar a voz humana com uma qualidade muito similar a de gravações humanas [Tan et al.(2022)]) na língua inglesa. Essa tecnologia oferece uma possibilidade de melhoria na acessibilidade digital. Recursos de acessibilidade de sintetização de voz a partir de textos (do inglês text-to-speech ou simplesmente TTS) são utilizadas a bastante tempo, mas ainda existe uma lacuna no impacto que ferramentas que se aproximam mais a dicção e a voz humana podem trazer na melhoria da qualidade de vida para pessoas com deficiência visual e dificuldade de compreensão de textos. Neste estudo pretendemos replicar técnicas de TTS no português do brasil.

1.1 Objetivo

Este artigo propõe explorar síntese de voz humanizada utilizando técnicas de deep learning. Buscamos desenvolver sistemas TTS com foco na naturalidade para tentar reproduzir a entonação, o ritmo e outras características da fala humana para o português brasileiro. Em um primeiro momento nos concentramos em tecnologias existentes para entender os desafios iniciais dessas técnicas.

1.2 Questões de Pesquisa

- Quais são as técnicas mais eficazes de deep learning para desenvolver TTS em português mais natural e humanizada (entonação, ritmo e outras características)?
- Como validar e avaliar o quão humanizada está a leitura do texto obtida pelo TTS?

2 Revisão da Literatura

2.1 Artigos Selecionados

2.1.1 Artigo 1: TACOTRON: TOWARDS END-TO-END SPEECH SYNTHESIS

Contexto e contribuições:

O contexto do artigo é de *text-to-speech* (TTS) ou síntese de voz a partir de texto em inglês. Nele é difinida uma arquitetura onde o modelo recebe caracteres como input e retorna um espectograma. A partir deste espectograma é utilizada uma outra técnica chamda de WaveNet [van den Oord et al.(2016)] para transformar o espectograma em audio utilizando uma reconstrução Griffin-Lim [Griffin and Lim(1984)].

Como a solução proposta pelo artigo foi avaliada?

A solução foi avaliada através de uma survey de falantes nativos com *Mean Opinion Score* (MOS) comparando com outros modelos de TTS: O modelo Parametric [Zen et al.(2016)] e o modelo Concatenativa [Gonzalvo et al.(2016)].

Pontos fortes do artigo:

- Os modelos são bem claros, e o processo é bem explicado.

Pontos que o artigo poderia melhorar:

- Artigo não acompanha código.
- Artigo não acompanha modelo para facilitar a validação.

2.1.2 Artigo 2: Conversão Texto-Fala para o Português Brasileiro Utilizando Tacotron 2 com Vocoder Griffin-Lim

Contexto e contribuições:

Neste artigo há uma aplicação do TACOTRON, mas dessa vez com foco no português brasileiro. O artigo consegue aplicar o português, contudo o resultado é bem aquem do esperado com diversas vozes extremamente robóticas. [Rosa and Silva(2021)]

Como a solução proposta pelo artigo foi avaliada?

Os autores decidiram por fazer uma avaliação de quantidade de erros de pronúncia e palavras puladas. No artigo não há uma definição do que são palavras que foram puladas mas assumo que sejam palavras cujo modelo não foi capaz de pronunciar e simplesmente ignorou.

Pontos fortes do artigo:

- O artigo acompanha um repositório (github) com códigos, ou seja, pode ser validado e replicado com certa facilidade.

Pontos que o artigo poderia melhorar:

- A validação parece ser uma estratégia boa para quando não se tem muitos recursos, mas ela é extremamente limitada a um modelo que espera-se que cometa erros muito fortes.
- Resultados das vozes geradas não são muito bons.

2.1.3 Artigo 3: NaturalSpeech: End-to-End Text-to-Speech Synthesis with Human-Level Quality

Contexto e contribuições:

O contexto também é de TTS, além de trazer a proposta de uma técnica de sintetização de voz, o artigo também se propõe a trazer uma maneira de definir e julgar o quão próximo o sintetizador de voz está do nível de qualidade de voz humana. É afirmado no artigo que se conseguiu algo muito próximo da qualidade humana. Um sintetizador de qualidade humana que é definido como: Um sintetizador que gera vozes em que não há diferença estatística entre uma gravação humana e a síntese de voz pela primeira vez em uma *MOS*.

Como a solução proposta pelo artigo foi avaliada?

A solução foi avaliada através de uma survey de falantes nativos com escala MOS(mean opinion score) comparando com outros modelos de TTS: O modelo Parametric [Zen et al.(2016)] e o modelo Concatenativa [Gonzalvo et al.(2016)]. Este artigo foi escolhido não somente por ser estado da arte, como também por ter compreensiva formalização matemática.

Atividade 1 de FPCC3: A formalização a seguir explica um autocodificador variacional. O objetivo de um autocodificador variacional é codificar uma entrada automaticamente em uma representação comprimida que pode depois ser decodificada para que a entrada possa ser reconstruída a partir do que foi codificado. Isso pode ser usado para explicar o processo de reconstrução de voz a partir de um treinamento utilizando este autocodificador. Em um primeiro momento é feita a codificação em algo que é chamado de vetor latente Z. Em um momento do artigo é feita a formalização matemática do autocodificador para explicar seu funcionamento.

X: Voz original

Y: Sequência de texto

$q(z \mid x)$: Encoder que parametriza a distribuição dos dados X para a a variável latente Z.

$p(x \mid z)$: Decoder que reconstrói o input que foi passado que no caso é o audio passado.

$p(z \mid y)$: Um segundo encoder que codifica uma sequência de texto Y em fonemas. E será passado ao Decoder posteriormente

$S : p(z \mid y) \rightarrow p(x \mid z)$: Voz Sintetizada a partir do processo de decodificação onde em um primeiro momento se obtém o vetor Z a partir do texto, para depois ser feita a decodificação do spectograma, o qual é utilizado em um segundo momento para alimentar um sintetizador de sons no domínio do tempo.

Pontos fortes do artigo:

- Acompanha código o que facilita a replicação.

Pontos que o artigo poderia melhorar:

- Muitos termos específicos da área, o que faz com que o artigo seja um pouco inacessível para alguém que está começando os estudos.

3 Metodologia

3.1 Tipo de Pesquisa

O tipo de pesquisa adotado que seguiremos é de experimentação e comparação. Em um primeiro momento testaremos múltiplas técnicas de síntese de voz, tentando entender os desafios para aplicá-las no contexto do português brasileiro e realizaremos a comparação dos resultados.

3.2 Estudos Selecionados

Alguns dos estudos selecionados estão contidos na seção de revisão sistemática da literatura. Atualmente o Estado da arte é o NaturalSpeech [Tan et al.(2022)], realizaremos experimentações com os modelos de deep learning do tacotron [Wang et al.(2017)] e de outros estudos a serem definidos.

3.3 Extração dos dados

Para popular os nossa base de dados de vozes, utilizaremos o portal Brasil [1], mas estamos em busca de outros bancos de vozes disponíveis gratuitamente online para este momento preliminar da pesquisa.

[1] https://gitlab.com/fb-audio-corpora

3.4 Validação

Parte dos nossos estudos iniciais vão ser para entender como validar as nossas experimentações. Alguns dos estudos que já mapeamos utilizam *MOS*, que são métricas subjetivos feitos através de pesquisas com falantes nativos da linguagem. Para fazer a experimentação em si, utilizaremos código aberto das técnicas discutidas no artigo e manipularemos parâmetros de entrada para fazer comparações entre as técnicas e seus resultados. Estudaremos também possibilidades de algoritmos de comparação entre a voz sintetizada e um baseline de voz humana gravada.

References

[Gonzalvo et al.(2016)] Xavi Gonzalvo, Siamak Tazari, Chun an Chan, Markus Becker, Alexander Gutkin, and Hanna Silen (Eds.). 2016. *Recent Advances in Google Real-time HMM-driven Unit Selection Synthesizer*. Sep 8–12, San Francisco, USA. http://www.isca-speech.org/archive/Interspeech_2016/pdfs/0264.PDF

[Griffin and Lim(1984)] D. Griffin and Jae Lim. 1984. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32, 2 (1984), 236–243. https://doi.org/10.1109/TASSP.1984.1164317

[Rosa and Silva(2021)] Rodrigo Rosa and Danilo Silva. 2021. Conversão Texto-Fala para o Português Brasileiro Utilizando Tacotron 2 com Vocoder Griffin-Lim. https://doi.org/10.14209/sbrt.2021.1570727280

[Tan et al.(2022)] Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, Frank Soong, Tao Qin, Sheng Zhao, and Tie-Yan Liu. 2022. NaturalSpeech: End-to-End Text to Speech Synthesis with Human-Level Quality. arXiv:2205.04421 [eess.AS]

[van den Oord et al.(2016)] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. arXiv:1609.03499 [cs.SD]

[(WAI)(2024)] W3C Web Accessibility Initiative (WAI). 2024. Text to speech. https://www.w3.org/WAI/perspective-videos/speech/

[Wang et al.(2017)] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017. Tacotron: Towards End-to-End Speech Synthesis. arXiv:1703.10135 [cs.CL]

[Wood et al.(2017)] Sarah Wood, Jerad Moxley, Elizabeth Tighe, and Richard Wagner. 2017. Does Use of Text-to-Speech and Related Read-Aloud Tools Improve Reading Comprehension for Students With Reading Disabilities? A Meta-Analysis. *Journal of Learning Disabilities* 51 (01 2017), 002221941668817. https://doi.org/10.1177/0022219416688170

[Zen et al.(2016)] Heiga Zen, Yannis Agiomyrgiannakis, Niels Egberts, Fergus Henderson, and Przemysław Szczepaniak. 2016. Fast, Compact, and High Quality LSTM-RNN Based Statistical Parametric Speech Synthesizers for Mobile Devices. arXiv:1606.06061 [cs.SD]