Comparação de Modelos TTS: XTTS vs. YourTTS

Aluno: Thomaz Diniz Pinto de Morais
thomaz.morais@ccc.ufcg.edu.br

Orientador: Herman Martins Gomes
hmg@dsc.ufcg.edu.br

1 Introdução

A síntese de voz (TTS) é uma tecnologia essencial para diversas aplicações, como assistentes virtuais, leitura automática de textos, e acessibilidade. Este estudo visa replicar parcialmente o artigo "Conversão Texto-Fala para o Português Brasileiro Utilizando Tacotron 2 com Vocoder Griffin-Lim" [Rosa and Silva(2021)]. Nele os autores se utilizam de um conjunto de dados chamado Common Voice para treinar o modelo Tacotron e gerar audio em português brasileiro.

2 Metodologia

Neste artigo utilizaremos uma implementação *TTS* disponibilizada pela *coqui-ai*[1] para comparar modelos de *TTS* multilinguais na lingua portuguesa do Brasil. Os modelos selecionados são o *XTTS* e o *YourTTS*, ambos modelos já capazes de sintetizar vozes em português brasileiro. Rosa et. al. [Rosa and Silva(2021)] avalia seu modelo através da contagem da quantidade de erros de pronúncia e de palavras ignoradas. Neste artigo avaliaremos utilizando a mesma contagem de erros e com as mesmas sentenças sintetizadas pelo modelo do artigo original. Para isto, sintetizamos as 200 sentenças que foram disponibilizadas pelo artigo de Rosa et. al. para os dois modelos que selecionamos (*XTTS* e *YourTTS*) e, posteriormente, fizemos a contagem da quantidade de erros de pronúncia e palavras ignoradas pelo sintetizador.

[1]Repositório no GitHub: https://github.com/coqui-ai/TTS

3 Reprodução

Para sintetização de voz utilizamos a implementação disponível no repositório do github da *coqui-ai*. Desenvolvemos também um passo-a-passo de como executar os códigos para sintetização das sentenças. O resultado disso está disponível em um repositório do github[2], bem como os áudios sintetizados, e uma página com resultados e demonstração dos áudios sintetizados. Ao finalizar a sintetização das sentenças, fizemos a avaliação das sentenças uma a uma manualmente. Escutando o resultado e verificando se houveram erros de português e palavras ignoradas pelo sintetizador de voz.

Os dois modelos , *XTTS* e *YourTTS*, foram avaliados com as mesmas 200 sentenças do artigo de Rosa et al. Fizemos a contagem de erros de pronúncia de forma manual e registramos os seguintes resultados:

Tipo de Erro	*XTTS*	*YourTTS*
Palavras puladas	0	0
Erros de pronúncia	10	50

Table 1: Comparação dos resultados entre *XTTS* e *YourTTS*

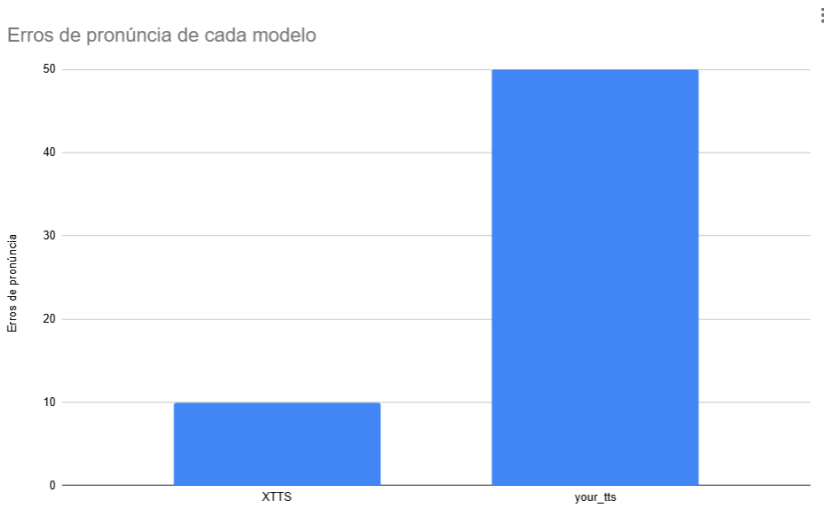

Figure 1: Comparação dos erros de pronúncia entre os modelos XTTS e YourTTS.

[2]Repositório no GitHub: `https://github.com/ThomazDiniz/tts`

4 Conclusões

Os resultados obtidos, bem como alguns dos áudios sintetizados podem ser observados em uma *github page* dedicada aos resultados desta reprodução[3]. Os resultados nos mostram que o modelo XTTS tem uma precisão de pronúncia significativamente melhor em comparação ao YourTTS. A baixa taxa de erros de pronúncia no XTTS indica uma performance superior, tornando-o uma opção mais viável que o YourTTS.

References

[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

[3]Página no GitHub Pages: https://thomazdiniz.github.io/tts/