Uma Perspectiva sobre a Clonagem de Voz e sua Integração em Sistemas TTS

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

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

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

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

***Resumo* — O presente relatório foi realizado para a disciplina de Processamento Digital de Sinais e Imagens. Nele, visou-se abordar um método de clonagem de voz multilingue que demanda recursos mínimos. Dessa forma, foi conduzida a abordagem teórica de seu funcionamento, bem como a aplicação prática do método em novos conjuntos de dados.**

Palavras chave— Low-Resource TTS, Zero-Shot Multispeaker TTS, Language-Agnostic Meta Learning (LAML).

I. INTRODUÇÃO

No atual panorama da tecnologia da informação, destaca-se a clonagem de voz como uma área de pesquisa inovadora e desafiadora. Diante dos avanços exponenciais em processamento de linguagem natural e inteligência artificial, a capacidade de reproduzir com precisão as características distintivas de uma voz humana se tornou uma realidade tangível. O propósito do presente relatório é investigar um método de clonagem de voz multilingue que demanda recursos mínimos, presente no artigo "*Low-Resource Multi-lingual and Zero-Shot Multi-Speaker TTS*" [1], com o intuito de elucidar seus procedimentos, sua arquitetura, bem como suas vantagens e desvantagens. Adicionalmente, será conduzida uma aplicação prática do mencionado método para ilustrar sua eficácia em situações específicas.

Em uma breve busca na online verifica-se a quantidade crescente de ferramentas, algoritmos e artigos sobre o tema. Segundo María Noel Espinosa, data scientist na Marvik, destaca-se como estado da arte em clonagem de voz as técnicas *ZSE-VITS: A Zero-Shot Expressive Voice Cloning Method Based on VITS*, *Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers* e *Low-Resource Multi-lingual and Zero-Shot Multi-speaker TTS*.

II. FUNDAMENTOS TEÓRICOS

Diante do desafio de desenvolver métodos eficientes para a clonagem de voz em um cenário multilíngue e de recursos limitados, destacado na introdução, o artigo 1 oferece soluções inovadoras. Estas incluem, o método de Aprendizagem Meta-Linguagem-Agnóstica (LAML) é um marco no treinamento de modelos text-to-speech (TTS), permitindo que eles se adaptem rapidamente a novos idiomas, mesmo com apenas 5 minutos de amostra de voz. Este avanço é complementado por modificações no codificador de TTS, que permitem um melhor manuseio dos dados multilíngues e a desvinculação de idiomas e falantes, crucial para a síntese de voz natural em diferentes idiomas.

Além disso, o estudo explora a representação de entrada articulatória, permitindo ao sistema lidar com fonemas sobrepostos em diferentes línguas, e a incorporação de fronteiras de palavras e pausas, fundamentais para a naturalidade da fala. A capacidade do modelo de TTS multi-falante Zero-Shot de transferir identidades de falantes em múltiplos idiomas, sem treinamento específico no falante, destaca a versatilidade do sistema. A combinação dessas técnicas representa um avanço notável no campo de TTS, especialmente para idiomas com escassez de dados, abrindo novas possibilidades para aplicações de síntese de voz em diversos contextos linguísticos e culturais.

III. METODOLOGIA

IV. RESULTADOS E CONCLUSÕES

REFERÊNCIAS

[1] F. Lux, J. Koch, and N. Thang Vu, "Low-Resource Multilingual and Zero-Shot Multispeaker TTS", University of Stuttgart, 21 de out. 2022.

[2] María Noel Espinosa, State of the art in Voice Cloning: A review, Marvik, 21 de março de 2023. Disponível em: https://blog.marvik.ai/2023/03/21/state-of-the-art-in-voice-cloning-a-review/.

[3] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples".

[4] K. Azizah, M. Adriani and W. Jatmiko, "Hierarchical Transfer Learning for Multilingual, Multi-Speaker, and Style Transfer DNN-Based TTS on Low-Resource Languages. ".

[5] K. Azizah and W. Jatmiko, "Transfer Learning, Style Control, and Speaker Reconstruction Loss for Zero-Shot Multilingual Multi-Speaker Text-to-Speech on Low-Resource Languages".

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.