Integração de Inteligência Artificial e Clonagem de Voz para Manter a Autenticidade e Aperfeiçoar a Interação do Robô Otto com Crianças com TEA

Maria Vittória S. R. Palma[1], Aldalice R. Dias[1], Nathalia D. Borges[1], Thais Reggina Kempner[1], Luciana C. L. de Faria Borges[2], Eunice P. dos Santos Nunes[2]

[1]Faculdade de Engenharia – Universidade Federal de Mato Grosso (UFMT)
Av. Fernando Correa da Costa, nº 2367 - Boa Esperança. Cuiabá MT – CEP 78060-900

[2]Instituto de Computação – Universidade Federal de Mato Grosso (UFMT)
Av. Fernando Correa da Costa, nº 2367 - Boa Esperança. Cuiabá MT – CEP 78060-900

{vitlia40, aldalice.rd, nathaliadborges, thaisrgk,
lucianafariaborges}@gmail.com, eunice@ufmt.br

Abstract. *Voice cloning through Artificial Intelligence (AI) has advanced significantly, with applications in entertainment, assistive technology, and education. This study, in particular, focuses on the use of voice cloning as a tool to improve the interaction and rehabilitation of children with Autism Spectrum Disorder (ASD). The objective was to explore how this technology can be used to expand the vocabulary of the therapeutic robot Otto, through software such as Eleven Labs and Audacity, which allow the creation of natural voices. Using AI techniques, a boy's voice was recreated in a way that maintained the naturalness and appropriateness of the audios, resulting in a richer vocal interaction and promoting significant advances in the treatment of children.*

Resumo. *A clonagem de voz por Inteligência Artificial (IA) tem apresentado avanços significativos, com diversas aplicações nas áreas de entretenimento, tecnologias assistivas e educação. Este estudo, em particular, foca no uso da clonagem de voz como ferramenta para melhorar a interação e reabilitação de crianças com Transtorno do Espectro Autista (TEA). O objetivo foi explorar como essa tecnologia pode ser utilizada para expandir o vocabulário do robô terapêutico Otto, através de softwares como Eleven Labs e Audacity, que permitem a criação de vozes naturais. Utilizando técnicas de IA, a voz de um menino foi recriada de forma a manter a naturalidade e adequação dos áudios, resultando em uma interação vocal mais rica e promovendo avanços significativos no tratamento das crianças.*

1. Introdução

A clonagem de voz humana por meio de Inteligência Artificial (IA) tem evoluído muito e oferecido uma gama de aplicações que abrangem desde o entretenimento até Tecnologias Assistivas (TA) e educacionais [Amador et al., 2021; Neekhara et al., 2021]. Tais inovações são relevantes ao considerarmos condições como o Transtorno do Espectro Autista (TEA). Segundo Axelsson et al. (2019), o TEA é uma condição com alterações no desenvolvimento neurológico que pode afetar a comunicação, interação social e comportamento. Em crianças, algumas dessas alterações se manifestam em dificuldades na formação de frases e sensibilidade auditiva, tornando-as mais suscetíveis a volumes altos e frequências específicas [Otto-Meyer et al., 2018]. A voz é fundamental para a comunicação diária e a falta desse recurso representa um grande

desafio para o neurodesenvolvimento. Portanto, a clonagem de voz por IA emerge como uma ferramenta promissora para auxiliar indivíduos em condições que prejudicam o uso da fala, como o TEA [Riera et al., 2023]. Tal clonagem utiliza a síntese vocal para permitir que o usuário converta textos em representações auditivas por meio de vozes artificiais, que replicam notavelmente a sua voz natural. (TTS, do inglês *Text to Speech*) [Galdino & Oliveira, 2023; Riera et al., 2023].

No âmbito da TA, o robô Otto foi projetado por nosso grupo para apoiar a reabilitação de crianças com TEA, utilizando cartões *RFID* que acionam áudios correspondentes a atividades de comunicação e desenvolvimento cognitivo [Rebouçaset al., 2023; Dias et al., 2023]. Um menino de 11 anos gravou em estúdio 170 frases, de forma a atender a sensibilidade auditiva das crianças [Rebouças et al., 2023]. Após testes em terapia com crianças, o robô recebeu avaliações positivas de terapeutas [Andrade et al., 2023].

Na pesquisa atual, há a necessidade de ampliar o vocabulário presente na vocalização do robô. Entretanto, diante da impossibilidade de novas gravações, devido à mudança vocal do menino e limitações orçamentárias, optou-se por clonar a voz do mesmo com IA. Visando atender às necessidades específicas do público-alvo, buscou-se estratégias para manter a naturalidade e adequação dos áudios [Chen & Jiang, 2023]. Nesse contexto, esta pesquisa visa investigar o potencial da clonagem de voz para atender as necessidades do robô assistivo Otto dentro deste estudo de caso. O estudo descreve o processo de seleção das técnicas aplicadas para recriar com alta fidelidade a voz de um menino, focando em aspectos como entonação, emoções e realismo na fala.

2. Metodologia

Essa pesquisa, baseada em um estudo de caso, por explorar a clonagem de voz de uma criança pela IA [Rosa, 2021], foi aprovada pelo Comitê de Ética em Pesquisa com Seres Humanos da UFMT. Adota o levantamento bibliográfico para embasar o seu foco, e considera manter a autenticidade e melhora da interação do robô com crianças com TEA, pela voz clonada. As bases de busca foram IEEE Xplore, Google Scholar e Scopus, com as strings "*Artificial Intelligence*" AND "*voice cloning*" AND "*autism*", AND "*voice synthesis*" AND "*robot interaction*" OR "*human-robot interaction*" AND "*autistic children*". Realizou-se também a análise de tecnologias de IA existentes, incluindo ferramentas de clonagem de voz como *Eleven Labs* (https://elevenlabs.io/), *Play HP* (https://play.ht) e *VoiceIA* (https://voice.ai/), para compreender suas capacidades e limitações, sendo escolhida a *Eleven Labs*, por algoritmos avançados e processamento de sinais que capturam com precisão os padrões de conversa e entonação de um indivíduo, índices de custo-benefício e satisfação dos usuários na educação e terapia. Para garantir uma voz clonada autêntica, foi essencial alimentar a IA com dados variados e de alta qualidade: as 170 frases originais do estudo, propiciaram uma clonagem mais coesa com a realidade e asseguram que o processo mantivesse a coerência emocional [Casanova et al., 2023]. Devido à necessidade de simplificação de alguns dos áudios do estudo, escolheu-se o software de edição de áudio gratuito e de código aberto *Audacity*, por seus recursos de edição e sua ampla disponibilidade em múltiplas plataformas [Jaworski & Thibeault, 2011]. A ferramenta permitiu a edição, o processamento e a aplicação de efeitos nos áudios existentes, além da manutenção dos

novos áudios criados pela *Eleven Labs*. Foram ainda utilizadas técnicas de retroalimentação dos áudios já gerados, para otimizar os resultados e expandir o banco de dados para clonagem.

3. Resultados e Discussões

Apesar da IA fornecer possibilidades de replicar características da voz humana para conseguir o realismo, é necessário um conjunto de dados treinados [Salviato, 2023]. O desempenho dessas tecnologias depende da qualidade dos dados utilizados para treiná-las e da otimização dos hiperparâmetros envolvidos [Zen et al., 2016]. A qualidade dos dados está diretamente relacionada com a confiabilidade dos modelos de IA, e os dados alimentados devem conter diferentes tons, sotaques e emoções para garantir que seja capturada uma ampla gama de variações vocais. A rotulagem ajuda a IA a reconhecer e adaptar variações sutis na entonação, como alegria e tristeza, melhorando a qualidade da síntese de voz. Sem essa etiquetagem, a voz sintetizada pode soar artificial ou inadequada [Seong et al., 2021]. Assim, foram explorados elementos fonéticos e linguísticos fundamentais para que uma voz clonada parecesse autêntica: com prosódia [Barbosa, 2012], emoção e personalidade vocal. Com isso, iniciou-se a clonagem com a plataforma *Eleven Labs*. Inicialmente, exigia um material de áudio base de alta qualidade, sendo então compilados os 170 áudios originais gravados em estúdio nos formatos WAV e MP3, visando à máxima qualidade. Esses áudios originais abordavam temas variados, como frutas e legumes, atividades domésticas, animais, alfabeto, objetos, cores, sensações e emoções, apresentando nuances emocionais e sotaque regional. Embora os áudios tivessem curta duração, sua qualidade se destacava. No entanto, o material original totalizava apenas 2 minutos e 50 segundos, sendo necessário duplicar os áudios para atingir o tempo mínimo exigido pela plataforma, resultando no primeiro teste, utilizando as configurações "*Eleven Multilingual v2*".

A seleção criteriosa dos dados de treinamento mostrou-se essencial, pois conjuntos de dados menores e bem selecionados podem gerar vozes com som mais natural [Kuo et al., 2018]. A diversidade de palavras e entonações enriqueceu o material utilizado pela IA, contribuindo para resultados significativos. Além disso, a forma de escrita precisa dos áudios, com pontuações específicas e caracteres, foi fundamental para garantir uma interação vocal mais natural e autêntica. Em seguida, para o processo de clonagem e produção de novos áudios pela *Eleven Labs* foi dedicado tempo ao aprendizado, manuseio e manutenção das funcionalidades da plataforma. Conforme a necessidade de edição dos áudios clonados pela *Eleven Labs*, utilizou-se a *Audacity* para realizar ajustes, como cortar, diminuir ou aumentar a velocidade dos áudios. Os testes iniciais foram realizados com falas do cotidiano. Por exemplo, frases como “Oi, meu nome é Otto! Vamos brincar?” foram usadas nos primeiros testes para familiarizar-se com a plataforma. Buscava-se entender como obter um resultado similar ao áudio original. Inicialmente, notou-se que a escrita da saudação “Oi, meu nome é Otto!!!”, com mais pontos de exclamação, transmitia emoções de felicidade e animação. Posteriormente, foram exploradas características lexicais, como escrever “oii”, resultando em um som mais próximo da realidade. À medida que novos áudios eram gerados, foram realizados vários testes regenerativos do mesmo texto. Observou-se que a utilização da linguagem permitia alcançar resultados variados. Nos primeiros meses,

os resultados apresentaram um tom robótico e sintético. Para melhorar, foram fornecidos *feedbacks* contínuos, sempre contendo palavras-chave relacionadas a “suavidade”, “naturalidade” e “animação”, visando alcançar resultados mais próximos à personalidade vocal original. No teste da frase “vamos brincar?”, apenas com *feedbacks* e configurações específicas foi possível ajustar o tom de voz necessário, que deveria transmitir animação e convite. Mudanças na escrita, como “vamos BRINCAR???”, não apresentaram resultados positivos. Constatou-se, então, a necessidade de estudar aspectos lexicais que modificassem a entonação da fala. Encontrou-se informações sobre linguística, em especial a relação entre fala e escrita, conforme Cristófaro-Silva e Guimarães (2013), que destacam a influência do léxico e da consciência linguística na forma como o sistema de linguagem é interpretado. Isso foi crucial para garantir que as entonações e as escolhas lexicais estejam alinhadas com as expectativas de uso em aplicações de texto para fala. Assim foi essencial apoio lexical durante a escrita.

Buscou-se conhecer também o alfabeto fonético para dar respaldo em suas atividades. A partir dessas ideias, foram realizados testes utilizando pontos, acentos e caracteres especiais para transmitir a ideia desejada, além de tentativas de aumentar as letras para produzir o efeito de alongamento do som. Um dos grandes desafios envolveu interjeições como “hum”, “ah” e “uau”, que são difíceis de expressar de maneira precisa. Por exemplo, no caso de “hum, não foi dessa vez”, o "hum" deveria transmitir tristeza e lamento, mas em alguns testes a IA retornaram a pronúncia da palavra com som de deboche ou sátira. Entonações de surpresa e ironia como “Ah! Que bom que você veio!” também se mostraram complexas de reproduzir na plataforma, exigindo várias gerações do mesmo áudio e técnicas lexicais de alteração de escritas. Percebeu-se também que, por trabalhar com o modelo Multilingual, a plataforma identifica o idioma conforme a inserção da frase, ou seja, se colocasse a palavra “animal”, sozinha, poderia ser entendida como em espanhol, o qual possui a mesma escrita mas diferente pronúncia. Assim, sempre optou-se por escrever uma frase completa como “O cachorro é um animal mamífero”. Com essa estratégia encontrou-se resultados muito significativos, pois assim, a plataforma poderia primeiramente identificar o idioma com clareza e aplicar entonações diversas de timbre no decorrer da frase, retornando de forma mais realística. Ao final, apenas realizava-se o corte do áudio na parte desejada com o *Audacity*.

Passou-se então para o processo de criação de vocabulário específico como objetos e formas geométricas. Palavras como “bicicleta” ou “prisma pentagonal” geraram resultados considerados satisfatórios após análise da equipe do projeto em timbres com apenas algumas regenerações. Apenas foi realizado um trabalho de redução de velocidade com o *Audacity* para chegar em um resultado mais natural. Era essencial que a dublagem de voz se aproximasse ao máximo do sotaque real, pois algumas crianças autistas possuem hipersensibilidade durante a reprodução dos sons [Gomes et al., 2008]. Como palavras solitárias não necessitam de emoções para a compreensão de seu significado, seus resultados foram bons. Em continuidade, foi explicitamente testada a comparação de frases de expressão de emoções para verificar a semelhança com o áudio original e vencer essa problemática relacionada à emoção. Os áudios originais de emoções existentes são: "Estou feliz", "Estou triste", "Estou cansado", "Estou com vergonha", "Estou com medo", "Estou com sono", "Estou com raiva" e "Está muito

barulho". Foram realizados testes para alcançar resultados similares em prosódia e emoção com essas frases. Buscou-se apoio morfológico e lexical para transmitir sentimentos. Versões como “estou cansado…” (transmitindo a ideia de cansaço e desânimo) e “está MUITO barulhoooo” (expressando desconforto e incômodo) foram testadas para alcançar resultados que refletissem as emoções associadas à frase. *Feedbacks* foram fornecidos continuamente pela equipe, tanto quando o resultado era satisfatório, quanto quando necessária a reinterpretação da frase.

4. Conclusão

Verificamos que as ferramentas *Eleven Labs* e *Audacity* possibilitaram um bom realismo para a clonagem de voz do menino para o robô Otto. Embora a entonação tenha sido positiva, pesquisas sobre emoção e suavidade ainda são necessárias. Áudios educativos tiveram bons resultados e com novos testes e *feedback* contínuo, espera-se melhorar a precisão emocional. Além disso, recomenda-se o desenvolvimento contínuo de inteligências artificiais que aprimorem a naturalidade e semelhança das vozes geradas, especialmente para dialetos específicos.

5. Referências

Amador, C., Dario Junior, R., Rossetes, R., Josue, J., Suárez, M., & Ángel, O. (2021). Implementación de clonador de voz en tiempo real para la lengua española usando algoritmos de aprendizaje profundo. Barranquilla, Universidad Del Norte.

Andrade, F., Fagundes, E. M., Van Der, I., et al. (2023). Resultado do uso do robô Otto em terapias com crianças autistas. Semana Acadêmica de Engenharia da Automação e Computação – SEMAC, Cuiabá. Even3, 1(978-85-5722-948-8), 1. https://www.even3.com.br/anais/semac2023/648480-resultado-do-uso-do-robo-otto-em-terapias-com-criancas-autistas/.

Axelsson, M., Racca, M., Weir, D., & Kyrki, V. (2019). A participatory design process of a robotic tutor of assistive sign language for children with autism. In 2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) (pp. 1-8). IEEE.

Barbosa, P. A. (2012). Conhecendo melhor a prosódia: aspectos teóricos e metodológicos daquilo que molda nossa enunciação. Revista de Estudos da Linguagem, 20(1), 11-27.

Casanova, E., Santos, V. G. D., Svartman, F. R. F., Leite, M. Q., Candido Junior, A., Marcacini, R. M., Aluísio, S. M. (2023). Recursos para o processamento de fala. Processamento de linguagem natural: conceitos, técnicas e aplicações em português.

Chen, W., & Jiang, X. (2023). Voice-Cloning Artificial-Intelligence Speakers Can Also Mimic Human-Specific Vocal Expression. https://doi.org/10.20944/preprints202312.0807.v1

Cristófaro-Silva, T., & Guimarães, D. O. (2013). A aquisição da linguagem falada e escrita: o papel da consciência linguística. Letras De Hoje, 48(3), 316–323.

Dias, A. R., Marques, F. A. P., Borges, N. D., Kempner, T. R., Borges, L. C. L. de F., & Nunes, E. P. dos S. (2023). Tecnologias assistivas: Cartões RFID como ferramenta de

auxílio na comunicação de crianças com TEA. 12ª Escola Regional de Informática de Mato Grosso (ERI-MT), Cuiabá/MT. Anais (pp. 97-106). Porto Alegre: SBC.

Galdino, J. C., & Oliveira Jr, M. (2023). Prosódia e síntese da fala: uma revisão integrativa da literatura. Revista da ABRALIN, 1-15.

Gomes, E., Pedroso, F. S., & Wagner, M. B. (2008). Hipersensibilidade auditiva no transtorno do espectro autístico. Pró-Fono Revista de Atualização Científica, 20, 279-284.

Jaworski, N., & Thibeault, M. D. (2011). Technology for teaching: Audacity. Free and open-source software. Music Educators Journal, 98(2), 39-40.

Kuo, F. Y., Aryal, S., Degottex, G., Kang, S., Lanchantin, P., & Ouyang, I. (2018, December). Data selection for improving naturalness of TTS voices trained on small found corpuses. In 2018 IEEE Spoken Language Technology Workshop (SLT) (pp. 319-324). IEEE.

Neekhara, P., Hussain, S., Dubnov, S., Koushanfar, F., & McAuley, J. (2021, November). Expressive neural voice cloning. Asian Conference on Machine Learning (pp. 252-267). PMLR.

Otto-Meyer, S., Krizman, J., White-Schwoch, T., & Kraus, N. (2018). Children with autism spectrum disorder have unstable neural responses to sound. Experimental Brain Research, 236, 733-743.

Rebouças, G. R. B. S., Neves, I. V. D. S., Lima, E. M., Kempner, T. R., Nunes, E. P. S., & Borges, L. C. L. F. (2023). O potencial da robótica no tratamento terapêutico de crianças com Transtorno do Espectro Autista. *SBC*.

Riera, PO, Passano, N., Paez, D., Bach, F., Pupkin, I., Sacerdoti, E., ... & San Martín, H. (2023). Implementação e Avaliação de um Sistema de Clonagem de Voz Rio de la Plata para Assistência na Comunicação Oral. Conferência de Acústica, Áudio e Som (JAAS), Universidade Nacional de Tres de Febrero .

Rosa, A. C. G. (2023). A tutela da voz no mundo da inteligência artificial: aspectos atuais da sua regulamentação no Brasil e na Europa (Trabalho de Conclusão de Curso, Universidade Federal do Rio de Janeiro). Pantheon.

Salviato, J. V. (2023). Geração semi-automática de audiodescrição : utilização de Inteligência Artificial na narração. Bdm.unb.br. https://bdm.unb.br/handle/10483/39256.

Seong, J., Lee, W., & Lee, S. (2021). Síntese de fala multilíngue para clonagem de voz. Em 2021 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 313-316). IEEE.

Zen, H., Senior, A., & Schuster, M. (2016). Listen, attend and spell: A neural network for large vocabulary speech recognition. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4778-4782).