Clonagem de Voz por IA para o Português Brasileiro em Contexto Assistivo: Desafios Comparativos em Cinco Ferramentas de Síntese de Fala

Luiz G. dos Santos[1], Nathalia Borges[2], Matheus Sales[2], Luciana C. L. F. Borges[1], Thais R. Kempner[2], Eunice P. S. Nunes[1]

[1]Instituto de Computação – Universidade Federal de Mato Grosso (UFMT), Cuiabá – MT – Brazil; [2]Faculdade de Engenharia – Universidade Federal de Mato Grosso (UFMT), Cuiabá – MT – Brazil

luizgdsantos91@gmail.com, nathalia@sou.ufmt.br, matheus2017mbs@gmail.com, lucianafariaborges@gmail.com, thaisrgk@gmail.com, eunice@ufmt.br

Abstract. Neural TTS models underperform for Brazilian Portuguese (PT-BR) regarding child voices, emotional prosody, and sublexical phonemes. This paper evaluates five AI voice synthesis tools in an assistive context using 170 studio recordings as corpus. Two challenges persisted across all approaches: emotional prosody and PT-BR phonemes absent from English. Incorporating the multilingual-phonemes-10k-alpha dataset into StyleTTS 2 consistently improved regional phonetic quality. Findings provide a comparative assessment and evidence of a low-cost strategy to improve PT-BR phonetic coverage.

Resumo. Modelos de síntese neural de fala apresentam desempenho reduzido para o português brasileiro (PT-BR) em vozes infantis, prosódia emocional e fonemas sublexicais. Este trabalho avalia cinco ferramentas de síntese por IA em contexto assistivo, com 170 gravações como corpus controlado. Dois desafios persistiram: prosódia emocional e fonemas do PT-BR ausentes do inglês. A incorporação do dataset multilingual-phonemes-10k-alpha ao StyleTTS 2 produziu melhoria consistente na qualidade fonética regional. Os achados oferecem avaliação comparativa e evidência de estratégia de baixo custo para ampliar a cobertura fonética do PT-BR.

1. Introdução

Apesar dos avanços recentes na síntese neural de fala, o desempenho observado em inglês não se transfere adequadamente para línguas de médio recurso como o português brasileiro (PT-BR), criando um descompasso entre estado da arte e aplicações reais. Essa assimetria decorre da escassez de datasets públicos em PT-BR com diversidade de falantes e anotação prosódica adequada [Casanova et al. 2022], e é agravada pelo fato de que modelos treinados predominantemente em inglês tendem a degradar a qualidade para vozes fora dos perfis de referência [Araújo et al. 2026]. Em aplicações assistivas, a autenticidade vocal impacta diretamente a efetividade terapêutica, pois variações de

timbre e entonação afetam o engajamento e a compreensão emocional em crianças com perfis neurodivergentes [Otto-Meyer et al. 2018].

A literatura ainda não oferece avaliações sistemáticas que integrem a síntese de voz infantil, a adaptação ao PT-BR em cenários de dados limitados e a reprodução fiel de características fonéticas e prosódicas em aplicações reais. Nesse contexto, este trabalho investiga essa lacuna a partir de um caso concreto: a necessidade de clonar e expandir o banco de voz de um robô educacional para tutoria infantil com crianças com Transtorno do Espectro Autista (TEA), Transtorno do Déficit de Atenção com Hiperatividade (TDAH) e Dislexia, população para a qual robôs sociais têm eficácia documentada como mediadores terapêuticos [Scassellati et al. 2012]. O banco original foi gravado em estúdio por um menino de 11 anos, totalizando 170 enunciados de vocabulário cotidiano, expressões emocionais e conteúdo educacional [Rebouças et al. 2021]. Com a mudança vocal decorrente da puberdade, tornou-se inviável manter a consistência tímbrica em novas gravações, o que motivou a investigação de clonagem de voz por IA. O sucessor pedagógico do robô, o BITI (Brinquedo Inclusivo para Treinamento Infantil), voltado para alfabetização pelo método fônico [Oliveira e Albuquerque 2021], acrescenta o requisito de fonemas isolados, uma classe de entrada para a qual os modelos neurais de TTS não foram originalmente projetados.

Diante dessas restrições, este trabalho tem como objetivo avaliar a capacidade de diferentes abordagens de síntese e clonagem de voz em reproduzir uma voz infantil em PT-BR com naturalidade e fidelidade fonética, considerando simultaneamente três dimensões críticas: clonagem tímbrica, controle prosódico emocional e síntese de fonemas sublexicais. Para isso, foi conduzido um estudo empírico comparativo ao longo de 18 meses, envolvendo cinco ferramentas.

As principais contribuições deste trabalho são: (i) uma avaliação comparativa controlada de cinco abordagens de síntese de fala aplicadas à voz infantil em PT-BR em um contexto assistivo real; (ii) a análise sistemática dos desafios fonológicos específicos do PT-BR em modelos majoritariamente treinados em inglês; e (iii) a demonstração do impacto da incorporação do dataset multilingual-phonemes-10k-alpha na melhoria da qualidade fonética regional, um resultado ainda pouco explorado na literatura. Os achados têm relevância prática direta para grupos que trabalham com síntese de fala em PT-BR sob restrição de dados, oferecendo tanto um mapa comparativo de ferramentas disponíveis quanto uma estratégia de baixo custo para ampliar a cobertura fonética regional.

2. Metodologia

A partir do objetivo de avaliar o desempenho de cinco ferramentas de síntese e clonagem de voz na reprodução de fala infantil em PT-BR sob severa restrição de dados, este trabalho, que se configura em um estudo experimental comparativo de natureza aplicada, foi conduzido ao longo de 18 meses. O estudo foi estruturado para analisar sistematicamente a fidelidade tímbrica ao locutor original, o controle prosódico

emocional e a precisão fonêmica, com ênfase em segmentos característicos do PT-BR. Todas as ferramentas foram submetidas às mesmas condições de entrada, incluindo corpus de treinamento, textos de teste e critérios de avaliação, de modo a permitir comparação equivalente. Os experimentos foram conduzidos em máquina com processador Ryzen 5 5500, GPU AMD RX 6750 XT (12 GB VRAM), 16 GB de RAM e armazenamento NVMe de 1 TB, executando Ubuntu 24.04.

2.1. Fundamentação na Literatura

O TTS-Portuguese Corpus [Casanova et al. 2022], com 10 horas e 28 minutos de um único falante adulto, foi por anos um dos principais recursos públicos para síntese end-to-end em PT-BR, estabelecendo MOS de referência de 4,03 com Tacotron 2. O panorama de recursos públicos de fala em PT-BR é hoje mais amplo, incluindo corpora multilocutor como o Multilingual LibriSpeech e o Common Voice, além de recursos mais recentes como o Certas Palavras [Araújo et al. 2026] e o capítulo de recursos de fala do livro PLN-BR [Casanova et al. 2026]; ainda assim, poucos desses recursos combinam falante único, qualidade de estúdio e pontuação textual — condições relevantes para síntese de fala. Entretanto, o desempenho originalmente reportado para o TTS-Portuguese Corpus não se sustenta em condições mais próximas do uso real [Araújo et al. 2026]. O modelo multilíngue YourTTS foi treinado com apenas um falante em português [Casanova et al. 2024], implicando degradação para vozes atípicas como a infantil com variação regional.

No campo da prosódia, Barakat et al. (2024), em revisão sistemática com 356 artigos, mostram que modelos de transferência de estilo suavizam extremos emocionais. Bott et al. (2024) propõem prompts em linguagem natural como alternativa de controle. Para fonemas sublexicais, Jurafsky e Martin (2023) explicam que modelos de TTS assumem coarticulação sentencial, tornando fonemas isolados uma entrada fora do regime de treinamento. Maia et al. (2023) confirmam que a geração prosódica é o componente mais frágil em TTS para PT-BR. Em conjunto, esses trabalhos evidenciam limitações recorrentes que a literatura ainda não integrou em avaliações sistemáticas sob restrição de dados, lacuna que orienta este estudo.

2.2. Corpus e Preparação dos Dados

O corpus base é composto por 170 enunciados gravados em estúdio por um locutor masculino de 11 anos com leve sotaque carioca, totalizando aproximadamente 2 minutos e 50 segundos de áudio [Rebouças et al. 2021]. O conjunto combina enunciados de palavra isolada (por exemplo, “alface”, “amarelo”, palavras-veículo como “b de bola”) com frases curtas de vocabulário cotidiano, expressões emocionais e conteúdo educacional, o que explica a curta duração média por enunciado. As gravações apresentam alta qualidade acústica; o áudio foi segmentado manualmente por enunciado, normalizado em intensidade e acompanhado de transcrições textuais revisadas. Quando necessário, os arquivos foram replicados em loop até atingir o volume mínimo exigido

pelas ferramentas, conforme recomendado na documentação do Piper [OHF-Voice 2025]. Em etapas posteriores, amostras sintéticas aprovadas pela equipe foram incorporadas para ampliar a cobertura fonética. Para experimentos de melhoria fonética em PT-BR, foi incorporado ao pipeline o dataset multilingual-phonemes-10k-alpha [StyleTTS2 Community 2024], com ênfase em segmentos sub-representados nos modelos avaliados.

2.3. Ferramentas e Paradigmas Avaliados

Foram avaliadas cinco ferramentas representativas de paradigmas distintos. O ElevenLabs [ElevenLabs 2024] é uma API comercial baseada em nuvem. O Coqui TTS com XTTS-v2 [Casanova et al. 2024] é um modelo open source de clonagem zero-shot multilíngue. O Piper TTS [OHF-Voice 2025] é um modelo VITS com fine-tuning supervisionado. O StyleTTS 2 [Li et al. 2023] e o Kokoro-82M [Hexgrad 2024] são modelos baseados em difusão de estilo. Por fim, o F5-TTS [Chen et al. 2024] combinado com RVC [RVC Project 2023] constitui uma abordagem híbrida de clonagem few-shot com conversão de voz. Cabe notar que StyleTTS 2 e Kokoro-82M são tratados neste trabalho como um único paradigma (difusão de estilo), mas reportados separadamente na Tabela 1 e na Seção 3.4 por apresentarem tempos de convergência e comportamentos distintos o suficiente para justificar análise individualizada. Todas foram configuradas para utilizar o mesmo corpus base, com ajustes mínimos para viabilizar a convergência em ambiente computacional disponível.

2.4. Protocolo de Treinamento

Os procedimentos de treinamento variaram conforme o paradigma de cada ferramenta e foram escolhidos com base na melhor adequação ao corpus disponível e às restrições do projeto. Para clonagem zero-shot (XTTS-v2, F5-TTS), as amostras de referência foram utilizadas diretamente, escolha justificada pela necessidade de clonagem rápida com volume reduzido de dados. Para fine-tuning supervisionado (Piper, StyleTTS 2, Kokoro-82M), os modelos foram adaptados ao corpus no formato requerido, abordagem adotada por permitir maior especialização ao timbre e padrão prosódico do locutor-alvo. Para o ElevenLabs, foi criado um perfil de voz via interface, dado que a plataforma não permite acesso aos pesos do modelo.

Nos modelos com treinamento iterativo, o processo foi monitorado por inspeção qualitativa periódica das amostras geradas. O treinamento iterativo foi conduzido até a estabilização da qualidade da síntese, definida como o ponto a partir do qual melhorias adicionais não eram detectáveis de forma consistente. Empiricamente, esse ponto ocorreu em torno de 15 horas para a maioria das arquiteturas. O tempo mínimo até resultado utilizável variou consideravelmente: RVC convergiu em 10 a 45 minutos; StyleTTS 2 e Kokoro-82M entre 1 e 3 horas; Piper TTS e F5-TTS entre 2 e 5 horas; e o XTTS-v2 demandou entre 24 e 72 horas para o mesmo corpus — tempo mais elevado que se explica pela escassez de dados em português no treinamento original do modelo, o que exige

maior adaptação da distribuição aprendida à nova voz (discutido em mais detalhe na Seção 5).

2.5. Protocolo de Avaliação Qualitativa

A avaliação dos áudios sintetizados foi conduzida por um painel de quatro avaliadores pertencentes à equipe de pesquisa, todos com experiência prévia em processamento de fala e interação humano-robô. Cada amostra foi avaliada segundo quatro critérios qualitativos. O primeiro foi a naturalidade, entendida como grau de semelhança com a fala humana do locutor original. O segundo foi a inteligibilidade, ou seja, a clareza e a compreensão do conteúdo pelo público-alvo. O terceiro foi a coerência emocional, referente à adequação da entonação ao estado afetivo pretendido. O quarto foi a precisão fonêmica, que avalia a fidelidade na realização de fonemas do PT-BR, com atenção a segmentos críticos. A aprovação exigia consenso entre os avaliadores em todos os critérios simultaneamente.

2.6. Protocolo para Fonemas Sublexicais

Para a análise de fonemas isolados exigidos pelo método fônico do robô educacional, foi definido um conjunto de segmentos-alvo representativos, incluindo consoantes plosivas (/b/, /d/, /p/, /t/), fricativas (/f/, /v/, /s/, /z/) e líquidas (/l/, /lh/, /nh/, /R/). A seleção foi orientada pelos fonemas de aquisição mais tardia na criança, que concentram as maiores dificuldades de aprendizado de leitura em crianças com dislexia e TEA, e pelos segmentos com maior probabilidade de falha nos modelos avaliados, especialmente aqueles sem equivalente no inglês. Como os modelos de TTS assumem contexto sentencial, foram testadas duas abordagens: a entrada direta do fonema como texto de síntese, para avaliar o comportamento do modelo diante de entrada sublexical, e a estratégia de palavra-veículo, com inserção do fonema em palavra de contexto controlado e extração manual do segmento em software de edição de áudio. A eficácia foi avaliada quanto à ausência de artefatos como inserção de vogais intrusivas, distorções ou substituições fonêmicas.

3. Ferramentas Avaliadas e Resultados Obtidos

3.1. ElevenLabs

O ElevenLabs é uma plataforma comercial de síntese e clonagem de voz em nuvem, avaliada por sua acessibilidade e suporte declarado ao português. Os controles expostos ao usuário limitam-se a três parâmetros (estabilidade, similaridade com a referência e exagero de estilo), cujo comportamento é não-linear. A geração é estocástica sem semente configurável, de modo que o mesmo texto com as mesmas configurações produz saídas distintas a cada execução.

Para fala neutra, os resultados foram satisfatórios. Para fala emocional, a ausência de controle prosódico direto obrigou o uso de estratégias tipográficas (reticências para ritmo mais lento, exclamações para animação, prefixos como “Com voz muito assustada:”),

com efeitos detectáveis em tristeza e cansaço, porém nulos para medo e surpresa. A calibração acumulada durante uma sessão não persistia entre solicitações de emoções distintas.

Para o método fônico, o ElevenLabs é inadequado. A plataforma não foi projetada para entrada sublexical e apresentou sistematicamente inserção de vogal intrusiva ou silêncio ao receber fonemas isolados como texto. A ausência de controle de semente aleatória torna ainda mais problemática a tarefa: mesmo que um fonema específico seja sintetizado corretamente em uma tentativa, não há garantia de reprodução nas seguintes.

3.2. Coqui TTS / XTTS-v2

O XTTS-v2 é um modelo de clonagem zero-shot multilíngue de código aberto que aceita referências de seis segundos e suporta português brasileiro [Casanova et al. 2024]. A instalação revelou-se mais complexa do que o esperado; incompatibilidades de versão entre PyTorch, torchaudio e CUDA demandaram dias de diagnóstico.

Para fala neutra, a qualidade foi satisfatória e a fidelidade tímbrica ao locutor de referência foi perceptível. Para fala emocional, o mecanismo de condicionamento por clip de referência, que consiste em fornecer um áudio original com a emoção desejada como âncora, atenuava sistematicamente os extremos prosódicos: “Estou com medo” entregue com voz tensa no clip chegava sintetizado próximo ao neutro. O problema decorre da ausência de separação explícita entre identidade do falante e estado emocional no espaço latente, conforme documentado por Cho et al. (2025): o modelo combina os dois sinais em vez de transferir apenas o componente emocional.

Para fonemas isolados, o modelo produziu os mesmos artefatos observados nas demais ferramentas: inserção de vogal intrusiva e troca de idioma no tokenizador multilíngue. A Coqui.ai encerrou as operações em dezembro de 2023; repositório é mantido por fork comunitário, o que representa riscos para projetos com horizonte longo.

3.3. Piper TTS

O Piper é uma ferramenta open source de síntese neural baseada em VITS, projetada para fine-tuning com corpus de falante único e exportação para ONNX Runtime [OHF-Voice 2025]. Diferentemente das ferramentas anteriores, não realiza clonagem zero-shot; requer treinamento supervisionado no formato LJSpeech. Foi avaliado pela estabilidade do pipeline de fine-tuning e pela execução completamente local.

Para fala neutra, o Piper produziu os resultados mais estáveis e reprodutíveis entre todas as ferramentas avaliadas: a mesma entrada de texto gerava saídas auditivamente consistentes entre execuções, o que facilita o controle de qualidade. O treinamento convergiu de forma mais estável do que o XTTS-v2 em situação equivalente, atingindo o platô de qualidade em torno de 15 horas.

Para fala emocional, a ausência de qualquer mecanismo de transferência de estilo ou condicionamento emocional implica síntese neutra por padrão, independentemente da formulação do texto de entrada.

Para o método fônico, o Piper é a ferramenta mais promissora entre as avaliadas, com ressalvas. A reprodutibilidade da saída e a ausência de dependência de GPU para inferência são vantagens diretas para a produção de um banco fonético estável. Contudo, fonemas isolados continuam sendo problemáticos na entrada direta, e a estratégia de palavra-veículo é mais eficaz com o Piper do que com os demais modelos justamente pela consistência entre gerações: o mesmo recorte pode ser reutilizado sem variação. A principal limitação para o BITI é a necessidade de retraining completo para incorporar novos fonemas, sem mecanismo de clonagem rápida.

3.4. StyleTTS 2 e Kokoro-82M

StyleTTS 2 [Li et al. 2023] e Kokoro-82M [Hexgrad 2024] são arquiteturas de difusão de estilo: em vez de condicionar sobre um clip de referência, modelam o estilo como variável latente amostrada durante a síntese, usando o WavLM como discriminador. Ambos foram treinados quase exclusivamente em inglês. O Kokoro, com 82 milhões de parâmetros, apresentou a convergência de fine-tuning mais rápida entre todos os modelos avaliados, atingindo o platô de qualidade em tempo inferior a 15 horas.

Para fala neutra em inglês, a qualidade de ambos é notável. Para fala em PT-BR, fonemas sem equivalente inglês não estavam adequadamente representados no espaço aprendido: vogais nasais (/ã/, /ẽ/, /ĩ/, /õ/, /ũ/) soavam mitigadas ou substituídas por aproximantes ingleses, e o ditongo /ão/ saía consistentemente comprometido. A incorporação do dataset multilingual-phonemes-10k-alpha [StyleTTS2 Community 2024] ao pipeline de treinamento do StyleTTS 2 produziu melhora perceptível e consistente na autenticidade fonética regional, descrita pelos avaliadores como audível. Esse resultado emergiu de uma modificação exploratória do pipeline e é discutido na Seção 5.

Para fala emocional, a difusão de estilo introduz variabilidade prosódica natural maior do que os modelos de referência. Contudo, sem controle explícito da dimensão afetiva no espaço latente, o estado emocional sintetizado é não determinístico e não corresponde necessariamente ao pretendido.

Para o método fônico, o StyleTTS 2 e o Kokoro apresentam o maior potencial futuro, mas a situação atual é limitada. Os fonemas nasais do PT-BR e o ditongo /ão/ são exatamente os segmentos mais relevantes para o método fônico brasileiro e os mais problemáticos nesses modelos. A incorporação do dataset multilíngue atenuou esse problema, mas não o eliminou. Ao mesmo tempo, a difusão de estilo permite maior variabilidade prosódica que pode ser explorada para a síntese de fonemas com entonação contextualmente adequada. Se o suporte ao PT-BR for ampliado nas versões futuras desses modelos, eles se tornarão os candidatos mais fortes para a aplicação fonética, dado o menor tempo de treinamento.

3.5. F5-TTS e Abordagem Híbrida com RVC

O F5-TTS [Chen et al. 2024] é um modelo de clonagem few-shot que opera por fluxo retificado sobre mel-espectrograma, requerendo apenas alguns segundos de referência. Para fala neutra, a fidelidade tímbrica foi satisfatória, comparável ao XTTS-v2. Para fala emocional, o problema de atenuação prosódica reproduziu-se com o mesmo padrão das ferramentas anteriores baseadas em referência. Para fonemas isolados, o comportamento foi igualmente problemático.

Em paralelo, foi investigada uma estratégia híbrida utilizando o RVC [RVC Project 2023], um conversor de identidade vocal, sobre saídas do XTTS-v2. O RVC recebe qualquer áudio de entrada e converte o timbre para a voz treinada, com alta fidelidade quando a entrada é voz humana gravada ao vivo. A estratégia proposta era usar o TTS para gerar o conteúdo textual e o RVC para transferir a identidade tímbrica. Na prática, quando a entrada para o RVC era saída de TTS, a artificialidade da síntese era amplificada pelo conversor, não atenuada. Com voz humana como entrada, os resultados melhoraram, mas a configuração vincula o pipeline ao timbre de uma pessoa específica, comprometendo a reprodutibilidade.

Para o método fônico, a abordagem RVC apresenta um problema específico: a neutralidade articulatória necessária para fonemas isolados depende de que o locutor de entrada produza o fonema de forma precisa e descontextualizada. Qualquer variação de entonação ou coarticulação na voz de entrada é transferida para a saída pelo conversor, tornando a padronização fonética difícil. A abordagem é ainda menos indicada para o BITI do que para o OTTO, pois os requisitos de precisão fonêmica são mais exigentes.

4. Desafios Transversais

4.1. Síntese Emocional

Os enunciados emocionais do banco original apresentam perfis prosódicos clinicamente distintos. “Estou com sono” inclui a qualidade vocal característica de bocejo, produzida pelo abaixamento do véu palatino, ampla abertura faríngea e ressonância nasal aumentada, resultando em timbre simultaneamente nasalizado e distendido. Nenhuma das cinco ferramentas reproduziu essa característica de forma reconhecível. “Estou com medo” exige tensão vocal e aceleração de ritmo que os modelos de transferência de estilo atenuavam sistematicamente. Trata-se de uma limitação com implicação terapêutica direta: o robô terapêutico é utilizado para treino de reconhecimento emocional com crianças que apresentam dificuldade nessa área [Otto-Meyer et al. 2018], e um áudio com expressão emocional fraca ou incorreta pode ser pedagogicamente inerte ou, em casos de hipersensibilidade sensorial, perturbador.

O problema é estrutural. Modelos treinados sem dados emocionais anotados para PT-BR sintetizam prosódia próxima ao centro da distribuição de entonação por padrão, suficiente

para fala neutra mas insuficiente para os extremos necessários à expressão emocional convincente, conforme previsto pela literatura [Barakat et al. 2024].

4.2. Fonemas Sublexicais e Nasalização

O método fônico exige áudios de fonemas isolados (/b/, /d/, /f/, entre outros) para exercícios de associação grafema-fonema no BITI. Essa entrada viola as premissas dos modelos neurais de TTS: Jurafsky e Martin (2023) explicam que esses modelos assumem contexto sentencial e coarticulação entre segmentos adjacentes, de modo que um fonema consonantal isolado produz resultados imprevisíveis, incluindo inserção de vogal intrusiva, silêncio ou troca de idioma no tokenizador multilíngue.

A estratégia parcialmente eficaz desenvolvida consistiu em embutir o fonema-alvo em uma palavra-veículo de contexto controlado e extrair o segmento via edição em software de áudio (por exemplo, sintetizar “afta” para isolar /f/, ou “bala” para /b/ inicial). A técnica é aplicável a fricativas intervocálicas, mas falha para /R/ vibrante múltipla, /lh/ e /nh/, três fonemas que continuam dependendo de gravação humana suplementar. Nos modelos baseados em inglês (StyleTTS 2, Kokoro), as vogais nasais e o ditongo /ão/ acrescentaram uma segunda camada de dificuldade, parcialmente mitigada pelo dataset multilingual-phonemes-10k-alpha [StyleTTS2 Community 2024].

A Tabela 1 sintetiza os resultados comparativos das cinco ferramentas nas dimensões avaliadas.

Ferramenta	Execução	Emoção	Fon. PT-BR	Método Fônico	Treino (platô)
ElevenLabs	API nuvem (pago)	Prompt (fraca)	Médio	Inadequado	N/A
XTTS-v2	Local, GPU ≥8GB	Ref. (atenuada)	Médio	Limitado**	~15h
Piper TTS	Local, CPU/GPU	Neutro	Médio	Mais indicado**	~15h
StyleTTS2+ML*	Local, GPU ≥8GB	Difusão latente	Bom*	Potencial futuro	~15h
Kokoro-82M	Local, GPU ≥8GB	Difusão latente	Médio	Potencial futuro	<15h
F5-TTS+RVC	Local, GPU ≥8GB	Ref. (atenuada)	Médio	Não indicado	~15h

*Tabela 1. Comparação das ferramentas avaliadas. N/A = dado não sistematizado nesta fase. * Com dataset multilingual-phonemes-10k-alpha. ** Estratégia de palavra-veículo parcialmente eficaz. *** StyleTTS2+ML e Kokoro-82M correspondem a um único paradigma (difusão de estilo), reportado em linhas separadas por apresentarem resultados distintos.*

5. Discussão

Os resultados deste estudo têm implicações que ultrapassam o contexto específico do BITI. A dificuldade em sintetizar prosódia emocional e fonemas nasais em PT-BR não é uma peculiaridade do caso avaliado: é uma consequência previsível de modelos treinados predominantemente em inglês, sem anotação emocional e sem representação adequada dos fonemas do português. Pesquisadores que trabalham com qualquer aplicação de

síntese de fala em PT-BR com requisitos além da neutralidade prosódica enfrentarão o mesmo conjunto de limitações, independentemente da ferramenta escolhida.

A assimetria nos tempos de treinamento revela algo importante sobre o custo real de adoção dessas ferramentas. O XTTS-v2, que exigiu entre 24 e 72 horas, corresponde a um modelo com poucos dados em português no treinamento original, o que implica que mais tempo de adaptação é necessário para ajustar a distribuição aprendida a uma nova voz. StyleTTS 2 e Kokoro, com 1 a 3 horas, chegaram ao platô mais rapidamente não porque são melhores, mas porque a difusão de estilo tem uma dinâmica de adaptação diferente. O platô em 15 horas, comum a todos os modelos com treinamento iterativo, indica que o corpus de 170 enunciados esgota sua capacidade de generalização em poucas horas. Essa conclusão tem implicação direta: ampliar o corpus é mais produtivo do que estender o treinamento.

O efeito do dataset multilingual-phonemes-10k-alpha é o resultado mais acionável deste trabalho para a comunidade. Trata-se de um recurso público, de incorporação simples ao pipeline de fine-tuning do StyleTTS 2, que produziu melhoria audível sem re-treinamento completo. A hipótese mais plausível é que os fonemas nasais do português presentes no dataset ampliam o espaço fonético coberto pelo modelo, reduzindo a tendência de substituição por aproximantes ingleses. Esse mecanismo, se confirmado por ablação, teria implicação para qualquer grupo que trabalhe com síntese em línguas cujos fonemas nasais não estão representados no inglês. Do ponto de vista prático, o Piper emerge como a opção mais adequada para produção de bancos fonéticos estáveis no curto prazo, enquanto StyleTTS 2 e Kokoro representam a aposta mais promissora à medida que o suporte ao PT-BR amadurecer.

6. Conclusões

Este trabalho contribuiu com três achados principais: uma avaliação comparativa de cinco paradigmas de síntese por IA aplicados à voz infantil em PT-BR com análise de adequação ao método fônico, dimensão ausente na literatura; a demonstração de que o dataset multilingual-phonemes-10k-alpha incorporado ao StyleTTS 2 produz melhoria consistente na cobertura fonética do PT-BR, resultado reproduzível com baixo custo; e o diagnóstico de que os desafios de síntese emocional e fonêmica são estruturais ao campo, o que reorienta o esforço de pesquisa de migração de plataforma para construção de corpora anotados.

Em termos práticos: ElevenLabs é adequado apenas para prototipagem neutra rápida; XTTS-v2 oferece clonagem local de qualidade quando há tempo de GPU disponível; Piper TTS é a melhor opção atual para bancos fonéticos estáveis, pela reprodutibilidade e inferência sem GPU; StyleTTS 2 tem maior potencial para PT-BR no médio prazo; Kokoro-82M é eficiente quando o tempo de treinamento é crítico; F5-TTS com RVC é indicado apenas quando entrada humana ao conversor é viável e a consistência tímbrica entre locutores é aceitável.

7. Limitações

A avaliação foi conduzida por consenso qualitativo de equipe multidisciplinar, sem protocolo MOS formal, avaliação cega ou sistematização de taxa de aprovação por ferramenta, percentual de erro fonêmico ou escala numérica de avaliação emocional. Em particular, a melhoria fonética atribuída à incorporação do dataset multilingual-phonemes-10k-alpha foi observada por inspeção qualitativa da equipe, sem ablação controlada ou medições objetivas de antes/depois; trata-se de uma observação promissora que ainda carece de validação estatística. O corpus de treinamento está abaixo do ideal para qualquer um dos modelos avaliados; o looping mitiga parcialmente essa restrição, mas pode favorecer overfitting em vez de ampliar genuinamente a cobertura fonética. Todas as avaliações foram realizadas em um único contexto de aplicação; generalizações exigiriam validação independente.

8. Considerações Éticas

A clonagem da voz de um menor foi conduzida com aprovação do Comitê de Ética em Pesquisa com Seres Humanos da Universidade Federal de Mato Grosso (UFMT) e consentimento livre e esclarecido do locutor e de seus responsáveis legais. Todo processamento foi realizado localmente, sem transmissão de dados biométricos a servidores externos, em conformidade com a LGPD (com exceção dos testes estritamente necessários e anônimos na API comercial avaliada, conduzidos sob termos de serviço e privacidade adequados).

9. Declaração de uso de IA Generativa

Durante a condução e redação deste trabalho, ferramentas de Inteligência Artificial Generativa baseadas em Grandes Modelos de Linguagem (LLMs) foram empregadas de forma delimitada e exclusiva como ferramentas auxiliares para a revisão ortográfica e gramatical e para a padronização e organização das referências bibliográficas. O planejamento experimental, a execução metodológica, a análise crítica dos resultados comparativos e toda a autoria intelectual subjacente às ideias apresentadas neste manuscrito são de responsabilidade exclusiva dos autores humanos. O texto não contém dados empíricos forjados ou resultados sintetizados artificialmente por sistemas de IA.

Referências

Araújo, G. E. et al. (2026). Certas Palavras: A 1980s–90s Brazilian Radio Corpus to Test TTS Models in Noisy Multi-Speaker Dialogues. In Proceedings of PROPOR 2026. ACL Anthology.

Barakat, H.; Turk, O.; Demiroglu, C. (2024). Deep learning-based expressive speech synthesis: a systematic review of approaches, challenges, and resources. EURASIP

Journal on Audio, Speech, and Music Processing. DOI: 10.1186/s13636-024-00329-7.

Bott, L. et al. (2024). Controlling Emotion in Text-to-Speech with Natural Language Prompts. In Proceedings of Interspeech 2024. ISCA.

Casanova, E.; Júnior, A. C.; Shulby, C.; Oliveira, F. S.; Teixeira, J. P.; Ponti, M. A.; Aluísio, S. (2022). TTS-Portuguese Corpus: a corpus for speech synthesis in Brazilian Portuguese. Language Resources and Evaluation, 56, 1043–1055. DOI: 10.1007/s10579-021-09570-4.

Casanova, E. et al. (2022b). YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone. In Proceedings of ICML 2022. PMLR.

Casanova, E. et al. (2024). XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. arXiv:2406.04904.

Casanova, E.; Santos, V. G.; Svartman, F. R. F.; Leite, M. Q.; Candido Junior, A.; Marcacini, R. M.; Rezende, S. O.; Aluísio, S. M. (2026). Recursos para o processamento de fala. In Caseli, H. M.; Nunes, M. G. V. (orgs), Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português. 4ª ed., v.1. BPLN. DOI: 10.5281/zenodo.19259398.

Chen, X. et al. (2024). F5-TTS: A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching. arXiv:2410.06885.

Cho, J. et al. (2025). ECE-TTS: A Zero-Shot Emotion Text-to-Speech Model with Simplified and Precise Control. Applied Sciences, MDPI, 15(9), 5108.

ElevenLabs. (2024). Voice cloning and text-to-speech API. Disponível em: elevenlabs.io.

Gomes, E.; Pedroso, F. S.; Wagner, M. B. (2008). Hipersensibilidade auditiva no transtorno do espectro autístico. Pró-Fono, v. 20, p. 279–284.

Jurafsky, D.; Martin, J. H. (2023). Speech and Language Processing (3rd ed. draft). Stanford University.

Li, Y. A. et al. (2023). StyleTTS 2: Towards Human-Level Text-to-Speech through Style Diffusion and Adversarial Training with Large Speech Language Models. In Proceedings of NeurIPS 2023. arXiv:2306.07691.

Maia, R. et al. (2023). Prosódia e síntese da fala: uma revisão integrativa da literatura em português brasileiro. Revista da ABRALIN.

Oliveira, J.; Albuquerque, F. E. (2021). Leitura e escrita em crianças com autismo: o trabalho psicopedagógico a partir do método fônico. Facit Business and Technology Journal, 1(29).

Otto-Meyer, S. et al. (2018). Children with autism spectrum disorder have unstable neural responses to sound. Experimental Brain Research, 236, 733–743.

Rebouças, G. et al. (2021). Modelagem 3D do Robô Otto para o atendimento de crianças com Transtorno do Espectro Autista. In: Anais da XXI Escola Regional de Informática de Mato Grosso (ERI-MT), p. 35-41. Porto Alegre: SBC.

Scassellati, B.; Admoni, H.; Mataric, M. (2012). Robots for Use in Autism Research. Annual Review of Biomedical Engineering, v. 14, p. 275–294.

StyleTTS2 Community. (2024). multilingual-phonemes-10k-alpha. Hugging Face Datasets. Disponível em: huggingface.co/datasets/styletts2-community/multilingual-phonemes-10k-alpha.

[Software] Chen, X. et al. (2024). F5-TTS. GitHub: SWivid/F5-TTS. Disponível em: github.com/SWivid/F5-TTS.

[Software] Hexgrad. (2024). Kokoro-82M. Hugging Face. Apache 2.0. Disponível em: huggingface.co/hexgrad/Kokoro-82M.

[Software] OHF-Voice. (2025). Piper 1 — Fast and local neural text-to-speech engine. GitHub: OHF-Voice/piper1-gpl. GPL.

[Software] RVC Project. (2023). Retrieval-based Voice Conversion WebUI. GitHub: RVC-Project/Retrieval-based-Voice-Conversion-WebUI.