



UNIVERSIDADE FEDERAL DE SANTA CATARINA
CAMPUS TRINDADE
CURSO DE GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Lucas Ferreira da Rosa

Arquitetura Modular para Agente Conversacional de Ensino de Inglês

Florianópolis
2026

Lucas Ferreira da Rosa

Arquitetura Modular para Agente Conversacional de Ensino de Inglês

Trabalho de Conclusão do Curso de Graduação
em Engenharia Elétrica da Universidade Federal
de Santa Catarina para a obtenção do título de
Bacharel em Engenharia Elétrica.
Orientador: Prof. Márcio Holsbach Costa, Dr.

Ficha catalográfica gerada por meio de sistema automatizado gerenciado pela BU/UFSC.
Dados inseridos pelo próprio autor.

Rosa, Lucas Ferreira da
Arquitetura Modular para Agente Conversacional de
Ensino de Inglês / Lucas Ferreira da Rosa ; orientador,
Márcio Holsbach Costa, 2026.
66 p.

Trabalho de Conclusão de Curso (graduação) -
Universidade Federal de Santa Catarina, Centro Tecnológico,
Graduação em Engenharia Elétrica, Florianópolis, 2026.

Inclui referências.

1. Engenharia Elétrica. 2. Agentes Conversacionais . 3.
Computer-Assisted Language Learning. 4. Automatic Speech
Recognition. 5. Goodness of Pronunciation. I. Costa,
Márcio Holsbach. II. Universidade Federal de Santa
Catarina. Graduação em Engenharia Elétrica. III. Título.

Lucas Ferreira da Rosa

Arquitetura Modular para Agente Conversacional de Ensino de Inglês

Este Trabalho Conclusão de Curso foi julgado adequado para obtenção do Título de “Bacharel em Engenharia Elétrica” e aceito, em sua forma final, pelo Curso de Graduação em Engenharia Elétrica.

Florianópolis, 04 de março de 2026.



Documento assinado digitalmente
ROBERTO FRANCISCO COELHO
Data: 05/03/2026 18:17:23-0300
CPF: ***.034.249-**
Verifique as assinaturas em <https://v.ufsc.br>

Prof. Roberto Francisco Coelho, Dr.
Coordenador do Curso de Graduação em Engenharia Elétrica

Banca Examinadora:



Documento assinado digitalmente
Marcio Holsbach Costa
Data: 04/03/2026 15:37:34-0300
CPF: ***.170.680-**
Verifique as assinaturas em <https://v.ufsc.br>

Prof. Márcio Holsbach Costa, Dr.
Orientador
Universidade Federal de Santa Catarina



Documento assinado digitalmente
Eduardo Luiz Ortiz Batista
Data: 04/03/2026 17:11:55-0300
CPF: ***.521.889-**
Verifique as assinaturas em <https://v.ufsc.br>

Prof. Eduardo Luiz Ortiz Batista, Dr.
Universidade Federal de Santa Catarina



Documento assinado digitalmente
Aline Lemos Pizzio
Data: 05/03/2026 15:14:53-0300
CPF: ***.377.880-**
Verifique as assinaturas em <https://v.ufsc.br>

Prof.^a Aline Lemos Pizzio, Dr.^a
Universidade Federal de Santa Catarina

AGRADECIMENTOS

Agradeço primeiramente aos meus pais, Júlio e Vanessa, por todos os sacrifícios que fizeram para que eu chegasse até aqui e pelo apoio para ir ainda mais longe. São exemplos de obstinação, parceria e amor, e toda qualidade que tenho é apenas um reflexo deles.

À minha irmã Laura, minha primeira amiga, que se faz presente mesmo à distância.

À minha irmã Luisa, que o convívio diário tornou minha melhor amiga, pela parceria em todos os momentos e amadurecimento que tivemos morando juntos.

Aos meus avós, que enfrentaram batalhas muito diferentes das minhas, e se estou aqui hoje é porque as venceram.

À minha namorada Camille, que me motiva a ser uma pessoa melhor todos os dias, com quem olho junto na mesma direção.

Aos meus amigos de infância, adolescência e vida adulta, pelos momentos vividos, a companhia nesta trajetória e o constante esforço de seguirem presentes enquanto a vida passa cada vez mais acelerada.

Ao meu orientador Márcio Costa, pela orientação ao longo do trabalho, sempre equilibrando leveza e rigor metodológico.

Aos professores que passaram pela minha trajetória no Colégio Elisa Andreoli, por estimularem a curiosidade e a determinação da criança que fui. Aos que tive no IFSC, pelos inúmeros desafios teóricos e práticos da formação técnica. Aos da UFSC, pelo rigor acadêmico e pela sólida base analítica que ajudaram a construir. A essas instituições, agradeço pela estrutura que tive e oportunidades que vivi, em especial as internacionais em Portugal e China, que sem elas nunca teriam sido possíveis.

Às empresas em que trabalhei durante a graduação, pelas habilidades técnicas que aprendi, o crescimento individual que tive e as pessoas que conheci.

A todos que sempre acreditaram em mim, que são tantos que não caberiam nesta página.

“Pois, dentre todas as coisas que pertencem a todas as coisas, somente ao homem foi dado o dom de falar”
Dante Alighieri

RESUMO

A baixa proficiência em língua inglesa no cenário brasileiro demanda novas abordagens tecnológicas que promovam a prática oral autônoma. Este trabalho apresenta o desenvolvimento e a validação de uma arquitetura modular para um Agente Conversacional voltado ao ensino de inglês assistido por computador, projetado especificamente para falantes nativos de português. O objetivo principal consiste na integração de sistemas de Reconhecimento Automático de Fala, algoritmos de Treinamento de Pronúncia Assistido por Computador, Modelos de Linguagem de Larga Escala e Síntese de Fala para fornecer realimentação pedagógica em tempo real. A metodologia envolveu a constituição de um corpus de avaliação inédito, composto por segmentos de fala controlada e espontânea de brasileiros. Realizaram-se experimentos de benchmarking com modelos de estado da arte, resultando na seleção do Whisper Large v3 Turbo como motor de transcrição devido ao seu equilíbrio entre acurácia (WER/CER) e latência. O módulo de avaliação de pronúncia, implementado via Wav2Vec 2.0 e alinhamento forçado para cálculo da métrica Goodness of Pronunciation, identificou padrões de erro característicos do falante brasileiro, sobretudo em vogais complexas e ditongos. Para o gerenciamento do diálogo, utilizou-se o modelo Llama 3.2 com ajuste de persona, o qual demonstrou, através de testes de similaridade de cosseno, resiliência semântica frente a erros fonéticos de transcrição. A devolutiva sonora foi operacionalizada pelo modelo Kokoro, validado por métricas de preferência humana. Os resultados obtidos confirmam a viabilidade técnica da arquitetura proposta, evidenciando que a orquestração desses componentes permite a criação de um tutor virtual capaz de mitigar a ansiedade linguística e auxiliar no aprimoramento da fluência e pronúncia.

Palavras-chave: Agente Conversacional. Ensino de Inglês. Reconhecimento de Fala. Avaliação de Pronúncia. Inteligência Artificial.

ABSTRACT

The low proficiency in English within the Brazilian context demands new technological approaches that promote autonomous oral practice. This work presents the development and validation of a modular architecture for a Conversational Agent aimed at Computer-Assisted Language Learning, designed specifically for native Portuguese speakers. The main objective consists of integrating Automatic Speech Recognition systems, Computer-Assisted Pronunciation Training algorithms, Large Language Models, and Text-to-Speech to provide real-time pedagogical feedback. The methodology involved the creation of a novel evaluation corpus, composed of controlled and spontaneous speech segments from Brazilians. Benchmarking experiments were conducted with state-of-the-art models, resulting in the selection of Whisper Large v3 Turbo as the transcription engine due to its balance between accuracy (WER/CER) and latency. The pronunciation assessment module, implemented via Wav2Vec 2.0 and forced alignment for the calculation of the Goodness of Pronunciation metric, identified error patterns characteristic of Brazilian speakers, especially in complex vowels and diphthongs. For dialogue management, the Llama 3.2 model with persona tuning was used, which demonstrated, through cosine similarity tests, semantic resilience against phonetic transcription errors. The audio feedback was operationalized by the Kokoro model, validated by human preference metrics. The results obtained confirm the technical feasibility of the proposed architecture, evidencing that the orchestration of these components allows for the creation of a virtual tutor capable of mitigating linguistic anxiety and assisting in the improvement of fluency and pronunciation.

Keywords: Conversational Agent. English Teaching. Speech Recognition. Pronunciation Assessment. Artificial Intelligence.

LISTA DE FIGURAS

Figura 1 – Documentos que mencionam Reconhecimento Automático de Fala na plataforma Scopus, organizados por ano de publicação	21
Figura 2 – Exemplo de modelo de Markov com quatro estados e probabilidades de transição	22
Figura 3 – Diagrama esquemático do sistema proposto	34
Figura 4 – Desempenho dos modelos no reconhecimento de áudio - WER	40
Figura 5 – Desempenho dos modelos no reconhecimento de áudio - CER	41
Figura 6 – Convergência de métricas no reconhecimento de áudios do Speech Accent Archive	42
Figura 7 – Convergência de métricas no reconhecimento de áudios do TED	43
Figura 8 – Diagrama de blocos representativo do sistema de avaliação de pronúncia	45
Figura 9 – Distribuição de pontuação de pronúncia	46
Figura 10 – Distribuição de pontuação de pronúncia dos 20 principais fonemas no Speech Accent Archive	47
Figura 11 – Diagrama representando o experimento de propagação de erros no mecanismo de elaboração de resposta	49
Figura 12 – Relação entre tipo de Erro e Similaridade de Cosseno	51
Figura 13 – Tempo de processamento médio por etapa	54

LISTA DE TABELAS

Tabela 2 – Composição final do Corpus de Avaliação Brasileiro-Inglês	37
Tabela 3 – Resultados do Open ASR Leaderboard ordenados por WER em português	38
Tabela 4 – Comparativo de Eficiência Computacional entre modelos de ASR . . .	42
Tabela 5 – Estatísticas de Convergência das Métricas no <i>Speech Accent Archive</i> . .	43
Tabela 6 – Distribuição quantitativa dos tipos de erro no conjunto de dados (N=482)	49
Tabela 7 – Ranking de desempenho de modelos TTS (Open Leaderboard)	53
Tabela 8 – Estatísticas descritivas do desempenho de síntese de áudio (20 frases) .	53

LISTA DE ABREVIATURAS E SIGLAS

AC	Agente Conversacional
ASL	Aprendizado de Segunda Língua
ASR	Reconhecimento Automático de Fala
CALL	Aprendizado de Idiomas Assistido por Computador
CAPT	Treinamento de Pronúncia Assistido por Computador
CER	Taxa de Erro por Caractere
E2E	End-to-End
GMM	Modelo de Mistura Gaussiana
GOP	Goodness of Pronunciation
HMM	Modelo Oculto de Markov
IA	Inteligência Artificial
IPA	Alfabeto Fonético Internacional
LLM	Modelo de Linguagem de Larga Escala
LoRA	Low-Rank Adaptation
RTF	Fator de Tempo Real
RTFx	Fator de Tempo Real Inverso
TTS	Text-to-Speech
WER	Taxa de Erro por Palavra

SUMÁRIO

1	INTRODUÇÃO	13
1.1	OBJETIVOS	14
1.1.1	Objetivo Geral	14
1.1.2	Objetivos Específicos	14
2	FUNDAMENTAÇÃO TEÓRICA	15
2.1	APRENDIZADO DE IDIOMAS ASSISTIDO POR COMPUTADOR	18
2.2	SISTEMAS DE RECONHECIMENTO AUTOMÁTICO DE FALA	20
2.3	TREINAMENTO DE PRONÚNCIA ASSISTIDO POR COMPUTADOR	26
2.4	<i>LARGE LANGUAGE MODELS</i>	28
2.5	SISTEMAS DE CONVERSÃO TEXTO-FALA	32
3	PROPOSIÇÃO DO SISTEMA	33
3.1	CONSTITUIÇÃO DO CORPUS DE AVALIAÇÃO	35
3.2	RECONHECIMENTO DE FALA	37
3.3	AVALIAÇÃO DE PRONÚNCIA	44
3.4	MECANISMO DE ELABORAÇÃO DE RESPOSTA	48
3.5	RETORNO PARA O USUÁRIO VIA ÁUDIO	52
3.6	DESEMPENHO DO CONJUNTO	54
4	CONCLUSÃO	56
	REFERÊNCIAS	58
	ANEXO A – VÍDEOS UTILIZADOS NO <i>CORPUS</i> PERSONALIZADO	64

1 INTRODUÇÃO

A aprendizagem de línguas estrangeiras é um desafio complexo que vai além do domínio de vocabulário e gramática, exigindo também o aperfeiçoamento da fluência oral e da pronúncia. Para falantes de português interessados em progredir no inglês, obstáculos como acesso limitado a interlocutores nativos, insegurança na fala e dificuldade em perceber erros fonéticos podem comprometer a evolução consistente.

O cenário brasileiro reforça a importância de novas abordagens tecnológicas para o ensino do inglês. De acordo com o EF English Proficiency Index da World Population Review (2024), o Brasil ocupa a 70^a posição entre 113 países avaliados, evidenciando um índice baixo de proficiência. Mesmo entre aqueles que planejam ir ao exterior, a fluência é restrita: uma pesquisa de 2024 da assessoria Viva América (que presta serviços para quem quer morar nos EUA) mostra que menos de 8% dos clientes possuem inglês fluente, com apenas 13,6% se avaliando como avançados (VivaAmérica, 2024).

Nesse contexto, soluções tecnológicas que ofereçam realimentação imediata, personalizado e de alta qualidade tornam-se extremamente relevantes, estimulando uma prática mais autônoma, dinâmica e motivadora. Paralelamente, os avanços em inteligência artificial, incluindo reconhecimento de fala, compreensão contextual e geração de linguagem natural, possibilitam o desenvolvimento de ferramentas acessíveis e interativas. Esses sistemas, por sua vez, permitem maior personalização dos conteúdos e adaptações às necessidades específicas de cada aprendiz, tornando a experiência de estudo mais alinhada com os objetivos individuais e as diferentes trajetórias de aprendizado.

A proficiência em inglês abre um vasto leque de oportunidades, sobretudo no universo acadêmico. Como língua franca global, o inglês domina mais de 49% do conteúdo da internet, frente a apenas 3,9% em português, (W3Techs, 2024). Ao mergulhar nesse novo repertório linguístico, é possível ter acesso a ideias, conhecimentos e referências que ampliam a visão de mundo e a capacidade de inovação.

Nesse sentido, desenvolver um Trabalho de Conclusão de Curso em Engenharia Elétrica na área de Processamento de Sinais aplicado ao aprimoramento do aprendizado de idiomas, revela-se uma oportunidade singular. Trata-se de unir a área tecnológica a uma das habilidades mais valorizadas globalmente, contribuindo para soluções de impacto real na educação e na formação de cidadãos mais preparados para os desafios do mercado e da vida acadêmica internacional.

O objetivo principal do presente Trabalho de Conclusão de Curso é desenvolver e avaliar o protótipo de uma ferramenta interativa que auxilie falantes de português no aprimoramento da pronúncia, compreensão e fluência em inglês, integrando reconhecimento de fala, modelos de linguagem e realimentação automatizada em tempo real.

1.1 OBJETIVOS

Nas subseções a seguir estão descritos o objetivo geral e os objetivos específicos deste TCC.

1.1.1 Objetivo Geral

O objetivo principal do presente Trabalho de Conclusão de Curso é propor e validar uma arquitetura modular para um Agente Conversacional (AC) voltado para Aprendizado de Idiomas Assistido por Computador (CALL), do inglês *Computer Assisted Language Learning*, integrando Reconhecimento Automático de Fala (ASR) (do inglês *Automatic Speech Recognition*) para transcrição, algoritmos de detecção de erros de pronúncia, com suporte de um Modelo de Linguagem de Larga Escala (LLM) (do inglês *Large Language Model*) para realimentação e sistemas de Text-to-Speech (TTS) (do inglês *Text-to-Speech*) para interação oral.

1.1.2 Objetivos Específicos

- a) Mapear, através de revisão bibliográfica, as tecnologias de estado da arte em ASR, TTS e LLMs aplicáveis ao ensino de idiomas;
- b) Realizar experimentos de validação (benchmarking) com modelos de estado da arte, avaliando métricas de erro (WER/CER) e latência em bancos de dados de falantes não-nativos;
- c) Propor o diagrama arquitetural da solução completa, definindo os requisitos de sistema para o processamento do fluxo de áudio bidirecional, analisando o impacto da latência ao longo das etapas;
- d) Implementar um protótipo mínimo para o sistema proposto;
- e) Avaliar o desempenho do protótipo.

2 FUNDAMENTAÇÃO TEÓRICA

Todas as sociedades humanas possuem linguagem, e entende-se que sempre possuíram, não constituindo um fenômeno criado por algum grupo e posteriormente espalhado, tais como a agricultura ou o alfabeto (Pinker; Bloom, 1990). Esta capacidade é exclusiva da espécie humana, fundamentada em princípios restritivos profundos e enraizada em sua natureza biológica, como pode ser demonstrado pela velocidade com que crianças aprendem regras gramaticais complexas por mera exposição a adultos, televisão, leitura ou mesmo em brincadeiras com outras crianças (Chomsky, 1959).

O uso de linguagem é considerado uma adaptação biológica complexa, moldada pela seleção natural para preencher o que é descrito por Tooby e DeVore (1988) como um “nicho cognitivo”. Segundo os autores, enquanto predadores e presas desempenham uma coevolução antagônica contínua, os humanos desenvolveram a capacidade de abstrair situações, levando a sequências de comportamentos cognitivamente organizados em busca de objetivos que permitiram a superação das defesas naturais de presas e plantas. Enquanto estas defesas dependem de lentas alterações genéticas para evoluir, a cognição humana gera soluções comportamentais imediatas e inéditas. Segundo os autores, ao ocupar este nicho, os seres humanos desenvolveram uma série de características zoológicamente anômalas, entre elas a linguagem.

Segundo Tooby e DeVore (1988), a linguagem distingue-se de outros sistemas de comunicação animal por permitir trocas de informação baseadas em relações de causa e efeito. Além de facilitar a interação imediata, os autores indicam que a linguagem desempenha um papel econômico crucial na evolução hominídea ao permitir a transmissão de modelos causais complexos entre indivíduos e gerações. De tal forma, os autores colocam a linguagem como base para a cultura, transformando a informação em um recurso não rival que compensa os riscos inerentes à aprendizagem por tentativa e erro. Com isso, os aprendizados individuais compartilhados tornam-se patrimônio coletivo da espécie, e não apenas uma vantagem passageira de um indivíduo.

Segundo Chomsky (2006), o uso criativo da linguagem permite a qualquer falante produzir e compreender um número indefinido de expressões inéditas e apropriadas a novas situações, operando livre do controle de estímulos externos detectáveis. Enquanto visões anteriores à década de 1960 tratavam o aprendizado verbal como a formação de hábitos via reforço repetitivo (em uma visão chamada de behaviourista), Chomsky (1959) refutou a hipótese de que a aquisição da linguagem dependesse exclusivamente de um treinamento metódico ou da imitação repetitiva, argumentando que tais explicações, derivadas de experimentos com animais, falham ao tentar justificar a capacidade humana de produzir e interpretar sentenças originais e complexas que jamais foram ouvidas anteriormente. Em contrapartida, Chomsky (1959) destaca uma capacidade inata de generalizar, criar hipóteses e processar informações em uma variedade de maneiras, associada a observações

casuais e curiosidade natural como principais fatores para aquisição de linguagem.

O processo de aquisição de um idioma na infância é distinto daquele que ocorre na vida adulta. Krashen (1982) trata por aquisição o processo subconsciente de internalização linguística, enquanto a aprendizagem corresponde a um conhecimento consciente de gramática e regras linguísticas. Enquanto alguns linguistas consideram que as crianças “adquirem” e que adultos “aprendem”, o autor defende que a aquisição ocorre em qualquer etapa da vida, ainda que adultos nem sempre consigam atingir níveis nativos em um segundo idioma. A língua materna, por sua vez, é internalizada de forma inconsciente de tal forma que muitas vezes há apenas um sentimento de que algo “soa” gramaticalmente incorreto, sem noção de qual erro específico foi violado (Krashen, 1982).

No processo de aquisição de um idioma, há uma ordem previsível entre estruturas gramaticais, onde algumas são absorvidas pelo estudante antes de outras, fenômeno definido por Krashen (1982) como hipótese da ordem natural. Para estudantes de inglês, por exemplo, vê-se que a internalização do sufixo -s para plurais, ou o morfema -ing, equivalente ao gerúndio no português, ocorre antes do marcador /s/ visto na terceira pessoa do singular (como em “He lives in New York”). Tal sequência não é necessariamente a mesma entre aqueles que adquirem o primeiro e o segundo idiomas, nem deve servir de base para guiar o aprendizado, sendo apenas um resultado percebido no processo de aquisição (Krashen, 1982).

Enquanto aquisição e aprendizado coexistem como dois processos separados no aprendizado de uma segunda língua, ambos dão-se de forma distinta. Segundo Krashen (1982), enquanto a aquisição é responsável por iniciar o processo e internalizar o idioma rumo à fluência, o aprendizado de regras e formalidades de forma consciente permite avaliar e corrigir a fala e a escrita após sua execução, naquilo que é definido pelo autor como hipótese do Monitor. De acordo com tal lógica, pode-se utilizar conceitos e padrões aprendidos para corrigir-se, desde que haja tempo para pensar, foco na forma de se expressar e entendimento das regras. A exacerbação desse controle monitorado pode resultar em inibição linguística: ao priorizar a correção gramatical absoluta, o estudante torna-se mais hesitante e corrige-se no meio de falas, comprometendo a naturalidade e fluidez do discurso (Krashen, 1982). Por outro lado, o autor aponta que indivíduos que não aprenderam regras gramaticais ou abstêm-se de usá-las costumam guiar-se apenas pelo sentimento de que algo “soa” correto ou não, atendo-se completamente ao sistema de aquisição.

A partir da hipótese do Monitor, Krashen (1982) estabelece a aquisição de idiomas como central, colocando o aprendizado como periférico, levando a um foco pleno no primeiro em detrimento do segundo. Com isso, o autor formula a hipótese do *input*, que afirma que o progresso da competência ocorre apenas quando o aprendiz é exposto a conteúdos ligeiramente acima de seu nível atual, permitindo avanços graduais na aquisição. Dessa forma, a transição de um estágio i de aprendizado para outro mais avançado $i+1$

deve dar-se quando o estudante já entende o significado de *inputs* que contenham $i+1$, mesmo que pelo contexto ou informações extra-linguísticas.

Em paralelo à exposição a *inputs* mais complexos, Krashen (1982) reconhece que fatores emocionais como motivação, autoestima e ansiedade modulam a forma como o idioma é absorvido pelo indivíduo, de modo que ambientes de baixa ansiedade e alta motivação são mais eficazes para a aquisição natural da segunda língua. Tal fenômeno é descrito pelo autor como um filtro afetivo que potencializa ou inibe a aquisição do idioma. As hipóteses postuladas por Krashen (1982), baseadas em uma dicotomia entre aquisição (implícita) e aprendizado (explícito) foram posteriormente contestadas por Gregg (1984), mas seguem sendo bases para hipóteses posteriores quanto à aquisição de idiomas.

Em contraposição à hipótese do *input* está a noção de que, tão importante quanto a exposição a conteúdos do idioma que adquire-se é a prática por parte do estudante. Segundo Swain e Lapkin (1995), falar permite compreender as principais deficiências de vocabulário, indicando os principais pontos a serem corrigidos. Segundo os autores, ao expressar-se em uma segunda língua verbalmente, os aprendizes frequentemente deparam-se com problemas linguísticos que os levam a perceber lacunas em seu conhecimento, impulsionando os mesmos a modificar a estrutura de frases, promovendo um processamento mais profundo e consciente e acelerando o processo de aprendizado.

Outro aspecto relevante na aquisição de idiomas é a exposição a diferentes situações e o estímulo a interações variadas. Segundo Long (1981), modificações interacionais são o componente essencial para a absorção, por permitirem ao aprendiz identificar lacunas em seu conhecimento e receber retroalimentação imediata. A Hipótese de Interação, proposta por Long (1981), postula que o ponto central na aquisição de segunda língua não está nas modificações do conteúdo linguístico a que é exposto o aprendiz, mas sim na variabilidade de interações, acompanhadas de pedidos de esclarecimento, confirmação e reformulações.

As variações nos paradigmas de aprendizado de idiomas são amplamente associadas a transformações socioeconômicas regionais e globais. Com os avanços da chamada economia da informação, intensificou-se a necessidade de processamento de grandes volumes de dados e de competências interculturais, à medida que o mundo torna-se cada vez mais interconectado. Nesse cenário de densidade informativa, o papel do docente foi reconfigurado, dado que em um contexto de hiperconectividade, o professor deixa de ser o detentor exclusivo do conhecimento, e materiais literários tradicionais deixam de ser o único referencial de conhecimento legítimo no ensino de línguas estrangeiras (Warschauer; Healey, 1998).

Dentro deste contexto de hiperconectividade global, o aprendizado de idiomas tem rumado fortemente em direção ao CALL, cuja utilização proporciona inúmeros benefícios, sobretudo em testes automatizados e na promoção da autoaprendizagem (Witt; Young, 2000), estando presente tanto como suporte a relações tradicionais de ensino de idiomas, quanto em estudos individuais e *online*.

2.1 APRENDIZADO DE IDIOMAS ASSISTIDO POR COMPUTADOR

O Aprendizado de Idiomas Assistido por Computador constitui o campo dedicado ao estudo e à aplicação de tecnologias voltadas ao aprimoramento das quatro habilidades linguísticas: fala, audição, leitura e escrita. Sua fundamentação teórica busca superar as limitações do ensino tradicional ao tentar emular as condições de interação e fornecimento de *input* compreensível necessárias para o desenvolvimento da competência subconsciente em uma segunda língua, seja como ferramenta complementar ao ensino tradicional, seja como uma substituta e apoio à aprendizagem individual.

O entendimento de que computadores podem ser utilizados no aprendizado de idiomas vem desde os primórdios do desenvolvimento dos mesmos, com o desenvolvimento do CALL podendo ser dividido em três fases principais: o CALL behaviorista, o CALL comunicativo e o CALL integrativo (Warschauer; Healey, 1998). Visto que se desenvolveu atrelado ao ensino de idiomas, o campo evolui conforme as convicções de cada período, impulsionado tanto pelo progresso tecnológico quanto pelo entendimento das melhores práticas na aquisição de línguas.

A primeira fase do CALL, conhecida como behaviorista, concebida nos anos 1950 e implementada na década seguinte, fundamentava-se em teorias de aprendizagem que priorizavam a repetição e o reforço, utilizando o computador essencialmente como um tutor para exercícios, com grande foco em repetição (Warschauer; Healey, 1998). Com o abandono dessas visões, a evolução para o estágio comunicativo, entre as décadas de 1970 e 1980, marcou uma transição pedagógica em que o foco se deslocou do treinamento formal para o uso da língua em contextos distintos. Com o maior desenvolvimento de outros dispositivos além do computador pessoal e a popularização da internet, chegou-se à fase do CALL integrativo ao final década de 1990, onde buscava-se combinar diversas mídias e ferramentas de comunicação para criar um ambiente de aprendizado mais holístico e socialmente engajado.

Posterior à fase de CALL integrativo, a etapa atual de desenvolvimento do campo é caracterizada pelo advento e consolidação da Inteligência Artificial, que introduz níveis inéditos de interatividade e personalização ao processo de ensino. O processamento de volumes massivos de dados, viabilizado por avanços em modelos de LLMs e sistemas robustos de reconhecimento automático de fala, permite que as ferramentas tecnológicas ofereçam interações contextuais ricas e adaptadas às trajetórias individuais de cada aprendiz. Nesse cenário, a utilização de um AC torna-se um componente pedagógico central, funcionando como um interlocutor automatizado que simula conversações naturais, auxiliando na mitigação da ansiedade linguística e fornecendo realimentação corretiva imediata em domínios como a fluência e a pronúncia (Xiao et al., 2023).

De acordo com Xiao et al. (2023), os agentes conversacionais são definidos como sistemas de diálogo que simulam conversações naturais com seres humanos, utilizando Inteligência Artificial (IA) para analisar significado e intenção em falas e para gerar

respostas apropriadas a cada contexto. Segundo os autores, esse processo dá-se através de quatro etapas sequenciais: a transcrição da fala em texto, a determinação do sentido da mensagem, a geração da resposta textual e, por fim, a conversão do conteúdo de volta em fala, seja por voz sintetizada ou gravações prévias. Tal sistema pode ser utilizado em uma miríade de aplicações, algumas das quais serão abordadas neste material.

Ao analisarem estudos abordando o uso de agentes no processo de aquisição de idiomas, Xiao et al. (2023) identificaram um potencial significativo na aplicação dos mesmos como parceiros conversacionais, impulsionando o aprendizado tanto dentro quanto fora de sala de aula. Considerando uma ferramenta promissora, apesar de recente, sugerem que agentes conversacionais podem ser utilizados como ferramentas de diagnóstico, assistentes de sala de aula e instrumentos complementares para a prática individual. No entanto, os pesquisadores apontam que os professores também devem considerar os meios de promover o uso contínuo e de adaptar o agente ao seu contexto de ensino. Existem duas modalidades predominantes na aplicação pedagógica de ACs no ensino de idiomas: a prática geral de comunicação, que visa a interação aberta, e o aprendizado baseado em tarefas, que insere o agente em atividades com objetivos definidos (Xiao et al., 2023)

Uma das aplicações dos agentes conversacionais consiste no fomento à prática comunicativa aberta, proporcionando aos aprendizes a oportunidade de estabelecer diálogos livres no idioma estudado. Por meio de sistemas integrados de ASR, tais agentes são capazes de processar e reagir a *inputs* de fala espontânea, replicando a interação interpessoal e consolidando-se como parceiros virtuais para estudantes de idiomas (Xiao et al., 2023). Tal prática tem como um dos principais benefícios a redução de ansiedade de fala quando comparada à interação com seres humanos. Entretanto, os estudos analisados por Xiao et al. (2023) indicam dificuldades entre os estudantes de manter o uso de ACs fora do ambiente de sala de aula.

Outro meio de utilizar ACs no processo de Aprendizado de Segunda Língua (ASL) é sua implementação dentro de tarefas, onde para serem completadas, o estudante deve comunicar-se com o agente, em uma oportunidade de trazer ludificação ao aprendizado. Dentro deste contexto, Tai e Chen (2022), ao investigarem a influência de assistentes no aprendizado de estudantes de inglês como segundo idioma, concluíram que estes favoreceram significativamente a compreensão auditiva dos participantes, sobretudo quando comparados ao grupo de controle, que utilizou apenas *Compact Disks* (CDs), prática usual do ambiente escolar no que tange ao ensino de inglês. Segundo os autores, o uso de assistentes interativos e tarefas ludificadas promoveu flexibilidade e engajamento, auxiliando tanto no aspecto quantitativo, quanto na motivação dos estudantes.

A despeito do potencial dos ACs em promover um ambiente de prática imersivo e de baixa ansiedade, a implementação dessas ferramentas como instrumentos educacionais completos requer uma infraestrutura complexa, composta por diferentes blocos de processamento e resposta. Para que o uso desses assistentes resulte em ganhos de aprendizagem

comprováveis, é fundamental que o sistema consiga entender e interpretar a fala do usuário. Tal necessidade relaciona-se com as dificuldades de mensuração no processo de ASL.

Mensurar o avanço no aprendizado de uma segunda língua possui desafios significativos, dada a variabilidade de motivações, estilo de aprendizagem ou contexto entre indivíduos. Adicionalmente, a avaliação da proficiência oral enfrenta obstáculos inerentes à subjetividade do avaliador humano e a dificuldade de fornecer realimentação consistente e imediata em grande escala (Witt; Young, 2000). Para superar tais barreiras e permitir que sistemas computacionais processem, analisem e avaliem a produção oral dos aprendizes, torna-se importante a aplicação de tecnologias capazes de decodificar o sinal acústico de forma autônoma e convertê-lo em dados linguísticos estruturados, função desempenhada pelos sistemas de ASR (Levis; Suvorov, 2013).

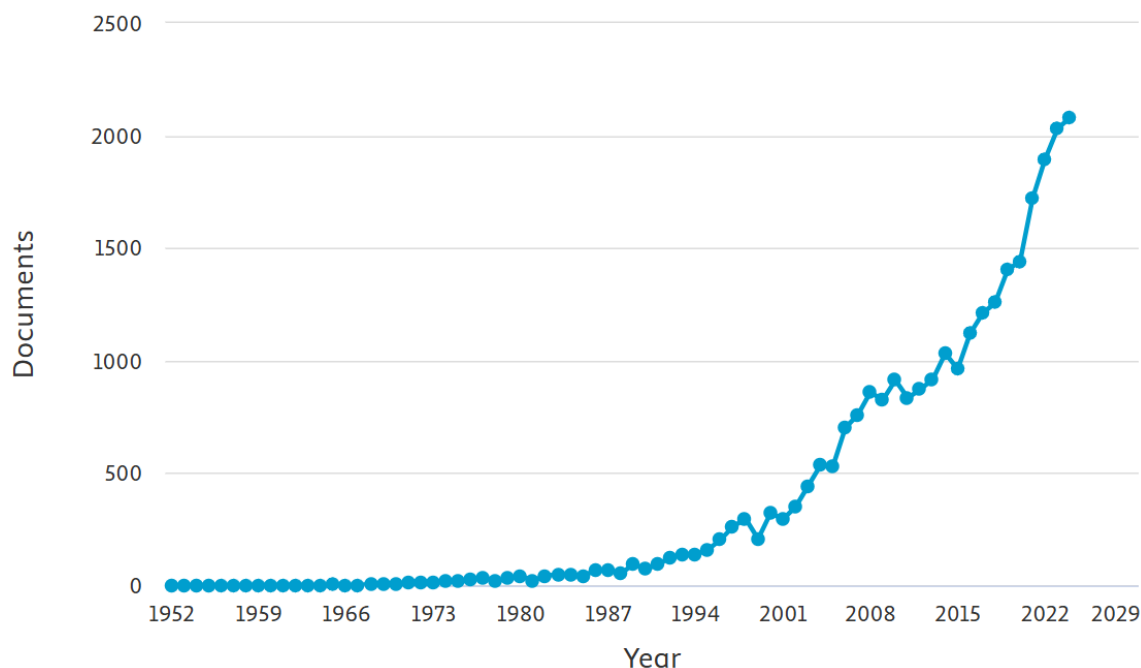
2.2 SISTEMAS DE RECONHECIMENTO AUTOMÁTICO DE FALA

Para que as potencialidades pedagógicas do CALL sejam efetivamente realizadas, especialmente no que tange à prática oral e à redução da ansiedade do aprendiz, torna-se fundamental a integração com sistemas robustos de processamento de sinais. Neste contexto, os sistemas de ASR atuam como o alicerce tecnológico que viabiliza a interação homem-máquina através da voz, permitindo que o computador analise a entrada acústica e produza uma saída em formato textual. No âmbito educacional, a tecnologia de ASR transcende a função de mera transcrição, sendo aplicada para identificar imprecisões na fala e oferecer realimentação instantânea, apoiando o desenvolvimento da pronúncia e da habilidade de fala de estudantes (Levis; Suvorov, 2013).

Os sistemas de ASR são processos responsáveis por decodificar e transcrever sinais sonoros de fala de forma autônoma (Levis; Suvorov, 2013). Para sua operação, o sinal sonoro de voz é captado por um microfone, convertido em sinais digitais e analisado tanto por modelos acústicos quanto por modelos de linguagem, de forma a identificar fonemas e, conseqüentemente, conectá-los para formar sequências inteligíveis no idioma-alvo, geralmente na forma textual (Lai; Karat; Yankelovich, 2003). Segundo Levis e Suvorov (2013), a implementação destes modelos possui diversos desafios associados, tais como: variabilidade da fala, reconhecimento de diferentes unidades (palavras e frases, sílabas, fonemas, etc), complexidade da linguagem, ambigüidade (em especial homófonos, fronteiras de palavras, ambigüidade sintática e semântica) e condições ambientais, como ruído de fundo ou várias pessoas falando simultaneamente.

Com o objetivo de superar os desafios supracitados, o desenvolvimento de estudos relacionados a sistemas de ASR tem crescido exponencialmente. A base de dados do Scopus indica mais de vinte e sete mil publicações entre 1952 e 2024, com um crescimento expressivo no século XXI, como expresso na Figura 1.

Figura 1 – Documentos que mencionam Reconhecimento Automático de Fala na plataforma Scopus, organizados por ano de publicação



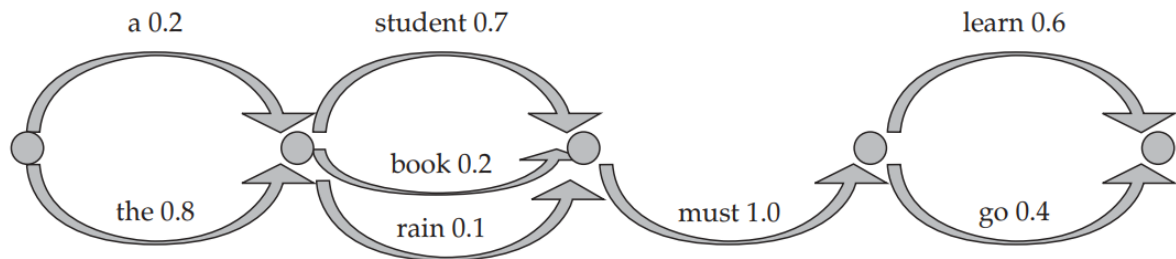
Fonte: Scopus (2025)

Os esforços iniciais apresentados nesse histórico foram guiados, em grande medida, pela teoria acústico-fonética, que buscava descrever como os sons básicos de uma língua se manifestam em um enunciado oral (Juang; Rabiner, 2004). Conforme a síntese histórica apresentada pelos autores, o primeiro sistema conhecido foi proposto em 1952 por Davis, Biddulph e Balashek, do Bell Labs, que desenvolveram um sistema de reconhecimento isolado de dígitos, utilizando frequências medidas nas regiões vocálicas de cada dígito em inglês. Na década de 1970, surgiu a primeira empresa focada em ASR, a Threshold Technology Inc., cujo sistema VIP-100 levou à criação do programa Speech Understanding Research (SUR) da ARPA. A partir da década de 1980, a evolução de métodos estatísticos culminou na adoção generalizada da estrutura de Modelo Oculto de Markov (HMM), associada a Modelo de Mistura Gaussiana (GMM).

O formalismo do HMM-GMM representou um avanço significativo em relação aos métodos acústico-fonéticos e de reconhecimento de padrões simples utilizados anteriormente em sistemas de reconhecimento automático de fala. Segundo Juang e Rabiner (2004), essa abordagem fundamenta-se na premissa de que, embora a estrutura linguística (pronúncia, sintaxe e gramática) seja muito semelhante, os sinais acústicos gerados por diferentes indivíduos ao proferirem a mesma palavra não são idênticos, apresentando variações consideráveis. Para tratar essa variabilidade, os modelos ocultos de Markov representam unidades linguísticas (como fonemas ou palavras) como uma sequência de estados com

probabilidades de transição entre cada um, tal como representado na Figura 2.

Figura 2 – Exemplo de modelo de Markov com quatro estados e probabilidades de transição



Fonte: Levis; Suvorov (2013)

Enquanto o HMM define quais estados são possíveis e estima a probabilidade de transição para o próximo estado, o GMM estima a probabilidade de o sinal de áudio atual ter sido gerado por um estado específico. Ou seja, no reconhecimento da palavra “casa”, o componente HMM é responsável pela estrutura temporal, indicando que após o estado do fonema /k/ há uma grande probabilidade de seguir-se o fonema /a/. Já o GMM analisa as características espectrais do fragmento de áudio atual e quantifica a probabilidade de aquele som corresponder, de fato, ao perfil acústico de um /a/.

Para operacionalizar essa lógica probabilística, sistemas clássicos baseados em HMM-GMM são estruturados em cinco componentes essenciais, conforme descrito por Nasereddin e Omari (2017):

- Extração de Características Acústicas: o sinal bruto proveniente do microfone, originalmente contínuo, é quantizado em janelas temporais fixas, o que viabiliza o processamento digital;
- Modelo de Linguagem: responsável por modelar estatisticamente as combinações plausíveis de palavras na língua-alvo, conferindo coerência semântica às sequências reconhecidas;
- Dicionário de Pronúncia: um repositório que relaciona cada vocábulo a uma ou mais representações fonéticas, fornecendo uma base de fonemas empregados pelos modelos acústicos;
- Modelo Acústico: associa os sinais acústicos aos fonemas do idioma, ajustando sotaques, dialetos e outras variações de contexto;
- Decodificador: núcleo decisório do ASR, no qual se estabelece a correspondência entre os vetores de características de cada quadro e os padrões previstos pelos demais módulos. A probabilidade de acerto aumenta à medida que mais quadros são analisados, permitindo o reconhecimento de padrões com mais facilidade.

Embora a estrutura de HMM-GMM tenha dominado o campo por décadas, a incorporação de Redes Neurais Profundas (DNNs) marcou uma significativa mudança de paradigma no campo. Inicialmente, as redes neurais foram introduzidas em arquiteturas híbridas (DNN-HMM), onde substituíam as GMMs na estimativa de probabilidades de emissão, mas ainda dependiam da estrutura de estados e alinhamentos dos HMMs (Jurafsky; Martin, 2024). Contudo, a evolução subsequente buscou eliminar a complexidade desses pipelines fragmentados em favor de arquiteturas End-to-End (E2E).

Ao contrário dos sistemas baseados em HMM, que exigem o treinamento separado de modelos acústicos, modelos de linguagem e dicionários de pronúncia para depois integrar todos os componentes, a abordagem E2E propõe o treinamento de uma única arquitetura neural capaz de mapear diretamente a sequência de entrada acústica para a sequência de saída textual (grafemas ou palavras), em uma unificação que simplifica o processo de engenharia por remover a necessidade de alinhamentos prévios e a construção manual de léxicos de pronúncia, permitindo que o modelo aprenda diretamente através dos dados brutos (Amodei et al., 2015).

O funcionamento fundamental das arquiteturas E2E contemporâneas baseia-se frequentemente no paradigma codificador-decodificador (encoder-decoder), também conhecido como sequence-to-sequence (Seq2Seq) (Jurafsky; Martin, 2024). Nesse modelo, o codificador processa o sinal de áudio, transformando-o em vetores de forma a extrair suas características e reduzir a taxa de amostragem, ajustando a grande quantidade de quadros do som ao formato mais compacto do texto (Jurafsky; Martin, 2024). Em seguida, o decodificador utiliza essas informações para gerar a transcrição.

Com a ascensão dos sistemas de aprendizado profundo, as etapas de codificação e decodificação eram realizadas por redes neurais recorrentes que processavam dados sequencialmente, tornando o treinamento lento e dificultando a compreensão de contextos longos. Para superar essas limitações, Vaswani et al. (2017) introduziram os Transformers ao substituir a recorrência pelo mecanismo de autoatenção (Self-Attention), permitindo que o modelo analise toda a sequência de áudio simultaneamente com paralelismo e capture relações entre trechos distantes da fala com muito mais eficiência.

Apesar de constantes avanços arquiteturais ao longo das décadas e um desenvolvimento significativo proporcionado pelos mecanismos de autoatenção, a tarefa de reconhecimento de fala permanece sujeita a desafios intrínsecos à variabilidade acústica e linguística. A complexidade da fala humana impede que a transcrição seja isenta de falhas, independentemente da robustez do modelo utilizado, trazendo à tona a necessidade de compreender e quantificar a acurácia desses sistemas, permitindo analisar a natureza das discrepâncias geradas.

O desempenho de um ASR está sujeito a três grupos de falhas, conforme apontado por Rodman (1999 apud Levis e Suvorov, 2012): erros no reconhecimento de fala discreta, erros no reconhecimento de fala contínua e erros na identificação de palavras (word

spotting). Dentro do grupo de erros de fala discreta, ocorrem casos de deleção, quando o sistema ignora um trecho devido à baixa intensidade da pronúncia; erros de inserção, quando ruídos são interpretados como unidades de fala; erros de substituição, quando um enunciado é identificado incorretamente (por exemplo, “mais” em vez de “mas”); e erros de rejeição, quando a palavra do locutor é recusada porque não integra o vocabulário do sistema. No grupo de erros de fala contínua manifestam-se os mesmos erros do grupo anterior, acrescidos de divisões, em que uma unidade de fala é reconhecida indevidamente como duas ou mais (por exemplo, “atualmente” por “a tua mente”); e fusões, quando duas ou mais unidades são percebidas como uma só (por exemplo, “a gente” reconhecido como “agente”). Por fim, quanto à identificação de palavras, registram-se falsos negativos, quando uma palavra presente no áudio não é detectada; e falsos positivos, quando uma palavra é identificada incorretamente.

De forma a mensurar o nível de erro de transcrições de áudio, a Taxa de Erro por Palavra (WER), do inglês *Word Error Rate*, é considerada como uma das principais métricas na avaliação de sistemas de ASR. Segundo Magalhães et al. (2022), deriva-se do cálculo de Distância de Levenshtein, métrica linguística que indica o mínimo de alterações necessárias para transformar uma sequência de palavras em outra. Através da WER, busca-se comparar uma transcrição de referência criada ou revisada por seres humanos com uma transcrição feita pelo modelo através da seguinte relação:

$$WER = \frac{S + D + I}{N}.$$

Onde S é o número de substituições (palavras erradas no lugar de corretas), D é o número de palavras ausentes (que estão na transcrição de referência, mas não na transcrição), I é o número de inserções de palavras extras, e N é o total de palavras na referência. Ou seja, normaliza a Distância de Levenshtein entre as duas transcrições, representando-a em termos percentuais. Por exemplo, ao comparar uma frase de referência “I want to eat pizza” e uma transcrição hipotética “I want eat a pizza”, tem-se:

I → I
 want → want
 to → (ausente, $D = 1$)
 eat → eat
 pizza → pizza
 a → (inserção, $I = 1$)

Com $N=5$, a WER resultante é de $(0+1+1)/5=0,4$, indicando que 40% da frase transcrita deve ser modificada para chegar à sequência correta de palavras.

Outra métrica disponível para avaliar modelos de ASR é a Taxa de Erro por Caractere (CER), do inglês *Character Error Rate*, defendida por Thennal et al. (2025)

como opção mais apropriada que a WER para modelos multilíngues. Segundo os autores, a WER é sensível demais a convenções ortográficas, a segmentações e à própria definição de “palavra”, em especial em idiomas onde não há espaçamentos claros entre palavras (como o mandarim, por exemplo), ou em casos em que há contrações ou variações morfológicas mais complexas. No português, por exemplo, palavras de mesmo sentido são escritas de formas diferentes a depender do contexto, como “para” e “pra”, ou “em um” e “num”, diferenças que aumentariam a WER de uma transcrição mesmo sem diferença semântica, apenas com pequenas variações ortográficas. Para atenuar esses casos, a CER é calculado de forma análoga à WER, mas em nível de caractere ao invés de palavra, o que indica resultados mais correlacionados à percepção de falantes nativos sobre quão corretas são as transcrições (Thennal et al., 2025). Utilizando a Distância de Levenshtein, a relação em nível de caractere é dada por:

$$CER = \frac{S_c + D_c + I_c}{N_c}.$$

Onde S_c , D_c , I_c representam as substituições, deleções e inserções de caracteres, e N_c é o número total de caracteres na referência. Retomando o exemplo da variação morfológica no português citado anteriormente, ao comparar a referência “para” com a transcrição “pra”, o impacto no WER seria de 100%, enquanto no CER o erro é proporcional à estrutura da palavra:

$$\begin{aligned} p &\rightarrow p \\ a &\rightarrow (\text{ausente}, D_c = 1) \\ r &\rightarrow r \\ a &\rightarrow a \end{aligned}$$

Com $N_c = 4$, a CER resultante é de $1/4 = 0,25$. Este resultado demonstra que, embora tenha havido uma omissão, 75% da estrutura de caracteres da palavra foi preservada, fornecendo uma métrica mais granulada e, frequentemente, mais justa para avaliar o desempenho de modelos em tarefas de transcrição de fala.

Além da qualidade de transcrição, é importante que um sistema de ASR seja capaz de operar com velocidade, especialmente em contextos de operação em tempo real, como em legendagem de videoconferências ou em ferramentas de acessibilidade voltadas a pessoas com deficiência auditiva. Para mensurar a rapidez de transcrição, utiliza-se o Fator de Tempo Real (RTF), do inglês *Real Time Factor*, que é calculado através da divisão do tempo de processamento do áudio pela duração do mesmo (Arriaga et al., 2025). Em um $RTF < 1$, o modelo consegue acompanhar o áudio, possibilitando o uso em tempo real. Caso $RTF > 1$, cada trecho de áudio demorará mais tempo para ser processado do que dura sua execução, gerando um atraso crescente e inviabilizando o uso em aplicações de tempo real. De forma análoga, utiliza-se muitas vezes o Fator de Tempo Real Inverso

(RTFx), inverso do RTF, que indica quantas vezes o processo de transcrição é mais rápido em relação à duração do áudio original.

Em síntese, a análise conjunta de WER, CER e RTF compõe o padrão de referência para a validação técnica de arquiteturas de reconhecimento de fala. Enquanto WER e CER mensuram a integridade da informação transcrita, o RTF valida a viabilidade de implementação em cenários de interação em tempo real. O equilíbrio entre esses indicadores permite definir a maturidade de um modelo no estado da arte, servindo como critério para escolha e classificações entre os melhores sistemas. Contudo, alcançar uma transcrição rápida e precisa é apenas um pré-requisito que baseia um sistema de ACs, havendo necessidade de outras tecnologias em conjunto.

Apesar de ser uma ferramenta com uma miríade de aplicações e uma robustez tecnológica em constante aprimoramento, os sistemas de ASR, *per se*, não configuram uma ferramenta pedagógica. Enquanto modelos de aplicações de ASR conseguem processar inúmeras variações de uma mesma palavra, o aprendizado de idiomas requer testes mais refinados de pronúncia, com heurísticas capazes de reconhecer e detectar quaisquer desvios daquilo que entende-se como pronúncia padrão (Shahin; Ahmed, 2024). Neste sentido, cabe um olhar mais apurado com relação a técnicas de detecção e diagnóstico de erros de pronúncia.

2.3 TREINAMENTO DE PRONÚNCIA ASSISTIDO POR COMPUTADOR

Diante da necessidade de transcender a mera transcrição textual e oferecer um suporte didático efetivo, consolida-se a área de Treinamento de Pronúncia Assistido por Computador (CAPT), do inglês *Computer Assisted Pronunciation Training*. Inserida no domínio mais amplo do CALL, essa vertente desloca-se da decodificação do que foi dito (conteúdo semântico) para a análise detalhada de “como” foi dito, focando em acurácia fonética. A partir de um texto de referência e do enunciado proferido pelo aprendiz, o sistema de avaliação de pronúncia deve determinar a correção da pronúncia e localizar trechos incorretos, fornecendo um ciclo de realimentação que auxilia no aprendizado do estudante (Tu et al., 2025). Tal avaliação difere-se daquela possível através de um ASR dado que, enquanto sistemas de reconhecimento de fala são capazes de ajustar pronúncias incorretas a partir do contexto (transcrevendo, por exemplo, “I hive a car” como “I have a car”), a checagem de pronúncia deve ser não-contextual, permitindo compreender deficiências na expressão no novo idioma.

A importância da avaliação de pronúncia faz-se relevante no contexto de ASL devido à influência de dois elementos: o princípio da natividade e o princípio da inteligibilidade, como enunciado por Levis (2005). Segundo o autor, o princípio da natividade defende que é possível e desejável alcançar uma pronúncia igual à de um falante nativo em uma língua estrangeira. Este paradigma, embora predominante até a década de 1960, perdeu força com o entendimento de que a natividade na pronúncia parece ser biologicamente

condicionada a ocorrer antes da idade adulta. Dessa forma, esperar pronúncia nativa de um não-nativo daquele idioma leva a um fardo irreal àqueles que buscam aprender um segundo idioma após a infância. Por outro lado, o princípio da inteligibilidade afirma que o objetivo do aprendiz é simplesmente ser compreensível, reconhecendo que a comunicação pode ser muito bem-sucedida mesmo quando o sotaque estrangeiro é perceptível ou forte.

Sob a ótica da inteligibilidade, portanto, o foco da análise desloca-se da eliminação do sotaque para a identificação de interferências da língua materna (L1) que possam comprometer a clareza da comunicação na língua alvo (L2). Essas interferências não são aleatórias, mas sim reflexos das restrições impostas pelo sistema nativo do aprendiz. No contexto específico de falantes de português aprendendo inglês, a distância fonológica entre os dois idiomas gera adaptações articulatórias previsíveis, que exigem atenção pedagógica dedicada para não se tornarem barreiras comunicativas.

Para falantes de português aprendendo inglês como segundo idioma, há desafios fonológicos derivados da diferença entre pronúncias de ambas as línguas. Cardoso (2008) destaca a dificuldade de produzir grupos consonantais iniciados por /s/ seguidos de outra consoante (como em *slip*, *snow*, *stop*), sequências inexistentes no português, que levam a introdução de uma vogal de apoio antes do /s/ (como “*eslip*”, “*esnow*”, “*estop*”). Ao investigar a aquisição desses fonemas por aprendizes brasileiros, o autor concluiu que a principal barreira não é o nível de exposição a cada fonema, e sim a complexidade articulatória de cada um, que contrapõe-se às articulações realizadas para a prática da língua portuguesa.

Tendo objetivos distintos do ASR padrão, a detecção de erros de pronúncia apresenta também métricas diferentes para auferir desempenho. A mais utilizada delas é o Goodness of Pronunciation (GOP), escore originalmente proposto por Witt e Young (2000) para avaliar a semelhança entre um fonema pronunciado e sua forma canônica esperada. A implementação prática desta medida envolve duas passagens de reconhecimento: uma via alinhamento forçado, que calcula a probabilidade do áudio corresponder exatamente ao texto alvo, e outra utilizando um laço de fonemas, que identifica a sequência acústica mais provável sem restrições gramaticais. A partir dessa comparação, o GOP mede a distância entre o “gabarito” e o fonema a ser avaliado, com valores próximos de zero indicando que o som produzido corresponde exatamente ao modelo acústico nativo esperado.

Para exemplificar o processamento do GOP, toma-se como cenário a pronúncia da palavra *think*, na qual frequentemente se observa a troca da fricativa dental /θ/ pela alveolar /s/ (como em “*caça*”). Supondo um áudio onde tal troca esteja presente, a primeira etapa, de alinhamento forçado, calcula a probabilidade do primeiro fonema ser efetivamente um /θ/. Como não foi o fonema enunciado, a verossimilhança é reduzida, gerando uma pontuação baixa. Simultaneamente, o laço de fonemas livre identifica o som efetivamente produzido — neste caso, o /s/, e atribui-lhe uma alta probabilidade de ocorrência. Subtraindo o primeiro do segundo, há uma alta disparidade, afastando o GOP

de zero e caracterizando o desvio fonético.

Para realizar o alinhamento entre o sinal de áudio e a transcrição fonética, utiliza-se o algoritmo de Viterbi. Conforme explicado por Forney (1973), este método funciona como uma busca por um caminho otimizado em um mapa de possibilidades conhecido como *trellis* (ou treliça). Ao invés de analisar todas as infinitas combinações possíveis de sons a cada milissegundo, o algoritmo descarta progressivamente as opções menos prováveis e mantém apenas a trajetória mais forte até aquele momento. Dessa forma, ao chegar ao final do áudio, o sistema consegue reconstituir a sequência única de estados que melhor explica o que foi dito, garantindo que o cálculo da nota de pronúncia seja feito sobre os trechos exatos onde cada fonema ocorre. Tal algoritmo, aliado a redes neurais profundas para reconhecimento dos fonemas, representa a base de muitos sistemas comerciais, que vêm sendo gradativamente substituídos por E2E com mecanismos de atenção.

Embora o GOP ofereça uma métrica quantitativa robusta para identificar a existência de falhas fonéticas, faz-se importante que seu uso venha acompanhado de uma devolutiva no contexto do AC, seja para correção, seja para entendimento do nível do estudante e ajuste de vocabulário por parte do Agente. Para preencher a lacuna entre a detecção do sinal acústico e a entrega de um retorno claro para o usuário, os próximos tópicos detalham como essa resposta é elaborada em texto e, em seguida, convertida em áudio para o usuário.

2.4 LARGE LANGUAGE MODELS

Com o objetivo de estruturar um sistema de aprendizado de idiomas interativo, faz-se necessária uma devolutiva para o usuário, ajustada a seu nível e ao contexto das interações, reafirmando as modificações interacionais mencionadas por Long (1981) como essenciais para o aprendizado de um segundo idioma. Neste sentido, os LLMs encaixam-se de forma apropriada, dado que constituem entes computacionais projetados para a interação conversacional, fundamentados na capacidade de prever sequências textuais a partir de um contexto prévio.

Segundo Jurafsky e Martin (2024), a essência dos LLMs reside em seu pré-treinamento, processo no qual o modelo é exposto a vastas quantidades de dados textuais para aprender a prever a continuidade de sequências. Para realizar essa tarefa, os modelos não operam diretamente sobre textos brutos, mas sim sobre *tokens*, que atuam como as unidades atômicas de entrada e saída da rede neural. Um *token* representa uma entrada de vocabulário finito e, dependendo do algoritmo de tokenização utilizado, pode corresponder a uma palavra inteira, a uma subpalavra ou apenas a um caractere. Dessa forma, o funcionamento fundamental de um LLM consiste em receber uma sequência desses elementos e calcular uma distribuição de probabilidade sobre todo o vocabulário disponível para determinar qual será o próximo *token*.

Arquiteturalmente, a maioria dos LLMs contemporâneos opera como modelos de

decodificação autorregressiva, inseridos no campo da Inteligência Artificial Generativa (Jurafsky; Martin, 2024). O funcionamento desses modelos baseia-se na intuição de que a capacidade de produzir *tokens* sequencialmente condicionados a um *prompt* inicial e às saídas subsequentes pode ser generalizada para resolver uma ampla gama de tarefas de Processamento de Linguagem Natural (PLN). Ao calcular a distribuição de probabilidade sobre o vocabulário possível e realizar a amostragem para selecionar o próximo termo, o modelo transforma a tarefa preditiva em um processo gerativo contínuo, ampliando a eficácia através de etapas posteriores de ajuste fino e alinhamento de preferências.

No processo de geração de texto, o comportamento do modelo não é determinado apenas pelos pesos aprendidos durante o treinamento, mas também por hiperparâmetros ajustáveis, como a temperatura. Segundo Jurafsky e Martin (2024), a intuição por trás dessa técnica é advinda da termodinâmica, onde um sistema em alta temperatura é mais flexível e suscetível a eventos menos previsíveis, enquanto em baixas temperaturas há uma tendência a convergir para estados estáveis e de menor energia. Na prática computacional, a amostragem com baixa temperatura aumenta suavemente a probabilidade dos *tokens* mais prováveis e diminui a dos raros, resultando em saídas determinísticas e gramaticalmente seguras. Em contrapartida, temperaturas mais altas alteram a distribuição de probabilidade de forma a permitir a escolha de termos menos frequentes.

No âmbito do processamento de linguagem natural, as arquiteturas de modelos dividem-se em três tipos de acordo com o fluxo de informação: decodificadores (decoders), codificadores (encoders) e codificadores-decodificadores (encoder-decoders). Os decoders são modelos generativos que recebem tokens como entrada e geram novos tokens sequencialmente, sendo a base de LLMs modernos, focados na previsão do próximo termo. Em contraste, os encoders processam uma sequência de entrada inteira para produzir uma representação vetorial (codificação) de cada token, sendo ideais para tarefas de classificação e compreensão, mas não para a geração de texto. Por fim, os encoder-decoders combinam essas duas estruturas para processar uma sequência de entrada e gerar uma sequência de saída distinta, permitindo uma relação flexível entre elas, o que os torna a arquitetura padrão para tarefas de transdução como a tradução automática (Jurafsky; Martin, 2024).

Independente da aplicação, LLMs modernas possuem Transformers como arquitetura padrão. A mesma representa uma ruptura com os modelos dominantes anteriormente, que tradicionalmente dependiam de redes neurais recorrentes ou convolucionais complexas contendo codificadores e decodificadores. Segundo Vaswani et al. (2017), o Transformer fundamenta-se inteiramente em mecanismos de atenção chamados de *self attention*, que dispensam a recorrência e as convoluções, em uma abordagem que permite que o modelo capture dependências globais entre entrada e saída de forma mais eficiente e viabiliza uma paralelização significativamente maior, reduzindo o tempo de treinamento necessário para alcançar o estado da arte em tarefas como tradução automática. Segundo Wang (2023), o mecanismo de self-attention processa a sequência inteira simultaneamente, gerando uma

nova representação onde cada elemento é uma soma ponderada de todos os outros. Para isso, o modelo utiliza projeções lineares de consulta (query), chave (key) e valor (value) para calcular a relevância entre cada par de posições. Essa abordagem supera limitações de processamento local, permitindo capturar dependências de longo alcance de forma eficiente e paralelizável.

A adaptabilidade e versatilidade das LLMs permitem que sejam utilizados no âmbito do aprendizado de idiomas. Lee et al. (2025), ao realizar uma revisão sistemática de quarenta e nove estudos empíricos publicados entre 2023 e 2024, abordaram uma lacuna na literatura ao focar em pesquisas baseadas em sala de aula. Os resultados indicaram que o ensino superior foi o cenário predominante, com o inglês figurando como a principal língua-alvo, sendo o foco de quarenta e seis estudos. Nestes, investigou-se principalmente as percepções dos aprendizes — tais como atitude, autoeficácia e motivação — enquanto os estudos sobre aquisição de linguagem concentraram-se primariamente nas habilidades de escrita.

No contexto do aprendizado de idiomas, LLMs possuem o potencial de desempenhar papéis complexos, atuando como provedoras de realimentação, tutoras de aprendizagem, estimuladoras cognitivas e parceiras de conversação (Lee et al., 2025). Contudo, para que esses modelos saiam de modelos genéricos de predição probabilística de palavras para agentes instrucionais efetivos, faz-se necessário o processo de *instruction tuning*, ou ajuste de instrução. Conforme detalhado por Bergmann (2025), modelos puramente pré-treinados não são otimizados para seguir comandos, tendendo apenas a completar sentenças de forma coerente, muitas vezes sem utilidade prática para o usuário. O ajuste de instrução supera essa barreira ao treinar o modelo com pares de entradas e saídas desejadas, permitindo que o sistema generalize a capacidade de seguir diretrizes pedagógicas e interaja de maneira mais direcionada.

Apesar dessa capacidade de adaptação técnica, a integração prática dessas tecnologias no ensino revela desafios multifacetados. Do ponto de vista tecnológico, persistem questões referentes à qualidade do conteúdo e problemas de consulta, uma vez que a natureza generativa dos modelos pode levar a imprecisões factuais ou "alucinações". Em paralelo, emergem desafios educacionais e éticos, incluindo o risco de dependência excessiva por parte dos estudantes, limitações na profundidade pedagógica das interações e preocupações com a integridade acadêmica (Lee et al., 2025). Portanto, a implementação de LLMs como ferramentas de suporte ao ensino exige um equilíbrio cuidadoso entre a automação e supervisão contínua.

A evolução dos LLMs tem sido um marco significativo no campo da Inteligência Artificial, com mudanças transformacionais em vários domínios (Ding et al., 2024). Em contrapartida, essa rápida expansão traz desafios, incluindo no que tange à sustentabilidade ambiental e eficiência de recursos. Embora essas arquiteturas apresentem capacidades cognitivas e generativas sem precedentes, sua implementação exige uma infraestrutura

computacional massiva, tornando importante analisar os fatores que regem o consumo de energia nessas arquiteturas, explorando a relação entre a complexidade algorítmica e o impacto ecológico, visando viabilizar uma inteligência artificial que seja tecnicamente robusta e ambientalmente responsável.

Segundo Ji e Jiang (2026), a sustentabilidade energética associada às LLMs não está apenas ligada a seus desenvolvedores, mas também a diversos outros agentes ao longo da cadeia de valor, incluindo centros de processamento de dados, operadores da rede elétrica e fabricantes de hardware, que compartilham a responsabilidade pelo aprimoramento da eficiência do setor. Portanto, a mitigação do impacto ambiental da inteligência artificial exige um esforço colaborativo e coordenado entre todos os integrantes desse ecossistema tecnológico, de forma a seguir avançando no tema. De forma a medir eficiência de uma LLM, inúmeros critérios podem ser utilizados, sendo seis os elencados por Ding et al. (2024):

- a) Número de parâmetros: atrelado aos pesos e vieses ajustados durante o treinamento do modelo, que permitem que o mesmo capture padrões nos dados a que é exposto. Possui como contrapartida uma maior complexidade do modelo, bem como *overfitting*, um sobreajuste que ocorre quando há uma adaptação excessiva aos dados de treinamento;
- b) Tamanho do modelo: refere-se ao espaço em disco em gigabytes (GB) ou megabytes (MB) necessário para armazenamento do modelo pré-treinado, constituindo fator determinante para a implementação em ambientes com restrições de espaço;
- c) Operações de Ponto Flutuante (FLOPs): contabilização de operações de ponto flutuante (como adição, subtração, multiplicação e divisão) que o modelo precisa executar para performar uma decisão. Apesar de não ser uma métrica absoluta e depender de outras variáveis como a arquitetura computacional utilizada, é uma métrica utilizada para estimar a complexidade de um LLM;
- d) Tempo de inferência: mede a latência prática e a velocidade de resposta em ambientes reais, sendo extremamente relevante para aplicações reais por estimar o tempo de processamento entre uma entrada e a saída do modelo;
- e) Ocupação de memória: medida em MB ou GB, expressa a necessidade de memória de acesso rápido (RAM) exigida para carregar e executar o modelo, considerando também variáveis intermediárias e estruturas de dados;
- f) Emissão de carbono: avalia o impacto ambiental de treinar e executar um modelo, sendo medida em quilogramas ou toneladas de carbono, estando atrelado inclusive às fontes de energia elétrica utilizada para treiná-lo, o que torna este critério o mais relacionado a fatores sociais e éticos do uso de LLMs.

Em suma, a análise criteriosa desses parâmetros de eficiência é determinante para a viabilidade de sistemas de aprendizado baseados em LLM. O desafio central reside em encontrar um equilíbrio entre complexidade computacional e eficiência de recursos. Assim, a escolha e a otimização da arquitetura do modelo não constituem apenas decisões técnicas de engenharia, mas pré-requisitos essenciais para garantir que a ferramenta pedagógica seja acessível e ambientalmente responsável, sem comprometer a qualidade da geração de texto.

2.5 SISTEMAS DE CONVERSÃO TEXTO-FALA

Os sistemas de TTS, ou conversão texto-fala, são tecnologias computacionais projetadas para sintetizar a fala humana a partir de um texto de entrada, fazendo com que também sejam conhecidos como sistemas de síntese de fala (Taylor, 2009). O objetivo fundamental de um sistema TTS é gerar um áudio inteligível e, idealmente, indistinguível da fala produzida por um ser humano.

A conversão de texto em algo que compreende-se como fala enfrenta desafios significativos, sobretudo devido às naturezas distintas de ambos. Conforme discute Taylor (2009), o texto de entrada não deve ser visto apenas como uma transcrição da fala, mas sim como um sinal visual ruidoso que precisa ser decodificado. Diferente da fala, que é contínua, a escrita contém elementos não linguísticos como números, abreviações e símbolos matemáticos, que não possuem uma correspondência fonética direta e imediata. O sistema, portanto, precisa realizar uma normalização para recuperar a forma linguística abstrata antes de qualquer geração de áudio. Além disso, elementos como entonação, emoção e ritmo não são completamente representados através da escrita, algo que para um leitor humano é inferido a partir da compreensão do significado do texto, mas implica em desafios significativos para um sistema de TTS. Sem uma interpretação correta, a leitura em voz alta resultante pode soar monótona ou artificial (Taylor, 2009).

Um sistema TTS moderno é composto, em essência, por três módulos fundamentais: um analisador de texto, um modelo acústico e um vocoder. Segundo Tan et al. (2021), o analisador textual é a “linha de frente” dos sistemas TTS, sendo responsável por transformar o texto de entrada em atributos linguísticos que codificam informações relevantes sobre pronúncia e prosódia, facilitando o processo de síntese da fala. Neste processo, há uma série de tarefas menores, como normalização textual, segmentação de palavras e previsão de prosódia. Os atributos linguísticos gerados pelo analisador de texto são entregues a modelos acústicos, que os transformam em formas de onda, que por sua vez são processadas pelo vocoder, responsável pela síntese da voz.

3 PROPOSIÇÃO DO SISTEMA

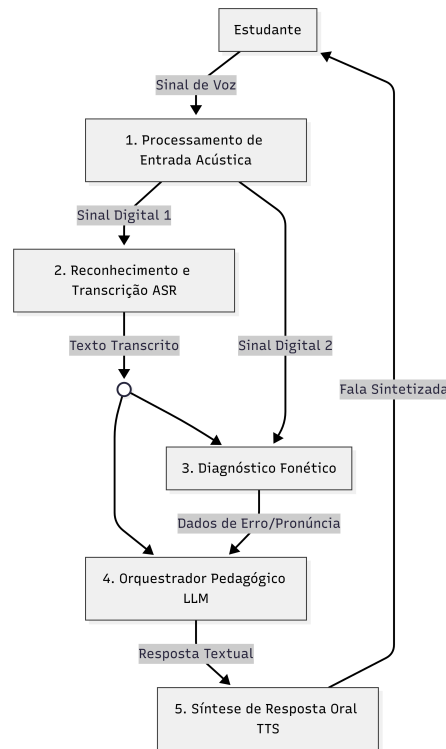
A revisão bibliográfica apresentada evidencia que a aquisição de uma segunda língua é um processo multidimensional, dependente não apenas de exposição a conteúdos compreensíveis (dentro da Hipótese do *Input*), mas também de oportunidades de produção oral (*output*) em ambientes de baixa ansiedade. Para operacionalizar os conceitos teóricos de Krashen (1982) e Long (1981) em uma ferramenta tecnológica, faz-se necessária uma engenharia que integre a flexibilidade conversacional dos ACs com algum rigor avaliativo.

Diante deste cenário, o desenvolvimento do sistema modular de ensino deu-se pela necessidade de atuar como um interlocutor empático através de LLMs e TTS para reduzir o filtro afetivo definido por Krashen (1982), e como um monitor acurado (via ASR e algoritmos de alinhamento) para garantir a inteligibilidade e acompanhamento do estudante. A seguir, detalham-se os requisitos e a arquitetura propostos para orquestrar essas tecnologias. A estruturação do sistema proposto fundamenta-se na decomposição de funcionalidades essenciais para um ambiente de CALL, com requisitos que foram divididos entre funcionais, que descrevem o comportamento do sistema, e não funcionais, que estabelecem os critérios de desempenho e qualidade técnica. São requisitos funcionais:

- a) Processamento de entrada acústica: o sistema deve realizar a captura do sinal de voz, sua digitalização e a extração de características acústicas para viabilizar o processamento digital;
- b) Reconhecimento e transcrição: a arquitetura deve decodificar o sinal sonoro de forma autônoma, convertendo-o em uma saída em formato textual inteligível;
- c) Diagnóstico fonético: comparar o enunciado do aprendiz com uma sequência fonêmica canônica para localizar desvios de pronúncia;
- d) Orquestração de realimentação pedagógica: através de LLMs, o sistema deve gerar respostas contextuais;
- e) Síntese de resposta oral: o sistema deve converter a realimentação textual gerada em fala sintetizada para manter o fluxo de interação oral bidirecional.

A consolidação destes requisitos funcionais materializa-se na arquitetura do sistema proposto, cujos componentes operam de maneira sequencial e integrada para processar o ciclo comunicativo. O diagrama de blocos da Figura 3 esquematiza essa organização, evidenciando as interfaces de comunicação e o fluxo de dados entre os módulos de entrada, processamento central e saída.

Figura 3 – Diagrama esquemático do sistema proposto



Fonte: Autor

A implementação dos requisitos funcionais não assegura, por si só, a efetividade pedagógica da ferramenta. Para que a interação se assemelhe a um diálogo natural e cumpra o objetivo de manter o filtro afetivo baixo, a execução dessas tarefas deve obedecer a restrições operacionais. Falhas de desempenho, como uma latência excessiva entre a fala do usuário e a resposta do sistema, podem quebrar a imersão comunicativa e gerar frustração, invalidando os benefícios teóricos do modelo. Nesse sentido, estabelecem-se os requisitos não funcionais, que delimitam os parâmetros de qualidade, desempenho e confiabilidade necessários para a viabilidade da solução:

- a) Operação em tempo real: para garantir a naturalidade da interação, o processamento deve apresentar uma latência reduzida, operando com um fator de tempo real $RTF < 1$;
- b) Acurácia de transcrição: o desempenho do módulo ASR deve ser mensurado através de WER e CER, buscando a minimização de erros de inserção, substituição e deleção;
- c) Confiabilidade do escore de pronúncia: a validação fonética deve basear-se na métrica GOP, fornecendo uma estimativa de verossimilhança estatística entre a fala do usuário e o padrão esperado;
- d) Modularidade e escalabilidade: a arquitetura deve ser modular, permitindo que

componentes individuais sejam atualizados ou substituídos sem comprometer a integridade do fluxo de dados;

Com o estabelecimento destes requisitos, consolida-se o alicerce técnico necessário para a concepção da arquitetura e seleção das tecnologias que integrarão a solução. Considerando a natureza heterogênea dos modelos envolvidos e a exigência de tempos de resposta rigorosos, a metodologia de desenvolvimento adotada nas seções subsequentes segue uma abordagem de validação por componentes. Esta estratégia permite isolar e verificar o desempenho de cada módulo individualmente, assegurando que cada etapa do *pipeline* cumpra os critérios de qualidade estabelecidos antes da integração final do sistema.

3.1 CONSTITUIÇÃO DO CORPUS DE AVALIAÇÃO

Considerando o objetivo de criar um Agente Conversacional para ensino de inglês a brasileiros, faz-se essencial entender como os principais modelos de ASR operam quando expostos a áudios de brasileiros expressando-se em língua inglesa nos mais diversos níveis. No entanto, a literatura não é vasta no que tange a *corpus* bem estabelecidos de “inglês brasileiro”, havendo maior vastidão entre materiais puramente em português, ou então com nativos de língua inglesa. Entre as bases de dados comumente utilizadas, há exemplos análogos ao Librispeech, composto de áudio livros entoados por narradores profissionais, nativos do idioma e com boa qualidade sonora, cenário diferente daquele encontrado em contextos de prática e ASL. Logo, para melhor avaliar os modelos de Reconhecimento de Fala frente ao contexto de um Agente Conversacional, optou-se por criar um *corpus* personalizado para fins de teste.

A construção do banco de dados para avaliação de modelos de ASR em falantes brasileiros de inglês surge da falta de materiais do tipo na literatura. Um dos raros exemplos existentes é o *Speech Accent Archive*, projeto que reúne gravações padronizadas de falantes nativos e não nativos de inglês, cada um lendo o mesmo parágrafo, com o objetivo de documentar e exibir uma grande variedade de sotaques e perfis demográficos (Weinberger, 2005). Funcionando como ferramenta de ensino e de pesquisa, permite comparar sotaques, analisar quais variáveis influenciam suas características e mostrar que sotaques são sistemas regulares, não erros de fala. Para falantes nativos de português, o banco de dados contém quarenta e oito segmentos totalizando aproximadamente 21 minutos, onde os voluntários leem o seguinte trecho:

Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station. (Weinberger, 2005)

Buscando uma maior variedade de contextos, optou-se pela construção de um pequeno corpus de teste baseado em vídeos de palestras públicas (como *TED Talks* e *Talks at Google*) disponíveis na plataforma YouTube, já previamente legendadas por voluntários. Como muitas dessas legendas disponíveis publicamente nem sempre possuem alinhamento temporal preciso com o áudio, foi desenvolvido um pipeline de validação automatizada para garantir a integridade dos segmentos. Compilou-se 14 vídeos diferentes, totalizando 234 minutos de áudio, que posteriormente foram reduzidos após processamento e filtragem.

O processamento dos dados brutos iniciou-se com a extração do áudio e sua segmentação com base nos timestamps das legendas oficiais, resultando em um conjunto inicial de 2215 segmentos candidatos. Contudo, legendas públicas frequentemente sofrem de desalinhamentos temporais ou simplificações linguísticas, o que levou à implementação de um algoritmo de validação em três estágios, desenvolvido em Python utilizando a API do Google Drive para gestão dos dados.

Adotou-se uma estratégia de transcrição de controle, submetendo cada segmento ao modelo OpenAI Whisper *tiny* (Radford et al., 2022). A escolha desta versão deve-se à sua alta velocidade de inferência e ao fato de que, para fins de validação, não se exige perfeição: se um modelo leve consegue transcrever o áudio de forma similar à legenda, infere-se que o áudio possui clareza acústica. As transcrições geradas foram comparadas às legendas de referência após um processo de normalização de texto, calculando-se a CER. Segmentos com CER elevado foram marcados para revisão.

O critério de corte estabelecido foi de um CER de 0.05, com segmentos que apresentaram uma taxa de erro inferior a esta sendo aprovados automaticamente devido à alta correlação entre o áudio e o texto de referência. Segmentos com CER superior a este limiar foram preliminarmente descartados, sob a premissa de que a divergência indicava desalinhamento temporal, baixa qualidade de áudio ou erro grosseiro na legenda original. Mesmo considerando que o ASR comete erros de transcrição, optou-se por um maior rigor quanto ao corte, de forma a evitar erros crassos de correspondência entre o áudio e a legenda associada.

Diferentemente de abordagens que descartam automaticamente segmentos com alto erro, implementou-se uma etapa de recuperação assistida por IA, utilizando o modelo Claude 3 Haiku, da empresa Anthropic (Anthropic, 2024). Através de engenharia de *prompt*, o modelo foi instruído a atuar como um avaliador semântico, distinguindo erros de transcrição aceitáveis de disparidades críticas, como por exemplo a presença de elementos explicativos da cena, como “Applause”. Esta etapa permitiu recuperar segmentos que seriam descartados por métricas puramente estatísticas como o CER, mas que eram pedagogicamente válidos.

Por fim, os segmentos filtrados pelas etapas automáticas passaram por uma verificação humana final para assegurar a ausência de cortes abruptos ou áudios mudos. O banco de dados resultante desse funil de qualidade totalizou 624 segmentos de alta fidelidade que,

somados aos dados do *Speech Accent Archive*, constituem um *corpus* robusto e balanceado, contendo tanto a fala lida e controlada quanto a fala espontânea de apresentações públicas.

Tabela 2 – Composição final do Corpus de Avaliação Brasileiro-Inglês

Sub-corpus	Fonte	Tipo de Fala	Segm.	Duração
Speech Accent Archive	Leitura Controlada	Script fixo	48	≈ 21 min
TED/Talks Personalizado	Apresentações Públicas	Espontânea/Planejada	624	≈ 35 min
Total	—	Misto	672	≈ 56 min

Fonte: O autor (2026)

3.2 RECONHECIMENTO DE FALA

Considerando-se a vertiginosa ascensão de modelos de ASR, há uma série de estudos visando à categorização e criação de ranqueamentos para os mesmos, com o objetivo de avaliar o desempenho em termos de acerto e latência. Dentre eles, Srivastav, Zheng et al. (2025) propõe o *Open ASR Leaderboard*, sistema reproduzível com a cobertura de mais de 60 sistemas abertos e proprietários, aplicando-os a 11 bancos de dados diferentes, seja em inglês, seja em outros idiomas.

Para assegurar a isonomia nas comparações entre modelos que possuem diferentes convenções de pontuação e capitalização, o *Open ASR Leaderboard* implementa um pipeline padronizado de normalização de texto, inspirado no modelo Whisper, que inclui a conversão de números por extenso e a remoção de palavras de preenchimento e pontuação antes da avaliação (Srivastav; Zheng et al., 2025). Além da WER para aferir a qualidade da transcrição, o material avalia também RTFx, permitindo uma análise direta da eficiência computacional e latência dos modelos.

Os resultados apresentados pelo *Leaderboard* no cenário multilíngue, que abrange alemão, francês, italiano, espanhol e português, indica que a especialização em inglês muitas vezes ocorre em detrimento da cobertura de outros idiomas, com modelos open-source demonstrando desempenho competitivo e frequentemente superior às soluções proprietárias avaliadas. O ranking, atualizado frequentemente na plataforma Hugging Face (Srivastav; Zheng et al., 2025), apresenta-se na Tabela 3 os resultados em idiomas testados além do inglês.

Tabela 3 – Resultados do Open ASR Leaderboard ordenados por WER em português

Modelo	WER Média	RTFx	Alemão	Francês	Italiano	Espanhol	Português
OpenAI Whisper Large v3	4,91	126,16	4,97	6,59	5,14	3,32	4,38
OpenAI Whisper Large v3 Turbo	5,44	188,6	5,86	7,01	5,86	3,77	4,54
Mistral AI Voxtral Mini 3B	5,18	58,93	5,36	5,96	5,88	3,81	4,80
Microsoft Phi 4 Multimodal Instruct	4,60	25,12	4,50	5,13	4,80	3,59	5,15
NVIDIA Parakeet TDT 0.6B v3	5,05	2154,2	4,90	5,38	5,58	3,72	5,95
NVIDIA Canary 1B v2	4,89	630,22	4,96	4,86	5,66	3,22	6,23
ElevenLabs Scribe v1 test	6,68	—	4,03	6,60	6,98	6,43	9,37
ElevenLabs Scribe v1	12,57	—	10,8	15,1	7,71	9,28	22,8

Fonte: adaptado de Hugging Face (2025)

Considerando-se o propósito de ensino de inglês para brasileiros, faz-se interessante a utilização de modelos multilíngues que consigam compreender o estudante caso o mesmo não consiga expressar-se em inglês. Neste cenário, o *Open ASR Leaderboard* evidencia a predominância e a eficácia de arquiteturas derivadas do Whisper, cujas variantes figuram consistentemente entre os melhores desempenhos tanto em transcrição de inglês quanto em suporte multilíngue (Srivastav; Zheng et al., 2025). Em seguida, destacam-se os modelos Voxtral Mini 3B da empresa Mistral AI (Liu et al., 2025), Phi 4 Multimodal Instruct da Microsoft (Microsoft et al., 2025), e o Parakeet TDT 0.6B v3 da NVIDIA (Sekoyan et al., 2025) entre os cinco principais de código aberto.

Desenvolvido pela OpenAI, o Whisper fundamenta-se em uma arquitetura *Transformer Encoder-Decoder*, e seu lançamento deu-se com um treinamento sob o paradigma de supervisão fraca em larga escala, processando um banco de dados de 680 mil horas de áudio coletado da web, das quais 117 mil horas cobrem 96 idiomas distintos além do inglês (Radford et al., 2022). Segundo os autores, tal exposição a dados diversos confere ao modelo uma capacidade de generalização próxima à humana, especialmente na versão Large v3 utilizada na avaliação, com 1550 milhões de parâmetros, tanto na versão comum quanto na Turbo, que caracteriza-se por menos camadas de decodificação que garantem ao modelo mais velocidade. Tal exposição a dados permite a ambos lidar com sotaques acentuados e ruído de fundo típicos de ambientes de aprendizado reais, sem a necessidade de re-treinamento específico.

Após os modelos Whisper Large, o melhor desempenho é do modelo Voxtral Mini 3B, desenvolvido pela Mistral AI. Este sistema distingue-se por adotar uma arquitetura multimodal que integra a robustez acústica de um codificador baseado no próprio Whisper large-v3 com a capacidade preditiva de um decodificador baseado em LLM próprio da empresa, o Ministral 3B (Liu et al., 2025). Para viabilizar essa integração, o modelo emprega uma camada adaptadora que reduz a resolução temporal dos vetores de áudio, dividindo-a por quatro, o que permite um processamento mais eficiente de contextos longos de até 32 mil *tokens* ou cerca de 40 minutos de áudio (Liu et al., 2025).

Dentro dos modelos multilíngues, o Phi-4 Multimodal Instruct, da Microsoft, destaca-se globalmente, liderando a classificação em termos de menor WER médio global (4,60). Diferentemente das abordagens anteriores que acoplam codificadores ao LLM via ajuste fino total, o Phi-4 baseia-se em um modelo de linguagem base de 3,8 bilhões de parâmetros que permanece congelado sem treinamento, enquanto adaptadores Low-Rank Adaptation (LoRA) específicos para áudio são treinados para processar os sinais provenientes de um codificador baseado em *Conformer* (Microsoft et al., 2025). Tal estratégia modular permite que o sistema processe fala e áudio preservando integralmente as capacidades de raciocínio textual do modelo original, resultando em um desempenho de ponta tanto em reconhecimento de fala multilíngue quanto em tradução.

Por fim, entre os modelos de código aberto de destaque, situa-se o Parakeet TDT 0.6B v3, desenvolvido pela NVIDIA e descrito por (Sekoyan et al., 2025). Segundo os autores, este modelo é projetado especificamente para tarefas de reconhecimento de fala com baixa latência e alto rendimento, com uma arquitetura que combina um codificador *FastConformer*, que processa o áudio com taxas de subamostragem otimizadas, a um decodificador *Token-and-Duration Transducer* (TDT). Segundo os autores, o mecanismo TDT inova em relação aos transdutores tradicionais ao prever simultaneamente o token textual e a duração temporal que ele ocupa no sinal, permitindo que o modelo pule quadros redundantes ou de silêncio durante a inferência. Treinado sob um regime supervisionado com um conjunto de 1,7 milhão de horas de áudio, o Parakeet v3 expande a capacidade de seu antecessor ao oferecer suporte a 25 idiomas, incluindo o português, apresentando o maior RTFx entre os modelos multilíngues avaliados.

Considerando os principais modelos avaliados em português, e com o objetivo de selecionar o motor de inferência que melhor atenda aos requisitos não funcionais de desempenho, conduziu-se uma avaliação comparativa entre três ASRs escolhidos a partir do *Open ASR Leaderboard*, bem como um modelo leve de referência. Considerando os de melhor desempenho e com código aberto, optou-se pelos seguintes modelos: OpenAI Whisper Large v3, OpenAI Whisper Large v3 Turbo, NVIDIA Parakeet TDT 0.6B v3 e OpenAI Whisper Tiny. Apesar de Voxtral Mini 3B e Phi 4 Multimodal Instruct desempenharam bem na classificação, apresentando WER de 4,80 e 5,15 respectivamente, ambos foram descartados por problemas de compatibilidade em bibliotecas com outros modelos, tornando a comparação mais enxuta.

O desenho experimental foi estruturado para avaliar a robustez dos modelos em dois cenários acústicos distintos, totalizando 348 segmentos de áudio e aproximadamente 41 minutos de gravação. A primeira situação, em cenário controlado, utilizou o subconjunto do *Speech Accent Archive*, com 48 amostras de falantes brasileiros lendo o mesmo texto padronizado. Tal corpus, totalizando cerca de 21 minutos, permite o isolamento de variáveis, visto que o conteúdo é fixo, restando apenas o sotaque como variável dependente. O segundo cenário fez-se através da seleção aleatória de 300 segmentos do corpus curado de palestras

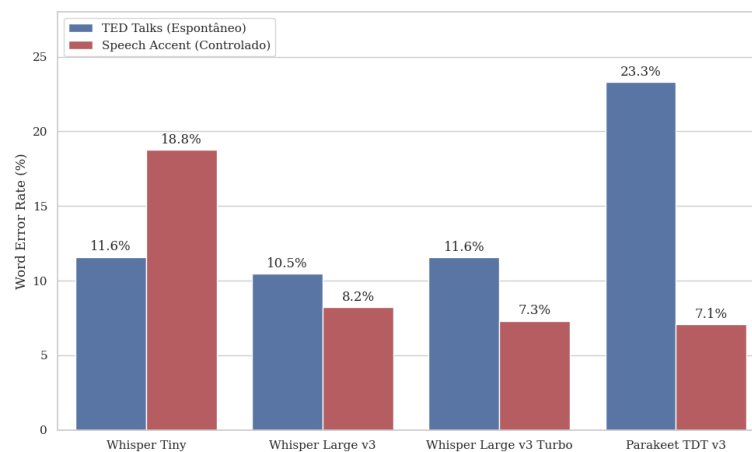
TED, totalizando aproximadamente 20 minutos. Tal conjunto introduz desafios de maior complexidade, como vocabulário técnico, prosódia de oratória, variação de velocidade de fala e ruído ambiental, simulando um ambiente de produção real.

A execução do protocolo ocorreu em ambiente controlado utilizando uma GPU NVIDIA T4 no Google Colab, com 16GB de memória RAM. O pipeline de inferência foi implementado em Python, utilizando o toolkit NeMo (NVIDIA Corporation, 2026) para os modelos da NVIDIA e o pacote oficial openai-whisper para a família Whisper (OpenAI, 2025), com cálculo de métricas através da biblioteca jiwer (Vaessen, 2025).

A análise quantitativa fundamentou-se na extração de três métricas-chave para a caracterização do desempenho: WER e CER, voltadas à mensuração da fidelidade da transcrição nos níveis de palavra e caractere, respectivamente, e o RTF, que permite entender a latência da transcrição, indicando o desempenho no caso de operação em tempo real. Diferente de abordagens que aplicam normalização agressiva (como remoção total de pontuação), optou-se por avaliar a capacidade dos modelos de gerar saídas prontas para uso, mantendo a estrutura original das sentenças sempre que possível, o que impõe um rigor maior às métricas de WER calculadas.

O desenho experimental estabeleceu uma estratégia de avaliação dividida em um modelo de referência (*baseline*) e modelos especializados de alto desempenho, permitindo uma análise comparativa mais ampla. Para compor a linha de base, selecionou-se o Whisper Tiny da OpenAI, que utiliza uma estrutura Transformer encoder-decoder treinada sob supervisão fraca. Com apenas 39 milhões de parâmetros, este modelo representa o limite inferior de consumo de recursos, servindo como comparativo para validar se o ganho de desempenho dos modelos maiores justifica o custo computacional adicional. Como resultado dessa avaliação, os modelos apresentaram o desempenho em termos de WER como exposto na Figura 4.

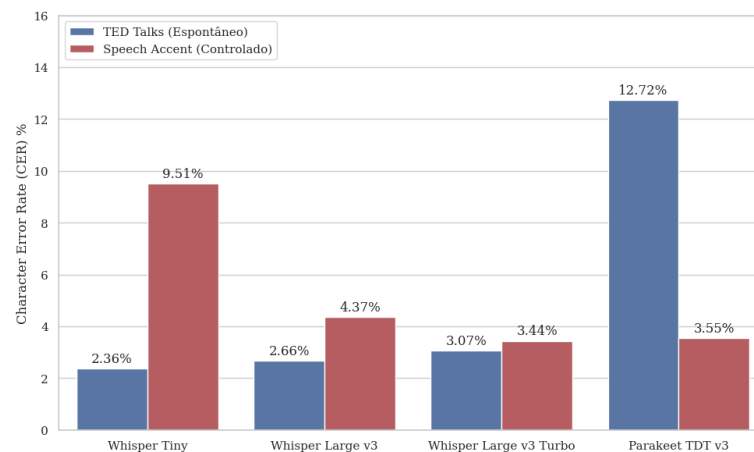
Figura 4 – Desempenho dos modelos no reconhecimento de áudio - WER



Fonte: Autor

É imperativo, contudo, contextualizar o desempenho do baseline no cenário TED sob a ótica do viés de seleção. Como o Whisper Tiny foi utilizado na etapa de pré-processamento para a validação automática do corpus, descartando segmentos onde sua própria transcrição divergia significativamente da referência, há evidente *overfitting* quando o modelo avalia esse mesmo grupo de dados. Portanto, o desempenho não enviesado do Whisper Tiny é melhor representado pelo cenário Speech Accent Archive (barra vermelha), onde não houve filtragem prévia baseada no modelo. A discrepância de quase 7 pontos percentuais entre os dois cenários para o Tiny (11,6% vs 18,8%) corrobora a existência desse viés, reforçando ainda mais a superioridade das arquiteturas Large e Turbo, que mantiveram consistência mesmo sem terem participado do processo de curadoria dos dados. Tal fenômeno é visto também na avaliação do CER:

Figura 5 – Desempenho dos modelos no reconhecimento de áudio - CER



Fonte: Autor

Adicionalmente, deve-se considerar o viés arquitetural compartilhado, visto que o baseline (Tiny) e os modelos de alto desempenho (Large v3 e Turbo) pertencem à mesma família algorítmica. Os três compartilham a mesma arquitetura Whisper, com a mesma lógica de tokenização e distribuições de dados de treinamento similares. Dessa forma, a filtragem realizada pelo Tiny selecionou segmentos acústicos intrinsecamente alinhados às capacidades indutivas dessa arquitetura específica.

Este fenômeno de pré-filtragem direcionada cria um cenário de avaliação desfavorável para o NVIDIA Parakeet, que opera sob um paradigma distinto. Ao ser testado no corpus TED, o Parakeet enfrentou um conjunto de dados do qual foram removidos justamente os exemplos que a arquitetura Whisper estranhou, mas que poderiam ser triviais para outro sistema. Já no corpus Speech Accent Archive, onde essa filtragem arquitetural não ocorreu, o Parakeet demonstrou sua real capacidade, superando os outros modelos em WER, e desempenhando levemente abaixo do Whisper Large v3 Turbo em CER, com uma diferença marginal da ordem de 0,1 pontos percentuais (3,55% para o Parakeet contra

3,44% para o Turbo). Dessa forma, vê-se que sua alta taxa de erro no cenário espontâneo foi exacerbada pela curadoria enviesada dos dados.

Por fim, a análise do RTF evidencia a disparidade de eficiência computacional entre as arquiteturas avaliadas. O modelo NVIDIA Parakeet TDT alcançou a menor latência global do experimento, registrando um RTF de 0,020, equivalente a um RTFx de 49,63 vezes o tempo real de fala. No extremo oposto, o Whisper Large v3, apesar de sua robustez semântica, apresentou o maior custo computacional, com um RTF de 0,179. Apesar disso, o mesmo apresenta um RTFx de 5,57, podendo ser usado em situações de transcrição em tempo real, como visto na Tabela 4.

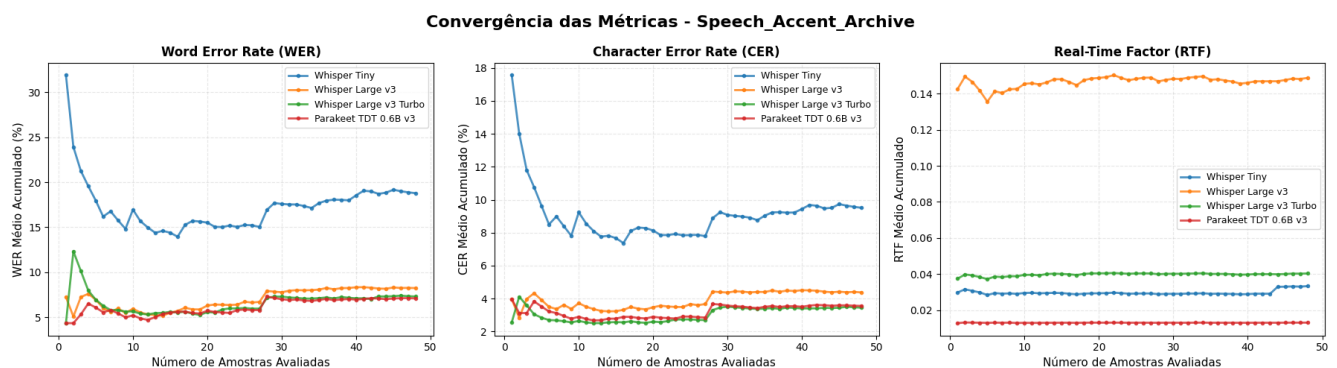
Tabela 4 – Comparativo de Eficiência Computacional entre modelos de ASR

Modelo	RTF	RTFx
Whisper Tiny	0,046	21,92x
Whisper Large v3	0,179	5,57x
Whisper Large v3 Turbo	0,071	14,15x
Parakeet TDT 0.6B v3	0,020	49,63x

Fonte: Autor

Estabelecido que o *Speech Accent Archive* representa o cenário de avaliação mais isento de vieses algorítmicos, faz-se necessário validar a consistência estatística dos resultados obtidos neste corpus. A Figura 6 a seguir apresenta a análise de convergência cumulativa das métricas ao longo das inferências. Observa-se que, apesar da menor amostragem (48 segmentos) em comparação ao corpus TED, as curvas de erro apresentam rápida estabilização. Tal fenômeno indica que o desempenho superior observado não é fruto de outliers, mas sim uma tendência consistente e representativa da capacidade do modelo em lidar com sotaques quando livre de filtragens adversárias.

Figura 6 – Convergência de métricas no reconhecimento de áudios do Speech Accent Archive

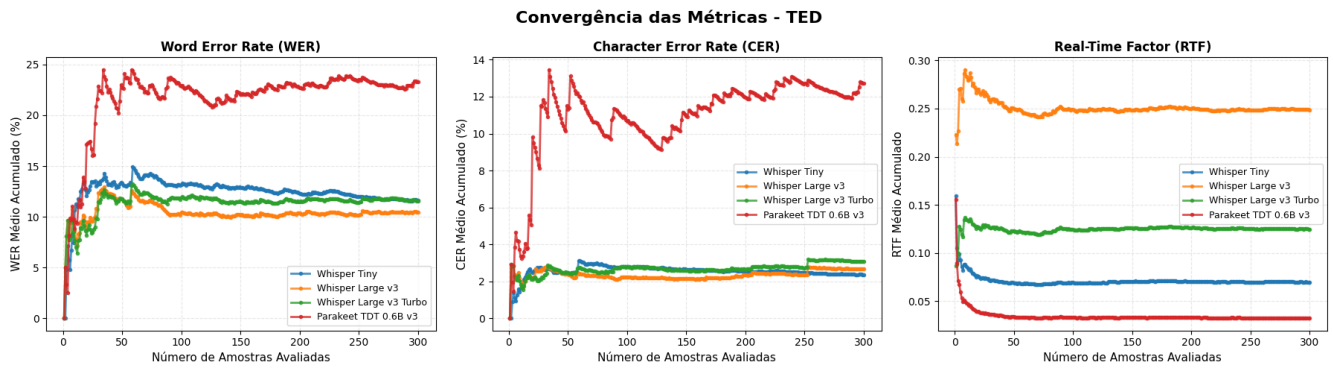


Fonte: Autor

Em contrapartida, a análise de convergência no corpus TED revela uma dinâmica

distinta, reflexo direto de suas características construtivas e aspectos inerentes à fala espontânea. Diferente da leitura cadenciada e uniforme do corpus anterior, os segmentos do TED introduzem variáveis de prosódia, hesitações e ruído de fundo que geram maior volatilidade inicial nas métricas. refletido na Figura 7. No entanto, há estabilidade nos modelos Whisper após as turbulências iniciais, enquanto o modelo Parakeet, além de situar-se em níveis mais elevados de erro, sofre com variações mesmo próximo ao final do banco de dados de teste.

Figura 7 – Convergência de métricas no reconhecimento de áudios do TED



Fonte: Autor

A variabilidade das métricas no *Speech Accent Archive* pode ser percebida também ao analisar-se a variação global dos indicadores nos últimos 20% dos registros, onde vê-se variações pequenas nos modelos mais robustos, inferiores a 2% em CER, como demonstrado na Tabela 5.

Tabela 5 – Estatísticas de Convergência das Métricas no *Speech Accent Archive*

Modelo	Amostras	Variab. WER (últimos 20%)	Variab. CER (últimos 20%)	Variab. RTF (últimos 20%)
Whisper Tiny	48	$\pm 12,61\%$	$\pm 6,54\%$	$\pm 0,054$
Whisper Large v3	48	$\pm 3,16\%$	$\pm 1,59\%$	$\pm 0,017$
Whisper Large v3 Turbo	48	$\pm 4,59\%$	$\pm 1,43\%$	$\pm 0,005$
Parakeet TDT 0.6B v3	48	$\pm 2,45\%$	$\pm 1,45\%$	$\pm 0,001$

Fonte: Autor

Diante dos resultados obtidos, o Whisper Large v3 Turbo foi selecionado como o motor de inferência definitivo para este trabalho. Sua escolha fundamenta-se na consistência de desempenho observada, mesmo considerando-se o viés da seleção de trechos do Corpus Espontâneo, que prejudicou o desempenho do modelo Parakeet. No entanto, dado que houve um desempenho de destaque também quando exposto ao banco de dados sem viés, entende-se que o Whisper Large v3 Turbo fornece uma operação confiável indepen-

dente do viés dos dados de entrada, ao mesmo tempo em que oferece latência compatível com a interatividade exigida.

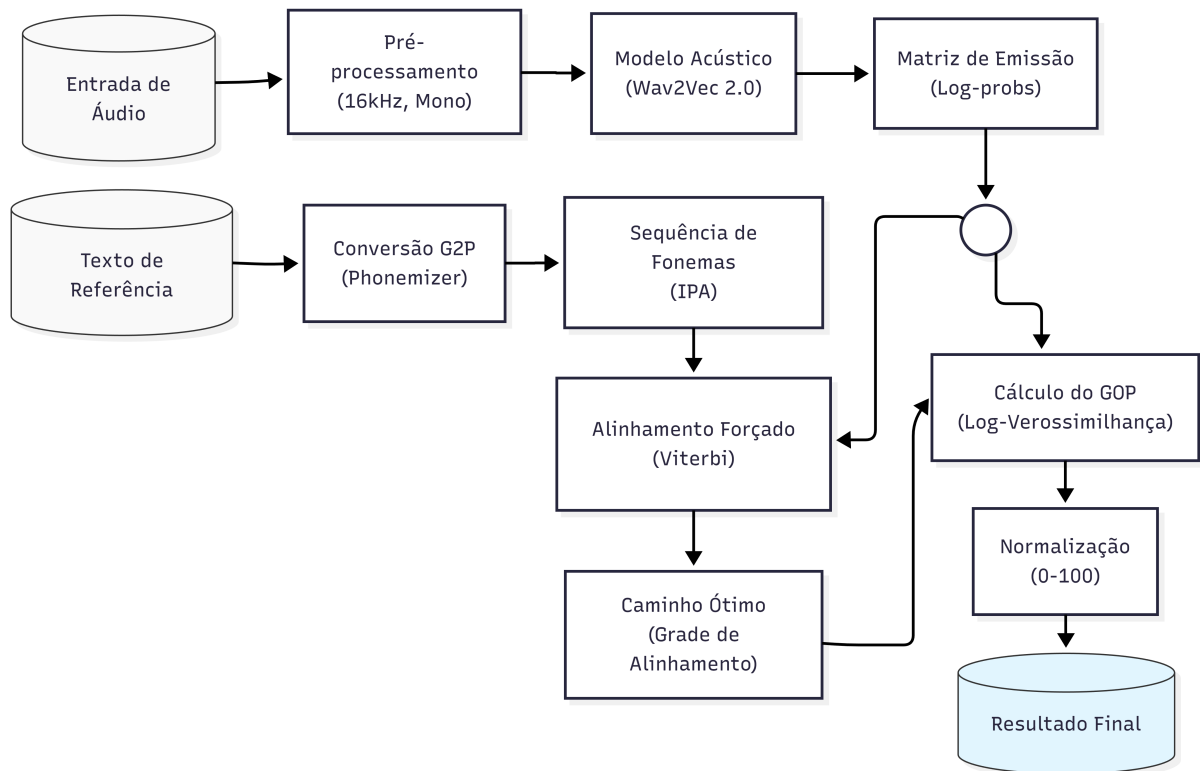
3.3 AVALIAÇÃO DE PRONÚNCIA

A CAPT impõe um desafio técnico distinto da simples transcrição de áudio, dado que exige uma análise granular da qualidade acústica de cada segmento da fala individualmente. Portanto, não basta identificar o que foi dito, sendo necessário alinhar temporalmente a execução do falante com a referência fonética ideal e mensurar a confiança do modelo acústico nesse alinhamento. De forma a operacionalizar essa análise, este trabalho implementa um pipeline fundamentado no algoritmo GOP, com o auxílio de um modelo acústico pré-treinado Wav2Vec 2.0 (Baevski et al., 2020) para extração de características e cálculo de probabilidades, acoplado a um algoritmo de alinhamento forçado via Viterbi (Forney, 1973). Tal abordagem permite confrontar a produção do estudante contra o modelo nativo, quantificando desvios fonéticos através de pontuações de verossimilhança (log-likelihood).

A escolha do framework Wav2Vec 2.0 fundamenta-se em sua capacidade de aprender representações de fala robustas diretamente do áudio bruto, mesmo com poucos dados rotulados. Diferentemente de arquiteturas Seq2Seq (como o Whisper) que podem mascarar erros de pronúncia por buscarem absorver o contexto, o Wav2Vec 2.0, quando ajustado com a função de perda do tipo Classificação Temporal Conexionista (CTC, do inglês (*Connectionist Temporal Classification*)) preserva o alinhamento temporal explícito necessário para o cálculo do GOP, fornecendo probabilidades acústicas exatas em uma matriz de emissão (Baevski et al., 2020). Portanto, este modelo é ideal para detectar desvios acústicos específicos em vez de apenas corrigir o texto final, permitindo sua utilização para avaliação de pronúncia de fonemas.

O fluxo geral do sistema implementado em Python, desde a entrada do áudio bruto até a geração das pontuações normalizadas, é apresentado a seguir, na Figura 8.

Figura 8 – Diagrama de blocos representativo do sistema de avaliação de pronúncia



Fonte: Autor

O primeiro bloco, de preparação de dados, consiste no pré-processamento simultâneo das entradas acústicas e textuais para padronizar os insumos do modelo. O áudio bruto é convertido para canal único (mono) e reamostrado para 16 kHz, alinhando-se à taxa de amostragem nativa utilizada no treinamento do Wav2Vec 2.0. Paralelamente, o texto de referência passa por uma conversão de Grafema para Fonema (G2P) utilizando a biblioteca Phonemizer (Bernard, 2021) com o *backend* eSpeak (Epaneshnikov; Dunn; Vitolins, 2026), em um processo que transforma a transcrição ortográfica em uma sequência de símbolos do Alfabeto Fonético Internacional (IPA), estabelecendo o gabarito fonético necessário para o alinhamento forçado e o subsequente cálculo da pontuação de pronúncia.

O áudio reamostrado para 16 kHz alimenta o Modelo Acústico Wav2Vec 2.0, que processa-o, gerando como saída uma Matriz de Emissão que contém as log-probabilidades de cada fonema possível para cada quadro de tempo do áudio, representando a interpretação bruta do modelo antes de qualquer contextualização linguística.

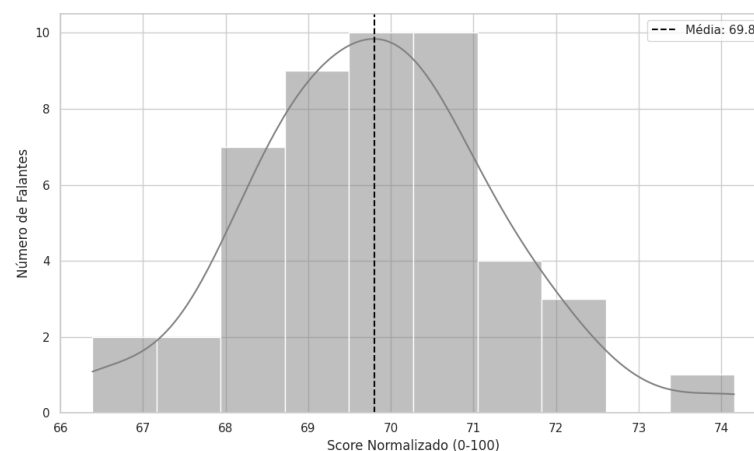
A integração entre os dados textuais e acústicos da Matriz de Emissão e da sequência de fonemas ocorre no bloco de Alinhamento Forçado. Utilizando o algoritmo de Viterbi, o sistema projeta a sequência de fonemas esperada sobre a Matriz de Emissão para encontrar o caminho ótimo na grade de alinhamento (trellis). O resultado deste processo é a segmentação temporal que delimita os quadros exatos de início e fim de cada fonema

pronunciado.

Com a segmentação temporal estabelecida, o sistema calcula o GOP relativo para cada fonema. Diferentemente de uma simples medida de confiança acústica, o GOP compara, em cada quadro do segmento, a log-probabilidade do fonema esperado com a log-probabilidade do fonema mais provável segundo o modelo. A média dessa diferença ao longo dos quadros do segmento constitui a pontuação do trecho de áudio, com valores próximos de zero indicando que o fonema esperado coincide com o mais provável, enquanto valores fortemente negativos indicam que o modelo identifica outro fonema como mais provável naquele trecho. Por fim, para garantir a interpretabilidade do resultado final, aplica-se uma normalização linear com truncamento, convertendo os valores brutos (logarítmicos) em uma pontuação percentual de 0 a 100, onde valores mais altos indicam maior fidelidade à pronúncia nativa.

Ao submeter os 48 segmentos de áudio do *Speech Accent Archive* ao pipeline de avaliação proposto, obteve-se a distribuição de *Scores* de Pronúncia apresentada na Figura 9. Os dados refletem as pontuações normalizadas em uma escala de 0 a 100, consolidando o desempenho global dos locutores do grupo amostral.

Figura 9 – Distribuição de pontuação de pronúncia



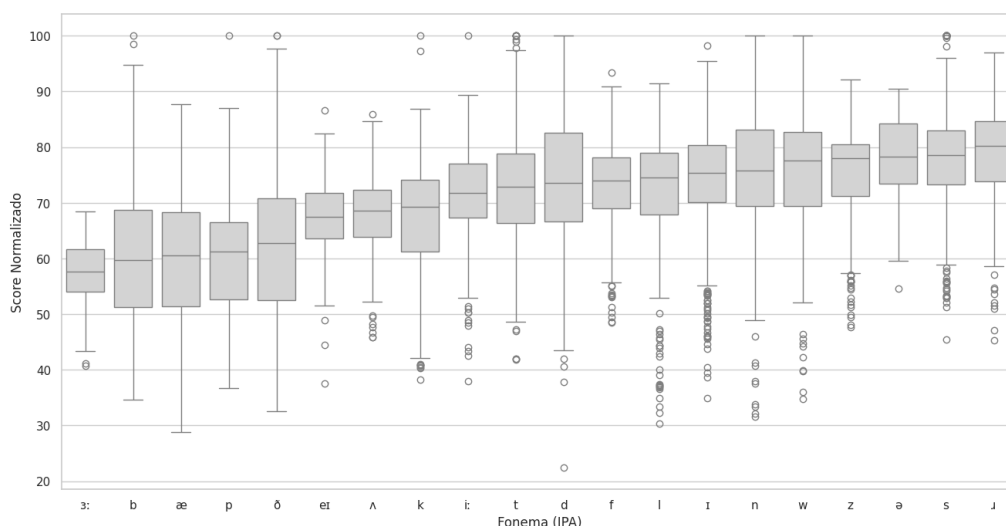
Fonte: Autor

A análise estatística descritiva aponta para uma notável homogeneidade na proficiência do grupo, com um *score* médio de 69,8 e um desvio padrão reduzido de $\pm 1,5$. A coincidência entre a média e a mediana (69,8), aliada a um intervalo de variação estreito (66,4 a 74,2), indica que os falantes avaliados apresentam níveis de proficiência muito semelhantes.

A análise granular dos 40 fonemas processados revela discrepâncias significativas dependendo da classe fonética. As maiores dificuldades articulatórias foram registradas nos fonemas /ɔɪ/ (como em *toy*, média 30,0), /oɪr/ (como em *store*, média 36,4) e /ɔɪr/ (como em *for*, média 46,7). Em contrapartida, observa-se uma alta taxa de acerto em

consoantes como /ɹ/ (de *red*, média 79,1), /v/ (de *five*, média 78,5) e /s/ (de *spoon*, média 78,5) figuram entre os mais bem avaliados. Uma distribuição dos principais fonemas que aparecem no banco de dados encontra-se na Figura 10, em um gráfico construído em Python com as bibliotecas Seaborn (Waskom, 2021) e Matplotlib (Hunter, 2007).

Figura 10 – Distribuição de pontuação de pronúncia dos 20 principais fonemas no Speech Accent Archive



Fonte: Autor

A análise da distribuição de variabilidade revela alguns padrões na proficiência dos falantes, em especial as dificuldades em vogais e ditongos inexistentes no sistema fonético do português brasileiro. O fonema /æ/ (presente em *cat* ou *bad*), por exemplo, apresentou uma das medianas mais baixas. Da mesma forma, o fonema /ɜ:/ (encontrado em *bird* ou *work*) impõe um desafio articulatório significativo, pois o português não possui essa vogal central longa e tensa. Outro aspecto interessante dá-se na pronúncia do /b/, com um número elevado de pontuações mais baixas, sobretudo quando o mesmo aparece ao final de palavras, como *job*.

Em contraste, a extremidade direita do gráfico demonstra os fonemas com maior facilidade de pronúncia, sobretudo por conta de uma correspondência acústica direta ou muito próxima no português, facilitando a transferência positiva. Um caso interessante é o do fonema /ɹ/ (o “r” retroflexo de *red* ou *car*), que embora seja um som característico do inglês e com muita dificuldade de pronúncia para falantes de outros idiomas, foi sujeito a uma alta pontuação, provavelmente devido à sua presença em dialetos regionais brasileiros, naquilo que é chamado muitas vezes de “r caipira”, o que permite aos falantes uma adaptação articulatória natural e precisa.

Por fim, observa-se uma zona de alta variabilidade em fonemas como o /z/ (de *boys* ou *music*), cujas caixas alongadas indicam falta de consenso no desempenho do grupo.

A dificuldade aqui é posicional: embora o som do “z” exista em português (*casa*), ele raramente ocorre em final de palavra, levando o falante a pronunciar o plural *boys* com som de “ss” (*boyss*, como em *lápiss*) em vez do som vibrante correto (*boyzz*).

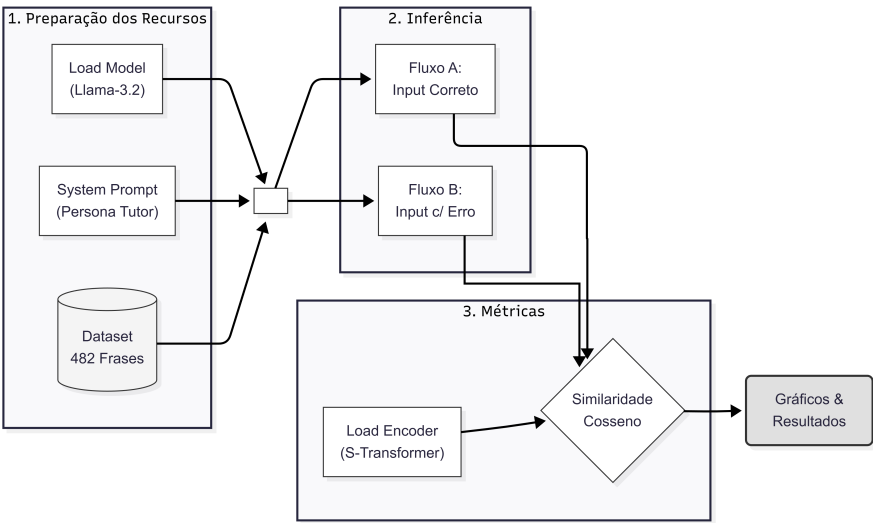
Cabe ressaltar, contudo, que a amostra proveniente do *Speech Accent Archive* apresenta um viés de seleção, dado que não é constituído por falantes inteiramente leigos no idioma. A própria natureza da coleta de dados, que envolve uma leitura voluntária de um parágrafo padronizado em inglês, pressupõe algum nível de escolaridade e exposição prévia à língua-alvo por parte dos participantes. Consequentemente, é natural que este grupo demonstre uma competência articulatória superior à média da população brasileira de aprendizes iniciantes, o que justifica, em parte, as pontuações elevadas observadas em fonemas historicamente desafiadores, como o retroflexo /ɺ/, que tenderiam a apresentar taxas de erro mais acentuadas em uma amostragem de proficiência elementar.

3.4 MECANISMO DE ELABORAÇÃO DE RESPOSTA

O mecanismo de elaboração de resposta constitui o núcleo conversacional da aplicação, encarregado de transformar a entrada transcrita do usuário em uma realimentação textual coerente e pedagogicamente alinhada. Mais do que um simples processador de texto, este componente atua como o mediador didático do sistema, projetado para sustentar o engajamento do estudante enquanto fornece correções linguísticas contextualizadas. A ideia central é simular uma interação natural, onde a inteligência do sistema deve ser capaz de interpretar a intenção comunicativa do estudante, sem a rigidez de respostas pré-programadas.

No entanto, a implementação prática desta arquitetura enfrenta um desafio crítico associado à qualidade dos dados de entrada. Como o texto processado provém de um sistema de reconhecimento automático de fala (ASR), ele é suscetível a erros fonéticos e ruídos de transcrição, que podem levar o mecanismo de interação a caminhos diferentes daqueles que o usuário gostaria de tomar no diálogo. De tal forma, para entender o impacto desses erros na raiz, criou-se um experimento com o objetivo de isolar e validar o do fluxo, conforme demonstra o diagrama da Figura 11.

Figura 11 – Diagrama representando o experimento de propagação de erros no mecanismo de elaboração de resposta



Fonte: Autor

Para validar a robustez do mecanismo de resposta, estruturou-se um conjunto de dados experimental composto por 482 pares de sentenças, em um processo de injeção controlada de ruído, desenhado para isolar a variável de imprecisão na transcrição. Cada registro do banco de dados constituiu-se de uma entrada de controle (Fluxo A), contendo uma sentença gramaticalmente e semanticamente íntegra, pareada com uma variante ruidosa (Fluxo B), alterada intencionalmente de forma a replicar algumas falhas comuns a um modelo de ASR. Tais desvios introduzidos no Fluxo B não consistem em erros de digitação aleatórios, mas sim em distorções baseadas em ambiguidade acústica, refletindo o comportamento real de transcritores automáticos em condições adversas. A tipologia dos erros distribui-se de acordo com as seguintes categorias:

Tabela 6 – Distribuição quantitativa dos tipos de erro no conjunto de dados (N=482)

Categoria do Erro	Ocorrências	Proporção
Erro Fonético Geral (<i>Phonetic</i>)	184	38%
Homofonia (<i>Homophone</i>)	100	21%
Erro Gramatical (<i>Grammar</i>)	68	16%
Fronteira de Palavra (<i>Word Boundary</i>)	60	12%
Nenhum/Sem Alteração (<i>None</i>)	60	12%
Total	482	100%

Fonte: Autor

A análise do conjunto de dados revela que as inconsistências se distribuem em categorias distintas, representando falhas comuns em transcrições automatizadas. A maior

parte da amostra é composta por erros fonéticos gerais (38%), caracterizados pela confusão entre fonemas semelhantes, gerando palavras com sonoridade similar, mas distintas da original (por exemplo, “think” transcrito como “thing”). Em seguida, 21% dos registros são casos de homofonia, caracterizados por termos que preservam a fonética original, mas divergem na grafia e semântica (por exemplo, “their” contra “there”), desafiando a capacidade do modelo de realizar a desambiguação pelo contexto. O grupo de erros gramaticais (16%), que inclui as reduções morfológicas, manifesta-se pela supressão de fonemas finais que levam à perda de marcadores de plural ou de concordância verbal (por exemplo, “she walks” transcrito como “she walk”). As falhas de fronteira de palavra (12%) ocorrem pela má interpretação dos limites lexicais na fala conectada, resultando em fusões ou separações indevidas de termos (por exemplo, “an ice cream” interpretado como “a nice cream”). Por fim, enquanto 12% dos registros não sofreram alteração, servindo como base de controle adicional.

Como motor de processamento para este experimento, utilizou-se o modelo de linguagem Llama 3.2 3B, desenvolvido pela Meta. Segundo o artigo associado a sua publicação, trata-se de um modelo autorregressivo baseado na arquitetura Transformer, otimizado para maior eficiência computacional e suporte a janelas de contexto estendidas de até 128 mil *tokens* (Meta, 2024). A escolha desta versão específica dá-se por basear-se em modelos mais robustos, treinados com até 15 trilhões de *tokens*, mas sem a necessidade de memória dos mesmos, tornando-o mais apropriado para o experimento executado.

De forma a operacionalizar a interação e garantir que o modelo atue estritamente dentro do escopo pedagógico proposto, a geração de respostas foi condicionada por um *prompt* de sistema onde através das instruções buscou-se configurar o agente como um tutor amigável, estabelecendo diretrizes de concisão e manutenção de um tom encorajador. A instrução exata fornecida ao modelo é apresentada a seguir:

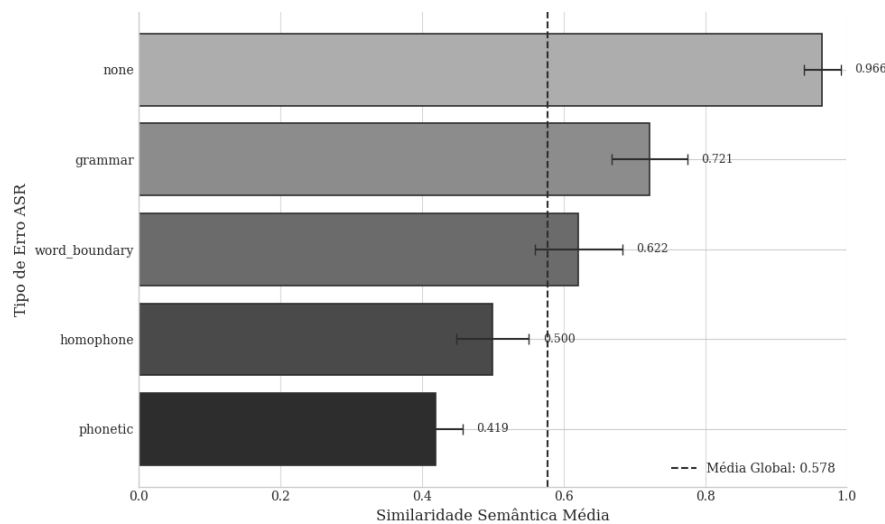
“You are a friendly English conversation tutor helping a Brazilian student practice speaking English. Your role is to:

- 1) Respond naturally to what the student says, continuing the conversation.
- 2) Keep responses concise (2-3 sentences max).
- 3) Be encouraging and keep the conversation flowing.
- 4) Do NOT correct grammar or pronunciation - just have a natural conversation Respond only with your conversational reply, nothing else.”

Apesar de não utilizar conceitos robustos de *prompt engineering*, a instrução busca dar um mínimo de contexto ao modelo, que deve compreender que está em meio a um diálogo e deve responder de volta. A partir das respostas geradas pelo modelo de linguagem para o Fluxo A e Fluxo B, compara-se ambos os *outputs* através de um cálculo de similaridade de cosseno, métrica que mede o quão alinhados dois vetores estão através do cálculo do cosseno do ângulo entre eles (Krantz; Jonker, 2024). Tal indicador é dado por

um valor entre -1 e 1, onde 1 indica máxima similaridade direcional, 0 indica ausência de relação direcional e -1 indica oposição completa. No contexto de linguagem, textos são convertidos em vetores, e a similaridade de cosseno é usada para comparar o quanto esses vetores apontam na mesma direção, permitindo uma comparação semântica entre trechos. Ao considerar-se a similaridade de cosseno a partir dos erros existentes no conjunto de dados, chega-se aos resultados vistos na Figura 12. No eixo vertical há o nome de cada grupo de erros em inglês, sendo *none* aqueles pares idênticos de frases, *grammar* aqueles com diferenças gramaticais entre pares, *word_boundary* os pares com erros nos limites entre palavras, *homophone* aqueles com erro de homofonia, e *phonetic* as duplas de frases com transcrições distintas em termos de fonética.

Figura 12 – Relação entre tipo de Erro e Similaridade de Cosseno



Fonte: Autor

Através dos resultados, é possível verificar que pares idênticos de frases geram *outputs* também quase idênticos (0,966), não sendo absolutamente idêntico devido ao indicador de temperatura utilizado de 0,01, que permite adicionar uma leve variação nas respostas. Em seguida, erros gramaticais levam a respostas também bastante similares, com o indicador de 0,721 demonstrando que o modelo consegue identificar o contexto com qualidade quando há tais disparidades. A má-identificação de limites de palavras também leva a um indicador acima da média, com o modelo entendendo o contexto e apresentando respostas semelhantes para os pares, em uma similaridade de cosseno de 0,622. Por outro lado, erros de homofonia e fonéticos apresentam diferenças acentuadas entre as duas respostas, com o primeiro apresentando similaridade de cosseno de 0,500, enquanto o segundo gera as maiores disparidades, com pontuação de 0,419.

As diferenças mais acentuadas ocasionadas por erros de homofonia e fonéticos dão-se pela característica de alteração semântica que ambos os erros carregam. No caso da

homofonia, dizer “the leek is fresh” quer dizer “o alho-poró está fresco”, enquanto “the leak is fresh” pode indicar algo como “o vazamento é recente”, gerando respostas extremamente diferentes do mecanismo de elaboração de resposta. Já na fonética, identificar incorretamente palavras semelhantes como “child” (criança) e “shield” (escudo) podem gerar também repostas bastante díspares. Apesar desses erros, uma média de 0,58 indica uma propagação de erro moderada, com apenas 12% dos pares apresentando erros críticos de impacto. Dessa forma, pode-se considerar que, apesar dos erros propagados, o sistema consegue operar com respostas coerentes ao usuário, mesmo que casos críticos possam levar a desvios no diálogo e posteriores correções de rota por parte do estudante.

3.5 RETORNO PARA O USUÁRIO VIA ÁUDIO

Concluída a transcrição da fala do usuário e a análise da pronúncia, e tendo-se a resposta textual elaborada pelo bloco anterior, o ciclo de interação do sistema tutor se encerra com a etapa de devolutiva acústica, operacionalizada através da Síntese de Fala (TTS). Enquanto o módulo de ASR atua na decodificação analítica do sinal de entrada, o subsistema de TTS desempenha o papel inverso de transformar a representação simbólica da realimentação textual e dos exemplos de correção em uma forma de onda contínua e inteligível. Esta etapa é crítica para a eficácia pedagógica da solução, pois fornece ao aprendiz não apenas uma correção descritiva, mas um modelo auditivo de referência, permitindo a comparação perceptiva imediata entre a sua tentativa e a pronúncia sintetizada ideal.

De forma a selecionar o melhor modelo de TTS de código aberto possível, buscou-se benchmarks abertos, chegando-se ao *TTS Spaces Arena*, hospedado na Hugging Face (Srivastav; Fourrier et al., 2025). Este ambiente experimental interativo tem como objetivo avaliar modelos de conversão de texto em fala de forma colaborativa através de comparações pareadas e avaliações cegas de naturalidade da fala sintetizada pelos usuários finais (Srivastav; Fourrier et al., 2025). Na plataforma, o usuário escuta duas amostras de áudio geradas a partir do mesmo enunciado em inglês e indica qual soa mais natural, sem saber o nome dos modelos.

Após o voto do usuário, a identidade dos modelos é revelada, alimentando um ranking público de desempenho entre diferentes modelos TTS, cujos principais modelos encontram-se descritos na Tabela 7. Para gerar a classificação, utiliza-se um sistema de Elo, comum em esportes como o xadrez, onde a pontuação de cada competidor quantifica sua força relativa. Segundo Beaume (2020), quando um competidor vence um duelo, seu Elo aumenta em proporção direta à classificação do oponente. Dessa forma, quanto mais “forte” for o adversário derrotado, maior será o ganho de pontos. Portanto, todo cálculo da expectativa de resultado é estruturado com base nas diferenças de pontuação entre os modelos avaliados, permitindo um ranqueamento dinâmico e ajustável.

Tabela 7 – Ranking de desempenho de modelos TTS (Open Leaderboard)

Posição	Modelo	Taxa de Vitória	Elo	Votos
#1	Kokoro v0.19	75,8%	1293	9.125
#2	Kokoro v1.0	71,4%	1238	16.456
#3	Orpheus 3B v0.1	66,2%	1174	11.144
#4	XTTS v2	57,0%	1166	11.206
#5	MetaVoice	57,4%	1166	787

Fonte: Adaptado de Open TTS Leaderboard

Dentro do ranking, os dois com maior taxa de vitória são versões do Kokoro (Hexgrad, 2025), modelo text-to-speech aberto de 82M parâmetros que converte texto em fala de alta qualidade, focando em ser leve, rápido e barato para uso em produção e projetos pessoais. Tem como diferencial seu tamanho reduzido, que não demanda poder computacional elevado, bem como um treinamento feito exclusivamente em dados não protegidos por direitos autorais, incluindo áudios sintéticos (Hexgrad, 2025). Por seu desempenho elevado frente aos outros modelos do Leaderboard, optou-se pelo seu uso e implementação em uma avaliação simples.

Apesar de emitir áudios com qualidade superior a modelos concorrentes, o Kokoro não atua como TTS sozinho, necessitando da biblioteca eSpeak (Epaneshnikov; Dunn; Vitolins, 2026) para converter a ortografia irregular do inglês na representação fonética padronizada IPA. Isso dá-se porque o Kokoro-82M foi treinado para mapear fonemas em espectrogramas, e não texto bruto em áudio. Dessa forma, incorporou-se ao script em Python este pacote, que permitiu a conversão de grafema para fonema (G2P). Em seguida, a geração e avaliação de áudios foi executada com 20 frases de teste predefinidas, variando de saudações simples a frases tecnicamente complexas.

Com os áudios gerados e a taxa de amostragem de 24kHz, foi possível estimar o RTF para cada frase individualmente. Vê-se um maior RTF na primeira execução, quando carrega-se o modelo, mas uma distribuição equilibrada nas 20 amostras, com um RTF médio de 0,13, como pode ser visto na Tabela 8.

Tabela 8 – Estatísticas descritivas do desempenho de síntese de áudio (20 frases)

Métrica	Tempo de Síntese (s)	Duração do Áudio (s)	RTF
Média	0,279	2,636	0,130
Desvio Padrão	0,729	0,755	0,377
Mínimo	0,090	1,825	0,031
Mediana (50%)	0,116	2,250	0,044
Máximo	3,376	4,175	1,731

Fonte: Autor

Os áudios resultantes foram então salvos sequencialmente em disco usando a bibli-

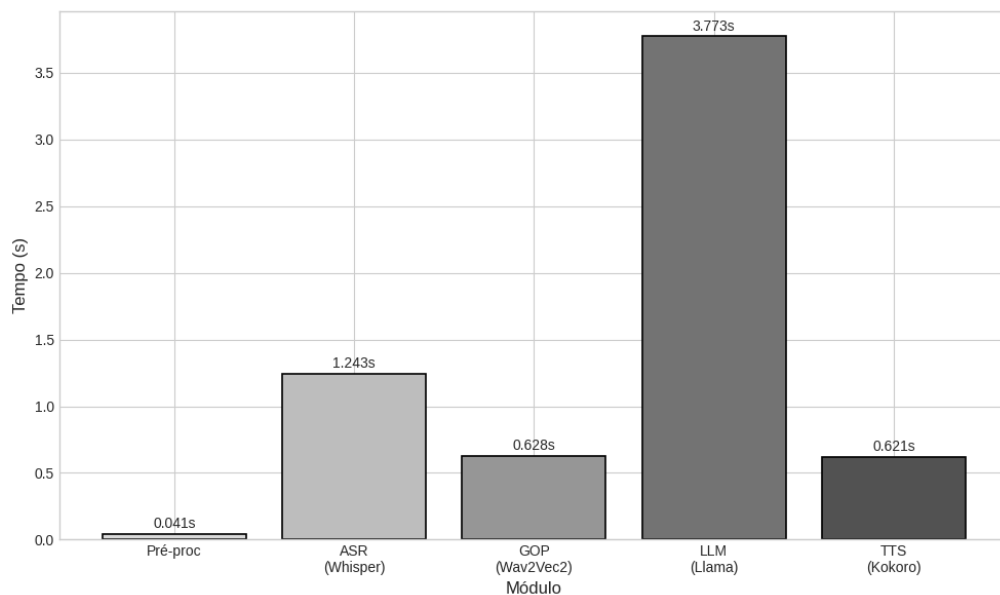
oteca soundfile (Bechtold, 2025), o que permitiu uma auditoria qualitativa posterior para confirmar que entonação e pausas soavam naturais aos ouvidos humanos.

3.6 DESEMPENHO DO CONJUNTO

De forma a entender a viabilidade do sistema como um todo, faz-se necessário avaliar o funcionamento dos blocos quando em conjunto, dado que a avaliação individual dos blocos não é suficiente para garantir a fluidez da interação. Para tal, fez-se um experimento simples buscando avaliar a latência das etapas unidas, buscando entender se o sistema sustenta uma operação em tempo real. Utilizou-se 10 segmentos de áudio do Speech Accent Archive, que passam por um reconhecimento de fala com Whisper Large v3 Turbo, seguido de uma avaliação de pronúncia com GOP, a elaboração de uma resposta textual através do Llama 3.2 3B e sua subsequente conversão em fala através do TTS Kokoro. Para viabilizar a execução em um ambiente com recursos limitados (GPU Tesla T4 com 16GB VRAM), adotou-se uma estratégia de gerenciamento dinâmico de memória, carregando e descarregando os modelos da GPU conforme a necessidade de cada etapa.

Os resultados obtidos foram positivos: o sistema atingiu um RTF médio de 0,206. Em termos práticos, isso significa que para cada segundo de áudio recebido, o sistema necessita de apenas cerca de 0,2 segundo para processar a entrada e gerar a saída, sendo 4,84 vezes mais rápido que o tempo real. A Figura 13 ilustra a contribuição de cada módulo para o tempo total. Observa-se que, apesar da complexidade dos modelos envolvidos, o tempo total médio de processamento foi de 6,31 segundos para áudios com duração média de 32,5 segundos.

Figura 13 – Tempo de processamento médio por etapa



Fonte: Autor

A geração de resposta é o componente dominante, sendo responsável por aproximadamente 60% do tempo total de processamento (3,77s). Isso se deve à natureza autor-regressiva da geração de texto, que produz *tokens* sequencialmente. Em contrapartida, os módulos de ASR e GOP, apesar de processarem o áudio bruto, mostram-se extremamente otimizados, com tempos de execução na ordem de 1,2s e 0,6s, respectivamente. O TTS Kokoro também apresentou desempenho notável, sintetizando a resposta em áudio em apenas 0,62s.

O RTF consolidado de 0,206 confirma a viabilidade técnica do sistema proposto. A margem de desempenho obtida oferece robustez suficiente para acomodar variações na carga computacional sem prejudicar a percepção de instantaneidade do usuário.

4 CONCLUSÃO

A baixa proficiência em língua inglesa no cenário brasileiro, aliada a barreiras emocionais como a ansiedade linguística e o acesso restrito a interlocutores nativos evidencia a urgência de ferramentas tecnológicas capazes de democratizar o ensino de idiomas. O presente trabalho propôs e validou uma arquitetura modular para um Agente Conversacional, demonstrando que a integração de modelos de estado da arte em Processamento de Sinais e Inteligência Artificial permite a criação de um tutor virtual capaz de fornecer realimentação pedagógica em tempo real.

Para viabilizar a interação oral, o estudo enfrentou inicialmente o desafio da transcrição de fala de não-nativos em inglês, que tivessem como língua materna o português. A constituição de um corpus inédito, composto por fala controlada do Speech Accent Archive e fala espontânea de palestras públicas, permitiu uma avaliação rigorosa de modelos de ASR. Os experimentos de *benchmarking* confirmaram que o modelo Whisper Large v3 Turbo oferece o equilíbrio ideal entre acurácia e latência, superando vieses observados em outras arquiteturas quando expostas a dados espontâneos e mantendo um desempenho robusto.

No que tange à qualidade da pronúncia, o módulo de avaliação baseado em GOP através do modelo Wav2Vec 2.0 provou-se eficaz na identificação de desvios fonéticos. A análise dos dados revelou padrões consistentes de dificuldades de pronúncia por parte dos falantes, ao mesmo tempo que evidenciou a transferência positiva de traços regionais, como o “r” retroflexo, facilitando a aquisição de certos fonemas do inglês. Essa capacidade de diagnóstico valida o sistema não apenas como ferramenta de transcrição, mas também como peça de monitoramento pedagógico.

Para uma devolutiva ao usuário, o mecanismo de elaboração de resposta baseou-se no modelo Llama 3.2, com testes de injeção controlada de erros que demonstraram resiliência semântica. A partir do experimento de avaliação semântica de *outputs* a partir da exposição de pares de entradas com erro, o agente manteve coerência de diálogo mesmo diante de erros fonéticos típicos de transcrição automática, preservando o fluxo da interação e o propósito instrucional. A interface de saída, operacionalizada pelo modelo de síntese de fala Kokoro, completou o ciclo comunicativo, fornecendo uma devolutiva sonora natural e eficiente.

Considerando os objetivos específicos estabelecidos, este trabalho mapeou com êxito as tecnologias de ASR, TTS e LLM aplicáveis ao ensino, realizando a validação quantitativa desses componentes através de métricas de erro (WER/CER) e latência. A implementação do protótipo e a orquestração dos módulos confirmaram a viabilidade técnica da arquitetura para operação com RTF compatível com a interação humana.

É importante ressaltar, contudo, as limitações do estudo. A validação concentrou-se em métricas técnicas e objetivas, não abrangendo testes com usuários humanos em

ambiente de aprendizado controlado, o que seria necessário para mensurar o ganho efetivo de fluência e a redução do filtro afetivo a longo prazo. Adicionalmente, a dependência de hardware de alto desempenho (GPUs) para manter a baixa latência dos modelos locais ainda representa um desafio para a escalabilidade da solução. Além disso, a construção do *corpus* próprio passou por desafios relacionados a *overfitting* que reduziram sua confiabilidade no que tange aos segmentos de áudio espontâneos, fazendo com que alguns não fossem utilizados em etapas posteriores do experimento.

O autor declara que ferramentas de IA foram utilizadas exclusivamente para fins de revisão gramatical, ortográfica e otimização estilística do texto. Ressalta-se que a concepção do tema, a estruturação dos capítulos, a análise crítica dos dados e a redação original de todas as ideias contidas neste trabalho são de autoria própria e integral do autor, assumindo este a responsabilidade total pelo conteúdo apresentado.

Para trabalhos futuros, sugere-se a implementação de uma interface gráfica interativa voltada para a experiência do usuário, bem como a realização de estudos de campo com estudantes de diferentes níveis de proficiência. Além disso, o uso de sistemas de filtragem e detecção de ruído podem auxiliar na operação em tempo real. Outra vertente com potencial de impulsionar o aprendizado é a ludificação do aprendizado baseada em pontuações, seja de pronúncia ou outros aspectos, bem como a expansão para outros idiomas.

Em suma, a arquitetura proposta estabelece uma base para o desenvolvimento de agentes conversacionais empáticos e tecnicamente acurados, aproximando o estudante brasileiro da fluência na língua inglesa.

REFERÊNCIAS

- AMODEI, Dario et al. **Deep Speech 2: End-to-End Speech Recognition in English and Mandarin**. [S.l.: s.n.], 2015. arXiv preprint arXiv:1512.02595. Disponível em: <https://arxiv.org/abs/1512.02595>. Acesso em: 6 jun. 2025.
- ANTHROPIC. **The Claude 3 Model Family: Opus, Sonnet, Haiku**. [S.l.], mar. 2024. Model card. Disponível em: https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.
- ARRIAGA, Carlos et al. Assessing Latency in ASR Systems: A Methodological Perspective for Real-Time Use. **IEEE Internet Computing**, Piscataway, NJ, 2025. In press. Disponível em: <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=11186205>. Acesso em: 30 nov. 2025.
- BAEVSKI, Alexei et al. **wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations**. [S.l.: s.n.], 2020. arXiv: 2006.11477 [cs.CL]. Disponível em: <https://arxiv.org/abs/2006.11477>.
- BEAUME, Cédric. **Understanding the Elo Rating System**. [S.l.: s.n.], 2020. <https://cbeaume.com/download/EloRatingSystem.pdf>. Comments and questions: ced.beaume@gmail.com.
- BECHTOLD, Bastian. **soundfile: An audio library based on libsndfile, CFFI and NumPy**. [S.l.: s.n.], 2025. <https://pypi.org/project/soundfile/>. Version 0.13.1, released 2025-01-25.
- BERGMANN, Dave. **What is instruction tuning?** IBM Think. 2025. Disponível em: <https://www.ibm.com/think/topics/instruction-tuning>. Acesso em: 15 fev. 2026.
- BERNARD, Mathieu. **phonemizer**. [S.l.: s.n.], 2021. <https://pypi.org/project/phonemizer/3.0.1/>. Versão 3.0.1, lançado em 18 de dezembro de 2021.
- CARDOSO, Walcir. The Development of sC Onset Clusters in Interlanguage: Markedness vs. Frequency Effects. In: SLABAKOVA, Roumyana et al. (ed.). **Proceedings of the 9th Generative Approaches to Second Language Acquisition Conference (GASLA 2007)**. Somerville, MA: Cascadilla Proceedings Project, 2008. p. 15–29. Disponível em: <https://www.lingref.com/cpp/gasla/9/paper1622.pdf>. Acesso em: 6 jun. 2025.
- CHOMSKY, Noam. A Review of B. F. Skinner's *Verbal Behavior*. **Language**, v. 35, n. 1, p. 26–58, 1959. Disponível em: https://chomsky.info/1967____/. Acesso em: 11 jan. 2026.
- CHOMSKY, Noam. **Language and Mind**. 3. ed. Cambridge: Cambridge University Press, 2006. p. 88–102. Disponível em:

<https://www.ugr.es/~fmanjon/Language%20and%20Mind.pdf>. Acesso em: 11 jan. 2026.

DING, Tianyu et al. The Efficiency Spectrum of Large Language Models: An Algorithmic Survey. **arXiv preprint arXiv:2312.00678**, New York, abr. 2024. Disponível em: <https://arxiv.org/pdf/2312.00678v2.pdf>. Acesso em: 10 jan. 2026.

EPANESHNIKOV, Alexandr; DUNN, Reece H.; VITOLINS, Valdis. **eSpeak NG: An Open Source Speech Synthesizer**. [S.l.]: GitHub, 2026. <https://github.com/espeak-ng/espeak-ng>.

FORNEY, G. David. The Viterbi Algorithm. **Proceedings of the IEEE**, v. 61, n. 3, p. 268–278, mar. 1973.

GREGG, Kevin R. Krashen's Monitor and Occam's Razor. **Applied Linguistics**, v. 5, n. 2, p. 79–100, 1984. DOI: 10.1093/applin/5.2.79. Disponível em: <https://academic.oup.com/applij/article-abstract/5/2/79/315322>. Acesso em: 11 jan. 2026.

HEXGRAD. **Kokoro-82M**. [S.l.]: Hugging Face, 2025. <https://huggingface.co/hexgrad/Kokoro-82M>. Text-to-Speech model. License: Apache-2.0. DOI: 10.57967/hf/4329.

HUNTER, J. D. Matplotlib: A 2D graphics environment. **Computing in Science & Engineering**, IEEE COMPUTER SOC, v. 9, n. 3, p. 90–95, 2007. DOI: 10.1109/MCSE.2007.55.

JI, Zhenya; JIANG, Ming. A systematic review of electricity demand for large language models: evaluations, challenges, and solutions. **Renewable and Sustainable Energy Reviews**, v. 225, p. 116159, 2026. DOI: 10.1016/j.rser.2025.116159. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1364032125008329>. Acesso em: 26 jan. 2026.

JUANG, B. H.; RABINER, Lawrence R. **Automatic Speech Recognition – A Brief History of the Technology Development**. [S.l.: s.n.], 2004. Disponível em: https://web.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/354_LALI-ASRHistory-final-10-8.pdf. Acesso em: 20 jun. 2025.

JURAFSKY, Daniel; MARTIN, James H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models**. Stanford, CA e Boulder, CO: [s.n.], 2024. Draft, Third Edition. Draft of August 20, 2024. Disponível em: https://web.stanford.edu/~jurafsky/slp3/ed3bookaug20_2024.pdf. Acesso em: 20 jan. 2026.

KRANTZ, Tom; JONKER, Alexandra. **O que é similaridade de cosseno?** IBM Think. 2024. Disponível em: <https://www.ibm.com/br-pt/think/topics/cosine-similarity>. Acesso em: 17 fev. 2026.

KRASHEN, Stephen D. **Principles and Practice in Second Language Acquisition**. [S.l.]: s.n., 1982. Disponível em: https://www.sdkrashen.com/content/books/principles_and_practice.pdf. Acesso em: 6 jun. 2025.

LAI, Jennifer; KARAT, Clare-Marie; YANKLOVICH, Nicole. Conversational Speech Interfaces and Technologies. In: JACKO, Julie A.; SEARS, Andrew (ed.). **The Human-Computer Interaction Handbook**. Mahwah, NJ: Lawrence Erlbaum, 2003. p. 698–716. Disponível em: URL_do_PDF_que_voc%C3%AA_usou. Acesso em: 20 jun. 2025.

LEE, Seongyong et al. Generative AI (GenAI) in the language classroom: A systematic review. **Interactive Learning Environments**, 2025. Ahead of print. DOI: 10.1080/10494820.2025.2498537. Disponível em: <https://www.tandfonline.com/doi/pdf/10.1080/10494820.2025.2498537>. Acesso em: 11 jan. 2026.

LEVIS, John; SUVOROV, Ruslan. Automatic Speech Recognition. In: CHAPPELLE, Carol A. (ed.). **The Encyclopedia of Applied Linguistics**. Malden, MA: Wiley-Blackwell, 2013. p. 316–323. Disponível em: https://www.researchgate.net/publication/261287458_Automatic_Speech_Recognition. Acesso em: 20 jun. 2025.

LEVIS, John M. Changing Contexts and Shifting Paradigms in Pronunciation Teaching. **TESOL Quarterly**, Malden, MA, v. 39, n. 3, p. 369–377, set. 2005. Disponível em: <https://www.jstor.org/stable/3588485>. Acesso em: 25 jun. 2025.

LIU, Alexander H. et al. Voxtral. **arXiv preprint arXiv:2507.13264**, 2025. Disponível em: <https://arxiv.org/abs/2507.13264>.

LONG, Michael H. Input, Interaction, and Second-Language Acquisition. **Annals of the New York Academy of Sciences**, New York, v. 379, p. 259–278, 1981. Disponível em: <https://nyaspubs.onlinelibrary.wiley.com/doi/10.1111/j.1749-6632.1981.tb42014.x>. Acesso em: 6 jun. 2025.

MAGALHÃES, Regis Pires et al. Evaluation of Automatic Speech Recognition Approaches. **Journal of Information and Data Management**, Porto Alegre, v. 13, n. 3, p. 366–377, set. 2022. Disponível em: <https://irislab.ce.gov.br/wp-content/uploads/2023/01/2514-Article-Text-11375-1-10-20220919.pdf>. Acesso em: 30 nov. 2025.

META. **Llama 3.1 8B**. [S.l.: s.n.], jul. 2024. <https://huggingface.co/meta-llama/Llama-3.1-8B>. Model card on Hugging Face.

MICROSOFT et al. **Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs**. [S.l.: s.n.], 2025. arXiv: 2503.01743 [cs.CL]. Disponível em: <https://arxiv.org/abs/2503.01743>.

NASEREDDIN, Hebah H. O.; OMARI, Ayoub Abdel Rahman. Classification Techniques for Automatic Speech Recognition (ASR) Algorithms Used with Real Time Speech

Translation. In: PROCEEDINGS of the Computing Conference 2017 (SAI). London: [s.n.], 2017. Disponível em: https://www.researchgate.net/profile/Hebah-H-0-Nasereddin/publication/325396450_Classification_Techniques_for_Automatic_Speech_Recognition_ASR_Algorithms_used_with_Real_Time_Speech_Translation/links/5b0b7160a6fdcc8c25345e28/Classification-Techniques-for-Automatic-Speech-Recognition-ASR-Algorithms-used-with-Real-Time-Speech-Translation.pdf. Acesso em: 20 jun. 2025.

NVIDIA CORPORATION. **NeMo Agent Toolkit**. [S.l.: s.n.], 2026. <https://developer.nvidia.com/nemo-agent-toolkit>.

OPENAI. **openai-whisper: Robust Speech Recognition via Large-Scale Weak Supervision**. [S.l.: s.n.], 2025. <https://pypi.org/project/openai-whisper/>. Versão 20250625, lançado em 26 jun. 2025.

PINKER, Steven; BLOOM, Paul. Natural Language and Natural Selection. **Behavioral and Brain Sciences**, New York, v. 13, n. 4, p. 707–784, 1990. Disponível em: https://stevenpinker.com/files/pinker/files/pinker_bloom_1990.pdf. Acesso em: 11 jan. 2026.

RADFORD, Alec et al. **Robust Speech Recognition via Large-Scale Weak Supervision**. [S.l.: s.n.], 2022. arXiv preprint arXiv:2212.04356. Disponível em: <https://arxiv.org/pdf/2212.04356>. Acesso em: 6 jun. 2025.

SEKOYAN, Monica et al. **Canary-1B-v2 Parakeet-TDT-0.6B-v3: Efficient and High-Performance Models for Multilingual ASR and AST**. [S.l.: s.n.], 2025. arXiv: 2509.14128 [cs.CL]. Disponível em: <https://arxiv.org/abs/2509.14128>.

SHAHIN, Mostafa; AHMED, Beena. Phonological-Level Mispronunciation Detection and Diagnosis. In: PROCEEDINGS of Interspeech 2024. Kos, Greece: [s.n.], 2024. p. 307–311. DOI: 10.21437/Interspeech.2024-2217. Disponível em: https://www.isca-archive.org/interspeech_2024/shahin24_interspeech.pdf. Acesso em: 3 jan. 2026.

SRIVASTAV, Vaibhav; FOURRIER, Clémentine et al. **TTS Arena 2.0: Benchmarking Text-to-Speech Models in the Wild**. [S.l.]: Hugging Face, 2025. <https://huggingface.co/spaces/TTS-AGI/TTS-Arena-V2>.

SRIVASTAV, Vaibhav; ZHENG, Steven et al. **Open ASR Leaderboard: Towards Reproducible and Transparent Multilingual and Long-Form Speech Recognition Evaluation**. [S.l.: s.n.], 2025. arXiv preprint arXiv:2510.06961. Disponível em: <https://arxiv.org/pdf/2510.06961>. Acesso em: 12 nov. 2025.

SWAIN, Merrill; LAPKIN, Sharon. Problems in Output and the Cognitive Processes They Generate: A Step Towards Second Language Learning. **Applied Linguistics**, Oxford University Press, Oxford, v. 16, n. 3, p. 371–391, 1995. Disponível em: <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=f8933309c4604e95933627054926d63257d6456c>. Acesso em: 6 jun. 2025.

TAI, Tzu-Yu; CHEN, Hao-Jan Howard. The Impact of Intelligent Personal Assistants on Adolescent EFL Learners' Listening Comprehension. **Computer Assisted Language Learning**, v. 35, n. 7, p. 1519–1543, 2022. DOI: 10.1080/09588221.2022.2040536.

Disponível em:

<https://www.tandfonline.com/doi/abs/10.1080/09588221.2022.2040536>. Acesso em: 11 jan. 2026.

TAN, Xu et al. **A Survey on Neural Speech Synthesis**. [S.l.: s.n.], 2021. arXiv preprint arXiv:2106.15561. Disponível em: <https://arxiv.org/pdf/2106.15561>.

Acesso em: 6 jun. 2025.

TAYLOR, Paul. **Text-to-Speech Synthesis**. Cambridge: Cambridge University Press, 2009. ISBN 9780521899277.

THENNAL, D. K. et al. Advocating Character Error Rate for Multilingual ASR Evaluation. In: FINDINGS of the Association for Computational Linguistics: NAACL 2025. Minneapolis, MN: Association for Computational Linguistics, abr. 2025.

p. 4941–4950. Disponível em:

<https://aclanthology.org/2025.findings-naacl.277.pdf>. Acesso em: 30 nov. 2025.

TOOBY, John; DEVORE, Irvén. The Reconstruction of Hominid Behavioral Evolution Through Strategic Modeling. In: KINZEY, Warren G. (ed.). **The Evolution of Human Behavior: Primate Models**. Albany, NY: State University of New York Press, 1988.

p. 183–237. Disponível em:

<https://www.cep.ucsb.edu/wp-content/uploads/2023/05/Reconst.pdf>. Acesso em: 11 jan. 2026.

TU, Huu Tuong et al. **Mispronunciation Detection and Diagnosis Without Model Training: A Retrieval-Based Approach**. [S.l.: s.n.], 2025. arXiv preprint arXiv:2511.20107. Disponível em: <https://arxiv.org/pdf/2511.20107v1.pdf>. Acesso em: 3 jan. 2026.

VAESSEN, Nik. **jiwer: Evaluate your speech-to-text system with similarity measures such as word error rate (WER)**. [S.l.: s.n.], 2025.

<https://pypi.org/project/jiwer/>. Versão 4.0.0, lançado em 19 jun. 2025.

VASWANI, Ashish et al. Attention is All You Need. **Advances in Neural Information Processing Systems**, v. 30, 2017. Disponível em:

<https://arxiv.org/abs/1706.03762>. Acesso em: 8 jan. 2026.

VIVA AMÉRICA. **Pesquisa: O inglês do brasileiro que migra para os EUA**.

Washington, D.C., jun. 2024. Disponível em: https://e1c27803-faa2-4842-9f65-62c163bc30a3.usrfiles.com/ugd/e1c278_68d97b46f4ee47dba6fb4f800edfaba2.pdf. Acesso em: 1 dez. 2024.

W3TECHS. **Technologies Overview: Content Language**. [S.l.: s.n.], 2024.

Disponível em: https://w3techs.com/technologies/overview/content_language.

Acesso em: 1 dez. 2024.

WANG, Bao. **Lecture 14. Self-attention Mechanism and Transformers**. Salt Lake City, UT: [s.n.], 2023. Lecture notes, Math 5750/6880, Fall 2023. Disponível em: <https://www.math.utah.edu/~bwang/mathds/Lecture14.pdf>. Acesso em: 10 jan. 2026.

WARSCHAUER, Mark; HEALEY, Deborah. Computers and Language Learning: An Overview. **Language Teaching**, v. 31, p. 57–71, 1998. Disponível em: https://education.uci.edu/uploads/7/2/7/6/72769947/computers_and_language_learning_-_an_overview.pdf. Acesso em: 10 jan. 2026.

WASKOM, Michael L. seaborn: statistical data visualization. **Journal of Open Source Software**, The Open Journal, v. 6, n. 60, p. 3021, 2021. DOI: 10.21105/joss.03021. Disponível em: <https://doi.org/10.21105/joss.03021>.

WEINBERGER, Steven H. **Speech Accent Archive**. Fairfax, Virginia: [s.n.], 2005. Online resource. Acesso em: 28 jan. 2026. Disponível em: <https://accent.gmu.edu/>. Acesso em: 28 jan. 2026.

WITT, Silke M.; YOUNG, Steve J. Phone-level pronunciation scoring and assessment for interactive language learning. **Speech Communication**, v. 30, n. 2–3, p. 95–108, 2000. DOI: 10.1016/S0167-6393(99)00044-8. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0167639399000448>. Acesso em: 3 jan. 2026.

WORLD POPULATION REVIEW. **EF English Proficiency Index by Country**. [S.l.: s.n.], 2024. Disponível em: <https://worldpopulationreview.com/country-rankings/ef-english-proficiency-index-by-country>. Acesso em: 1 dez. 2024.

XIAO, Feiwen et al. Conversational agents in language learning. **Journal of China Computer-Assisted Language Learning**, De Gruyter, Berlin, v. 4, n. 2, mar. 2023. DOI: 10.1515/jccall-2022-0032. Disponível em: <https://www.degruyterbrill.com/document/doi/10.1515/jccall-2022-0032/html>. Acesso em: 10 jan. 2026.

ANEXO A – VÍDEOS UTILIZADOS NO *CORPUS* PERSONALIZADO

De forma a construir o *corpus* personalizado com brasileiros falando inglês, buscou-se os seguintes vídeos já legendados no YouTube:

- ATALA, Alex. **D.O.M.: Rediscovering Brazilian Ingredients**. Talks at Google. Disponível em: <https://www.youtube.com/watch?v=3RdKiPxPiUY>.
- AZEVEDO, Tasso. **Hopeful lessons from the battle to save rainforests**. TED. Disponível em: https://www.youtube.com/watch?v=u70RZtCJ_rI.
- CARVALHO, Ilona Szabó de. **4 lessons I learned from taking a stand against drugs and gun violence**. TED. Disponível em: <https://www.youtube.com/watch?v=sYy-LEdYr8A>.
- DASS, Angélica. **The beauty of human skin in every color**. TED. Disponível em: <https://www.youtube.com/watch?v=NiMg0klgeos>.
- EFROMOVICH, Sid. **5 techniques to speak any language**. TEDxUpperEastSide. Disponível em: https://www.youtube.com/watch?v=-WLHr1_EVtQ.
- MEDICI, Patrícia. **The coolest animal you know nothing about ... and how we can save it**. TED. Disponível em: <https://www.youtube.com/watch?v=RbNqU2K0qzo>.
- MUNDANO. **Trash cart superheroes**. TED. Disponível em: <https://www.youtube.com/watch?v=jT8G44XPYW8>.
- NICOLELIS, Miguel. **A monkey that controls a robot with its thoughts. No, really**. TED. Disponível em: https://www.youtube.com/watch?v=CR_LBcZg_84&ab_channel=TED.
- NICOLELIS, Miguel. **Brain-to-brain communication has arrived. How we did it**. TED. Disponível em: <https://www.youtube.com/watch?v=HQzXqjT0w3k>.
- OROFINO, Alessandra. **It's our city. Let's fix it**. TED. Disponível em: <https://www.youtube.com/watch?v=zMGE3mbS9NY>.
- SEMLER, Ricardo. **Radical wisdom for a company, a school, a life**. TED. Disponível em: <https://www.youtube.com/watch?v=k4vzhwe0efs>.
- TORTURRA, Bruno. **Got a smartphone? Start broadcasting**. TED. Disponível em: https://www.youtube.com/watch?v=2r_qTWY0Iog.
- YAWANAWÁ, Nixiwaka. **We are all connected with nature**. TEDxHackney. Disponível em: https://www.youtube.com/watch?v=xk0-yebNA_o.

- YAWANAWÁ, Nixiwaka. **Why Brazil's indigenous people fight for the Amazon rainforest**. TEDxBedford. Disponível em: https://www.youtube.com/watch?v=hk1t2xW2DCg&ab_channel=TEDxTalks.